《人文及社會科學集刊》
第十一卷第二期 (88/6), pp. 155-182
©中央研究院中山人文社會科學研究所

Evolution of Cooperation with a Knowledgeable Mutant

Chun-lei Yang*
ISSP, Academia Sinica

ABSTRACT

How a society manages to solve the problem of cooperation in a prisoners-dilemma situation has been a major theme in economic theory. In the present paper, we develop an evolutionary model with a structural mutant called a knowledgeable (or cautious) dove, which is a dove in nature but one programmed to invest in gaining the ability to identify hawks and therefore avoiding their exploitation. We show that cooperative behavior can persist within a large class of admissible and compatible dynamics in the long run, either in the form of an asymptotically stable mixed population equilibrium or in the form of a stable cycle. Discussing other conceivable mutants, we argue that this persistent result is robust, which also matches our everyday observations regarding both the persistence and differentiation of cooperative behavior quite well.

Key Words: Prisoners' dilemma, cautious dove, dynamic, attracting sets, persistence of cooperation

* I would like to thank Erwin Amann, Heinz Holländer, Wolfgang Leininger, Karl Schlag and Jochen Weimann for comments and conversation, and for the Deutsche Forschungsgemeinschaft for financial support.

Received: June 15, 1998; Accepted: October 13, 1998

Introduction

The history of human society and its economic progress has been a history of ongoing specializations and cooperations of individual activities. How to realize the potential gains from trade is particularly important a problem in the situation of prisoners' dilemma (PD), where it is seemingly individually rational (i.e. it is a dominant strategy) to cheat on the partner with the consequence that nobody cooperates in equilibrium. The very puzzle to economists is however that in most of these situations cooperative behavior has been exhibited so frequently that potential gains from trade have been realized to some quite remarkable extent indeed, inspite of the coexistence of noncooperative behavior. How can this be explained?

The traditional game theory has attempted to rationalize this observation via introducing repetition of the basic one-shot PD game played by rational players. If repetition is infinite, the famous folk theorem is valid which says that virtually all admissible payoff vectors can be sustained in subgame perfect equilibrium, particularly the cooperative ones, (see e.g. Fudenberg and Maskin (1986)). The famous "gang of four" paper by Kreps *et al* (1982) tells us that finite repetition may be sufficient to provide credible threats to enforce cooperation when there is some slight inherent uncertainty about the rationality of the opponent such that he may be believed to be a Tit-for-Tat automata.

To explain the emergence and viability of the private judges system on Champagne Fairs in the Middle Age and the role of institution in information transmission, Milgrom, North and Weingast (1990) have a model where players are randomly matched in each period to play the one-shot PD game. No player in any period can fall back on any private information about the opponent such that the usual punishment strategies for cooperation do not apply. They argue that the presence of the private judges system, where records on breaches of contract are gathered and offenders are sentenced to fines on a voluntary basis, provides enough incentives for the cooperative outcome to prevail. Kandori (1992) develops some collective, so-called "contagious" punishment scheme that sometimes can sustain cooperation in the above setting of anonymous random matching but without an institution like the afore-mentioned private judges system. By introducing noises attached to actions, Ellison (1993) can show that cooperation through contagious punishment is not as vulnerable a result as it seems to be in the original setting by Kandori.

However, this rationality-based approach suffers from two general shortcomings. First, it ignores the fact that human players are only of (often very) bounded rationality. In addition, the solution concept of subgame perfection is logically not consistent with rationality as evidenced in the famous centipede-game by Rosenthal (1981). Thus, as argued by Mailath (1992), evolutionary models provide a suitable framework in this case. Friedman (1992) counts two basic requirements as responsible for the properness of an evolutionary modeling. First, there must be some inertia in individual reactions due to lack of information, adjustment delay, other decision costs, etc., which lead to a more or less smooth aggregate population dynamic. Second, players should not systematically attempt to influence others' future actions, possibly due to lack of means or to their bounded- rationality induced myopia.[1] This means that players treat the environment of their decision as of some more or less regular but independent nature, quite similar to price-taking consumers in the general equilibrium theory.

The most crucial term in an evolutionary analysis is the dynamic motion of aggregate behavior. At this aggregate level, the difference to biological models of evolution, where individual players are carriers of some inherited genotypes, lies only in the specific form of the dynamics under consideration. Common to both, however, is the property that shares of fitter strategies (respectively genotypes) in the population should increase relative to less fit ones, regardless of how the population dependent average fitness for a strategy (genotype) may be specified.

In this spirit, Axelrod (1984) provides the idea that cooperation in repeated PD game can prevail in the long-run in form of the simple Tit-for-Tat strategy which is to stick to cooperation with a built-in forgiving punishment against defection.[2] Assuming that the complexity of automata enters the fitness consideration in a lexicographic order next to the usual

[1] Myopic players have been blended into a reciprocity model of cooperation in PD-like contexts by Ghosh and Ray (1997) where non-myopic players have the option to stay matched with the same partner for a long-term relationship. The presence of myopic players serves as a disciplining device to make defections from a long-term cooperation a very costly adventure. While the ongoing process of trust building towards cooperation is illuminated vigorously, the population of the myopics has not been endogenized which is exactly the missing link to a population-dynamic model.

[2] See also discussions in Milgrom (1984).

"limit-of-the-mean" payoff,[3] Binmore and Samuelson (1992) show that a population of merely the Tit-for-Tat type can not be evolutionarily viable as it can be invaded by the simple cooperation-forever type. A nasty simple strategy called Tat-for-Tit, however, is viable. The latter starts with defection and switches to cooperation as soon as the opponent defects as well. This cooperation phase ends whenever the opponent defects, which will jump-start the automaton anew. The very virtue of this nasty machine is its ability to exploit the naive cooperating type but not exploited by defecting types while entering a cooperation with peers.

This basic feature of Binmore and Samuelson (1992) is also related to the so-called "secret-handshake" model by Robson (1990), where only the one-shot PD game is played in any matching period. It is obvious that defection-only is the only long-run stable state in this setting. This state, however, can be (locally) invaded by a mutation type that bears the code to cooperate if and only if the opponent is of the same mutant type. Peers of this group can for example be identified by a red spot on the forehead. Obviously, this mutant would drive the simple defectors to extinction, which implies that only some kind of conditioned cooperation would survive in the long-run. This however is a fallacy as Robson (1990) points out. An imitator of the first mutant who carrys the same red spot but never cooperates has substantial fitness advantage that will drive the original mutant to extinction. In fact, Banerjee and Weibull (1993) show in a generalized setting allowing arbitrary conditional strategies that only those population states that correspond to symmetric Nash equilibria of the basic game without the sophisticated conditional types can ever survive evolutionarily. Hence, the only stable state in a PD game modified likewise is the pure population of defectors.

An essential feature in those models is that sophistication be *costless* such that the extension with conditional types is analogous to the cheap-talk extension in the traditional game theory.[4] Conceivable is yet the emergence of mutant types that are a bit more sophisticated but have to bear some cost accordingly,[5] so that, fitness-wise, they are not automatically

[3] This means that bōunded rationality results from the cost of complexity.

[4] I am grateful to Jörgen Weibull for this interpretation.

[5] See also Rosenthal (1993) for justifications of introducing "ad hoc" cost of complexity in bounded rationality context and for more evidences of the predominance of routined rather than optimized individual behavior.

superior to any other existing type. Moreover, the nature of information used for conditional actions is important for the mutant to survive against invasions of further alien conditional types.

Frank (1987, 1988) has a completely different approach. His main concern is to explain the observation that people often do things which hurt themselves without expectation of any apparent (reciprocal) future advantage from their deeds. E.g. people risk their own life to save others' from burning houses, iced rivers etc., or they leave tips to waiters in restaurants of remote cities they never will visit again. In his theory, all these are subconscious investments for people to increase their ability of creditibly signaling that they are honest and cooperative, which help to overcome the commitment, respectively hold-up, problem in cooperation. Viable signals occur e.g. in form of sentiments which can not be imitated costlessly and perfectly, in contrast to Robson's (1990) red spot. If the signal observed indicates that the partner is more likely to be a dishonest person, a rational individual is likely to take his outside option of no trade and to wait for the next potential partner to come around. Frank shows that only a mixed population can exist in any equilibrium.

In the present paper, a slightly sophisticated type called Knowledgeable Dove (or likewise "cautious" dove) is introduced into the world of the simple types of Hawk (=defection) and Dove (=cooperation) in the PD game. Its carriers are initially dove-patterned but have somehow developed the ability to detect the programmed action of their matched partner. The effort of collecting necessary information reduces their fitness to some extent in exchange for the ability to evade "exploitations", for instance by rejection of trade with a hawk. Depending on the current population state, every type may have the highest average fitness, in contrast to the basic evolutionary PD game and its Robson modification. Within this setting, there is always an equilibrium of completely mixed population which may be even the unique one. There are constellations in which it is asymptotically stable. Even if it is not, there can be a cycle around it that is asymptotically stable. This means that cooperation will persist forever. Particularly, all types will be observed in the long run which meets our real world observations quite well: Naive, gullible people often feel no need for higher-level sophistication as far as there are enough cautious doves around such that exploitation by hawks is not too severe a problem on avarage. There is a sound symbiosis among different cooperative types. Compared to Frank (1987), our model can be interpreted as stressing on the cost of decoding signals, so that honest people are divided into two classes, the cautious and the naive ones.[6]

This paper is organized as follows. The basic evolutionary framework of PD game and its modification with the knowledgeable dove will be discussed in the next section where static equilibria are also characterized. Section 3 is devoted to relevant notions of dynamic analysis in an economic model of evolution with emphasis on admissibility and compatibility. The persistence of cooperation in the long run and its robustness within a class of compatible and admissible dynamics will be shown in section 4. This occurs either in form of asymptotic stability of the completely mixed equilibrium or in form of cycles flowing around it. Moreover, we discuss the kinship-modified best-response dynamic as an example. Section 5 deals with issues of robustness of the present model and of the persistence result obtained as a whole. The last section contains some further concluding remarks.

Models with Knowledgeable Doves

It is assumed that there is a continuum of members of a society who are hosts of either of the behavior patterns, *H* (Hawk) or *D* (Dove). There is also a continuum of time periods in which members of that society are matched pairwise and randomly, say, to enter some joint venture with potential gains from trade. The fitness respectively payoff a member receives from this match in a period is shown in Figure 1.

	D	*H*
D	3	0
H	6	1

Figure 1. Fitness matrix in PD game

It says that *D* gets a payoff 3 respectively zero if he is matched to *D* respectively *H*, while *H* receives 6 respectively 1 when meeting *D* respectively *H*. This is a specific version of the famous Prisoners' Dilemma (PD) which is characterized by the property that *H* is the dominant strategy for a

[6] In rationality-based models of PD aimed at rationalization of cooperation one generally obtains some cooperation-only result.

rational player in a two-person game, being aware of the payoff consequences illustrated in Figure 1.[7]

Now, let $p_H \in [0,1]$ respectively $p_D = 1 - p_H$ be the proportion of Hs respectively Ds in the population at some arbitrary time. The average fitness for hawk is then $6p_D + p_H$ while that for dove is $3p_D$ in the PD-game. Assume that types of higher average fitness thrive over time while the proportions of types of lower average fitness shrink, it is obvious that the dove-type will die out as it always yields lower average fitness in whatever state. In the real world, however, honesty has been frequently observed in the PD context, which is apparently incompatible with this theoreticalconclusion. Our goal in the present paper is to offer an explanation by introducing a mutant behavior pattern: The Knowledgeable dove (K).

Let us assume that the carrier/host of K is programmed to invest some fixed amount of fitness $\delta > 0$ to investigate the programmed action of the partner matched to her as to whether he is an exploiting type or not and condition her action accordingly.[8] If her opponent is identified as a cooperator, i.e. of either K- or D-type, she will play D as well. If the opponent is believed to be a hawk, K is assumed to avoid the exploitation, via ways we will discuss below, such that the encounter with K yields H the same payoff as with a hawk. K's payoff in this match is assumed to be higher than D's and nonincreasing in H's share in the population. To attain her knowledge, K may be able to properly decode some hardly imitatible signals like expressions of sentiment, as in Frank (1987). Or she may conclude from some pre-trade interaction with the matched partner.

Let $p = (p_H, p_D, p_K) \in \Delta := \{x \in IR^3_+ : \sum_{i \in I} x_i = 1\}$ denote the state of population with $I := \{H, D, K\}$, and $c(\cdot) \in C^1$ be a nondecreasing, continuously differentiable function of p_H with $c(0) \geq 0$, the (expected) fitness functions for the types are given by

$$\begin{cases} f_H(p) = 6p_D + (1 - p_D) \\ f_D(p) = 3(1 - p_H) \\ f_K(p) = 3 - \delta - c(p_H). \end{cases} \tag{1}$$

A population vector $p \in \Delta$ is called a (symmetric) *Nash equilibrium* iff $f_i(p)$

[7] This specific form is taken from Binmore (1992). More general forms will not change the essence of our results.

[8] As information and the act of conditioning is costly, it is reasonable to assume that Ks are not necessarily distinguishable from Ds.

$=\max_j f_j(p)$ whenever $p_i>0$. The static implications of introducing the K-type are summarized in the following result.

Lemma 1 *Suppose there is a unique* $\bar{p}_H \in (0, 2/3)$ *that satisfies* $3\bar{p}_H - c(\bar{p}_H) = \delta$. *Then, the set of Nash equilibria is:*

$$NE=\begin{cases}\{M\} & \Leftrightarrow c^{-1}(2-\delta)>1\\ \{M, N\} & \Leftrightarrow c^{-1}(2-\delta)=1\\ \{M, N, \lambda_H\} & \Leftrightarrow c^{-1}(2-\delta)<1\end{cases}$$

where $M=(\bar{p}_H,\ \bar{p}_D := \frac{1}{5}(2-\delta-c(\bar{p}_H)), 1-\bar{p}_H-\bar{p}_D) \in \mathring{\Delta}$,

$N \in \{p : p_D=0\}, \quad \lambda_H=(1, 0, 0)$.

Proof: The lemma is a simple consequence of the following comparisons, which are necessary and sufficient for determination of Nash equlibria.

$$\begin{aligned} f_H(p) \lesseqqgtr f_D(p) &\Leftrightarrow 5p_D+3p_H-2 \lesseqqgtr 0\\ f_D(p) \lesseqqgtr f_K(p) &\Leftrightarrow c(p_H)-3p_H+\delta \lesseqqgtr 0\\ f_K(p) \lesseqqgtr f_H(p) &\Leftrightarrow -5p_D-c(p_H)+2-\delta \lesseqqgtr 0\end{aligned}$$

It is worth noticing that Lemma 1 remains valid for slight perturbations of the assumed structure of expected payoff (1). Hence, K's knowledge does not need to be perfect to attain the validity of all the results claimed in the present paper.[9] A graphic illustration of Lemma 1 is implicitly included in Figure 2 in the next section. Note that the condition in Lemma 1 is a restriction on the c function, for which our model analysis works. The existence of such c functions is ensured by the three cases in the following.

The form of K's fitness function has been taken fairly general which allows for different motivations as to how Ks are to react when matched with a hawk, how the costs of sophistication accrue to her, or which form of interaction underlies the cooperation problem. Some special cases are discussed here.

Case 1: $\boldsymbol{c(\cdot)=\delta p_H}$. Suppose that a voluntary trade in form of a PD game at a random match is the concern here. While D and H are programmed to enter the trade unconditionally, K will reject doing so with a hawk. Assume that a period lasts so long till all members of the society have eventually

[9] See Amann and Yang (1994) for a related model with imperfect knowledge for K.

accepted some trading partner. The K-type has to invest the amount δ for each round before settling for a partner. After the first round there are just equal numbers of Ks and Hs left without a trading partner as a consequence of K's behavior. After infinite rounds, every K must finally have found a D or K to trade with. By subtracting expected costs of sophistication from expected payoff from the trade entered, we have

$$f_K(p)=3-(1-p_H)\delta-p_H\delta\left(\sum_{t=0}^{\infty}2^{-t}\right) \tag{2}$$

$$=3-\delta-\delta p_H \tag{3}$$

It can be easily seen that if $\delta<1$ there is a unique NE $M=(\frac{\delta}{3-2\delta}, \frac{2}{5}-\frac{3}{5}p_H, 1-p_H-p_D)$, while three Nash equilibria are present with $N=(\frac{2-\delta}{\delta}, 0, 1-p_H)$ if $\delta\in(1, 3/2)$.[10]

Case 2: $\boldsymbol{c(\cdot)=2p_H}$. Think of the K-type as sticking to the principle "an eye for an eye, a tooth for a tooth". There is no outside option. Therefore, she will use the strategy H when randomly matched with a hawk. In this case, there are always three Nash equilibria. Alternatively, one can think of there being some option for K to remain autark, as in Frank (1987). It is then reasonable to assume that the autarky payoff is not lower than that when two Hs meet. In the extreme case they are actually the same, $c(\cdot)=2p_H$ results from this story. [11]

Case 3: $\boldsymbol{c(\cdot)=0}$. If the above discussed rematching does not incur further sophistication cost, we have $c(\cdot)=0$. The only symmetric NE is $M=(\delta/3, (2-\delta)/5, (9-2\delta)/15)$. We can imagine a range of models with re-matching cost between 0 and δ_{PH} where only one NE exists. A degenerate case is $\delta=0$ which implies that 100% K is the only NE, as K is superior to D everywhere and able to invade an H-society successfully.

[10] Similar time feature allowing infinite matching rounds in an infinitesimal unit of period can also be found e.g. in Binmore and Samuelson (1992). I am grateful to Erwin Amann for pointing out the correct algebraical form of this "evasion of exploitation" interpretation to me.

[11] Holländer (1993) provides an example of cooperated hunting among two prehistoric hunters which suits this interpretation quite well. Also see Amann and Yang (1994).

Preliminaries for Dynamic Analysis

Evolutionary models of multiperson interaction are interested in the aggregate behavior pattern of the whole population and its evolution over time which is determined by average fitness the subpopulations receive. Throughout this paper, we will deal with *continuous time, continuous state* dynamics only, which can be represented by some autonomous system of ordinary differential equation (ODE) $\dot{p}=F(p), F:\Delta\rightarrow IR^3$, where $\dot{p}=(\dot{p}_H, \dot{p}_D, \dot{p}_K)$ is the vector of time rates of change of the state variables $\dot{p}_H, \dot{p}_D, \dot{p}_K$. Let us start with a brief discusion as to properties any economically motivated dynamic should have.

Admissible dynamics[12]

To understand an evolutionary process is to determine the state the underlying dynamic system is in at any time, given the starting state. An *admissible dynamic system* on some domain D is defined as a one-parameter family of *phase flow* $\varphi_t: D\mapsto D, t\in[0,\infty)$, or equivalently, $\varphi_t:[0,\infty)\times D\mapsto D$, where

$$\varphi_{s+t}(x)=\varphi_s(\varphi_t(x)), \forall x\in D, \forall s, t\in[0,\infty) \tag{4}$$

and φ is continuous. Moreover, φ is differentiable in t from the right with $\dot{\varphi}^+(t,x):=\lim_{\substack{\tau\rightarrow t\\ \tau>t}}\frac{\varphi(\tau,x)-\varphi(t,x)}{\tau-t}$.

A dynamic system is called C^1 if φ is defined for $t\in(\infty,\infty)$ and differentiable in t.[13] And $F(x):=\frac{d}{dt}\Big|_{t=0}\varphi(t,x)=:\dot{x}\in T_xD$ is called the *vector field* generated by the flow φ where T_xD denotes the tangent space of D at x. F:$D\mapsto TD$ is also what we call an *(autonomous) ODE*. Arnold (1973, p.8) illuminatingly points out that "... the task of theory of ODE is to reconstruct the past and predict the future of the process (φ) from a knowledge of the local law of evolution (F)." The fundamental theorem of ODE (see Appendix) solves exactly this task for an open domain D.

Evolutionary dynamics applied in economics are, however, often discontinuous in nature such that this fundamental theorem of ODE is not sufficient for the analysis. For any arbitrary set S, let $\mathring{S}$, ∂S, and $\overline{S}$ denote the

[12] Readers with minor interests in the formal arguments may ignore this notion.

[13] See Hirsch and Smale (1974).

interior, the boundary, and the closure of S. Let D be the compact simplex Δ of an Euclidean space. An economically motivated dynamic on $\mathring{\Delta}$ often can not be continuously extended to its boundary, e.g. the linear dynamic discussed in Friedman (1991). Frequently, it may not even be continuous in $\mathring{\Delta}$, e.g. the best-response dynamic by Gilboa and Matsui (1991). Nevertheless, these specific dynamics (i.e. characterizations of the local law of evolution) can entail meaningful solution paths that correspond exactly to some admissible dynamic systems as defined above, but not C^1. The major difference is that, while the future in these systems is still uniquely defined, different histories to the same future are possible. Moreover, the path $\varphi_t(x)$ is no longer differentiable in t everywhere. This however deos not bother economists very much, as they are mainly interested in the system's long-run behavior in the future only.

Definition 1 *A dynamic (respectively vector field)* $F:\Delta\rightarrow IR^3$ *is called* admissible *if*

1. $\sum_{i\in I} F_i(x)=0,\ \forall x\in\Delta$, *i.e.* $F:\Delta\rightarrow T\Delta$.
2. $F_i(x)\geq 0$ *if* $x_i=0$, *i.e.* F *is inward pointing on* $\partial\Delta$.
3. *There is a unique admissible dynamic system* φ *satisfying* $\dot{\varphi}^+(t,x)=F(\varphi(t,x))$, *for all* $x\in\Delta$ *and* $t\geq 0$.

A set of sufficient conditions for F to be admissible, which substantiate the meaning of admissibility wherever referred to in the rest of the paper, is discussed in Appendix. F has essentially to be a C^1 map on Δ upto some nowhere dense set, and to behave there well enough to allow the pieced together solution system φ to be continuous in the state variable x.

Economically compatible dynamics

In most economic models, fitness is to be interpreted as some kind of (more or less) abstract, expected payoff: profit, utility, market position, etc. or some mixture of them. Dynamic changes are interpreted as aggregate results of individuals' imitation of more successful strategies (in particular those of closer relatives), adaptation to mainstream, learning via information, etc., besides the biological interpretation of natural selection. Some fundamental inertia is assumed such that the aggregate process attains the form of a more or less smooth flow.[14] It is often difficult to justify any

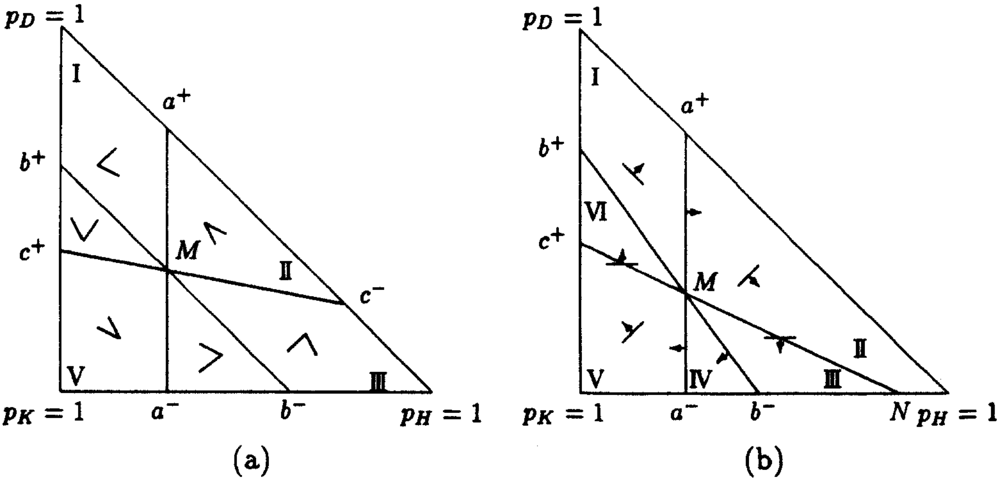
[14] Cf. Friedman (1992) and Mailath (1992) for further discussions on this issue.

specific form for the underlying dynamic. However, economists generally agree on some minimum requirement a reasonable dynamic in an economic model has to satisfy: populations of fitter strategies should increase relative to less fit ones over time — analogously to evolution of genotypes in biology.

Definition 2[15] *A dynamic F is called* (weakly) compatible, *respectively* order-compatible, *with fitness function f, if*

1. $\langle F(p), f(p)\rangle \geq 0\ \forall p \in \Delta$, *respectively* $F_i(p) > F_j(p)$ *whenever* $f_i(p) > f_j(p)$, *and*

2. $F(p) = 0$ *iff*, $\forall i \in I$, $p_i \neq 0$ *implies* $f_i(p) = \max_{j \in I} f_j(p)$.



(a)NE$=\{M\}$. Order-compatibility.

(b)NE$=\{M, N, \lambda_H\}$ and compatible vector fields.

Figure 2. Fixed point and compatibility

A graphical illustration of compatibility in our model can be found in Figure 2. Here, a^+a^-, b^+b^- and c^+c^- correspond to equations $f_K = f_D$, $f_H = f_D$ and

[15] Our definitions of compatibility are slightly different from the ones by Friedman (1991) as revival of extinct strategies is not allowed in his, which entails the unpleasant property that additional corner fixed points arise that are not (Nash) equilibrium of the underlying fitness function, and, that the boundary of Δ becomes artificially invariant. For this reason, we also stress on admissible dynamics in the Appendix that do not induce "alien" fixed point.

$f_H=f_K$ and fitness relations in different regions are (I) $f_H>f_D>f_K$, (II) $f_H>f_K>f_D$, (III) $f_K>f_H>f_D$, (IV) $f_K>f_D>f_H$, (V) $f_D>f_K>f_H$, (VI) $f_D>f_H>f_K$. It is easy to see in Figure 2.(b) that points N and λ_H are also FPs.

Since $\dot{p}_K=-\dot{p}_H-\dot{p}_D$, compatibility $\langle\dot{p},f(p)\rangle\geq 0$ implies $(\dot{p}_H,\dot{p}_D)^T(f_H-f_K,f_D-f_K)\geq 0$. The arrows correspond to the normal vector (f_H-f_K,f_D-f_K) in different state p. Graphically, compatibility is to require that the dynamic has to point to any direction on the positive halfspace generated by the according normal vector, i.e., on the arrow-pointed side of the orthogonal vector. According to the signs in the normal vector, Δ can be devided into four different areas by merging (I) with (VI) and (III) with (IV).

While the space of a compatible dynamic varies in each region (I) to (VI), dependent on the normal vector, that of an order-compatible one remains constant. In fact, each of the lines $p_H=p_D$, $p_H=-2p_D$, $p_H=-\frac{1}{2}p_D$ divides the plane into two halves corresponding to $\dot{p}_H\lesseqgtr\dot{p}_D$, $\dot{p}_H\lesseqgtr\dot{p}_K$, $\dot{p}_D\lesseqgtr\dot{p}_K$. Together, they partition the tangent plane into six pieces, each of which corresponds to one of the regions (I) to (VI) as a consequence of order-compatibility. Notice however that order-compatibility implies compatibility.[16]

Notions for stability analysis

Let $p(t,x)$ subsequently denote the solution path (trajectory, flow, orbit) of F starting in an initial state $x\in\Delta$, i.e. $p(\cdot,\cdot)$ is used instead of φ. An arbitrary set $B\subseteq\Delta$ is called *invariant* if $\forall x\in B$, $\forall t\in[0,\infty)$, $p(t,x)\in B$, i.e. the solution path starting from any arbitrary $x\in B$ is contained in B completely.

Let $L_\omega(x):=\{y\in\Delta\mid\exists\{t_n,n\in IN\}\subset[0,\infty):p(t_n,x)\xrightarrow{n\to\infty}y\}$ be the *ω-limit set* starting from x. An invariant set $B\subseteq\Delta$ is called *asymptotically invariant* (or *attracting*) if for all open set $U_1\supset\bar{B}$ there is an open set U_2 with $\bar{B}\subset U_2\subset U_2$ such that $\forall x\in U_2:L_\omega(x)\subseteq\bar{B}$ and $p(t,x)\in U_1\forall t\geq 0$.

If x is the only element in an invariant set, it is called a *fixed point* (FP, *equilibrium* or *stationary* point) of the dynamic F which has the property that $p(t,x)=x$ $\forall t$, i.e. the unique solution path starting in x is to stay in x forever. This happens iff $F(x)=0$ which is true for a compatible dynamic iff x is a Nash-equilibrium of the fitness function. A FP x^* is called *asymptotically stable* if x^* is an asymptotically invariant set. The *basin of*

[16] For more general (geometric) discussions of compatibility, see Friedman (1991, 1992).

attraction of x^* is then the set $B_{x^*}=\{x'\in\Delta: \lim_{t\to\infty}p(t, x')=x^*\}$. Obviously, B_{x^*} is an invariant set as well.

Let us define a *closed orbit* γ to be a flow with the property that whenever $p(\lambda, x)=x$ for some $\lambda>0$ it follows that $p(n\lambda, y)=y$ $\forall n\in IN$, $y\in\gamma$. γ is invariant. A *limit cycle* is a closed orbit γ such that $\gamma\subset L_\omega(x)$ for some $x\notin\gamma$. A limit cycle γ is *asymptotically stable* if γ is an asymptotically invariant set.

Persistence and Asymptotic Stability of Cooperative Behaviour

From a static point of view, the very feature induced by the introduction of the K type is that there is now a new, completely mixed equilibrium M. It is however straightforward to show that M is not an evolutionarily stable strategy in the sense of Maynard-Smith (1982). In what follows we show that cooperation can still be asympotically persistent nevertheless, which is precisely what we observe in reality.

A glance at Figure 2, tells us that close to the mixed FP M the dynamic system should flow around M clockwise. In general, M itself may not be asymptotically stable. Even if it is asymptotically stable, what turns out to be true under some further conditions, it does not mean that the (simplified) society will stay there in the long-run. There may exist stable cycles, other attractors or even chaos. If M is the only FP, the persistence of cooperation in the long-run is ensured for all compatible dynamics, since the state $p_H=1$ is never stable in this case. Unfortunately, this unconditional persistence of cooperation is not available if there are three FPs. Hence, we have to look for conditions under which this persistence is a robust property, i.e. M is either (locally) asymptotically stable or contained in an asymptotically invariant set.

A very useful tool for our analysis is the classical result by Poincaré-Bendixson which tells us that in the plane the limit set of a flow of a smooth dynamic must be either an equilibrium or a closed orbit or some mixture of both.

Proposition 1 *Let M be the unique FP, F be C^1 on $\mathring{\Delta}$, admissible, compatible, and possess only finitely many cycles, then either is M asymptotically stable or there is an asymptotically stable closed orbit surrounding it.*

Proof: Basicly, apply the Poincaré-Bendixson Theorem in the manner of Hirsch and Smale (1974). In details, suppose M is not asymptotically stable.

This implies that $\exists x\in\Delta{:}M\notin L_\omega(x)$. From the generalized Poincaré-Bendixson Theorem (appendix), $L_\omega(x)$ must be a closed orbit, since M is the only FP of F and $L_\omega(x)$ is obviously nonempty and compact as Δ is compact. It is well-known that there is always a FP in the open set enclosed by any closed orbit in Δ (see Hirsch and Smale (1974), Theorem 2, p.252). Hence, M must be in the interiors of all closed orbits which are ordered by their distances to M. Since the set $A_\gamma:=\{y|\gamma=L_\omega(y), y\notin\gamma\}$ is open for any closed orbit γ, if U is the open set between two neighbour closed orbits γ, α, then either $U\subset A_\gamma$ or $U\subset A_\alpha$. Let γ_i respectively U_i, $i=1,\ldots,n$, denote the given closed orbit respectively the open sets they enclose, where γ_i is closer to M than γ_j whenever $i<j$. It is obvious that $A_{\gamma_1}\cap U_1\neq\emptyset$ and $A_{\gamma_n}\cap(\Delta\setminus U_n)\neq\emptyset$ if $\gamma_n\neq\partial\Delta$. Hence, there necessarily exists some $i\in\{1,\ldots,n\}$ such that $A_{\gamma_i}\cap U_i\notin\emptyset$ and $A_{\gamma_i}\cap(U_{i+1}\setminus U_i)\notin\emptyset$ which implies $A_{\gamma_i}=U_{i+1}\setminus(U_{i-1}\cup\gamma_i\cup\gamma_{i-1})$, i.e. γ_i is asymptotically stable. If $\gamma_n=\partial\Delta$, it may happen that $\partial\Delta$ is the only asymptotically stable closed orbit.

The 3-FP case is a bit less favourable as the hawk-only FP is always asymptotically stable. Some conditions ensuring local robustness of persistence of cooperation can still be characterised. Let us call a closed orbit γ *locally repelling*, if $\exists x\in\gamma$, $\varepsilon>0$ and an open neighbourhood of x, U_x, such that diam $(L_\omega(x'), \gamma)>\varepsilon$, $\forall x'\in U_x$. We then have the following result.

Proposition 2 *Consider the 3-FP case. Let F be C^1 on $\mathring{\Delta}$, admissible, compatible, and of finitely many closed orbits. Then, λ_H is asymptotically stable with the basin of attraction B_H. Moreover, if M is not asymptotically stable but there exists some locally repelling closed orbit, then there exists some asymptotically stable closed orbit $\gamma\subset\Delta\setminus B_H$. The third FP, N, is never stable.*

Proof: If λ_H were not asymptotically stable there must exist a state x very close to λ_H such that $p(\cdot, x)$ is leaving the boundaries. This means $\dot{p}(t,x)=F(p(t,x))$ can not be compatible with f at some time t. For the same reason, N can not be stable.

If there is a closed orbit, then it must be contained in $B_H^C:=\Delta\setminus(B_H\cup\{M, N\})\neq\emptyset$ enclosing M in its inner side because the FP N is on the boundary. As argued in Proposition 1, if there are only finite closed orbits then one and only one of the two neighboring closed orbits must be the limit cycle for all point in the open set between them. If one closed orbit is locally repelling, then due to the finiteness of closed orbits, it can not be the limit cycle for either of the neighboring open sets. Similar argument as in Proposition 1 ensures the existence of an asymptotically stable closed orbit enclosing M.

Compatibility has so far been the only requirement imposed on dynamics we consider meaningful for our problem. In this sense, the robustness of persistence of cooperation is very striking a result. However, as the conditions required (particularly in Proposition 2) are quite abstract, it is not self-evident that the set of dynamics covered is not empty. This nonemptyness results from our subsequent analysis of a variant of the well-known best-response dynamic which is also of interest to the theory for its own sake.

Kinship-Modified Best-Response Dynamic

Gilboa and Matsui (1991) analyse a very simple dynamic for economic settings that can be easily dealt with geometrically — the so-called best-response (BR-) dynamic. Let $\tilde{\beta}(p)$ be the best response correspondence of f, i.e.

$$\tilde{\beta}(p) = \text{simplex spanned by } \{j : f_j(p) = \max_i f_i(p)\}.$$

For arbitrary $\alpha \in (0, 1)$, this dynamic is defined as

$$\dot{p} = F^{\beta}(p) = \alpha(\beta(p) - p)$$

where $\beta(p)$ is some selection of $\tilde{\beta}(p)$, such that $\beta(p) = p$ whenever p is a Nash equilibrium of f. The interpretation for this is that each of the continuum of players is inclined to behave myopically so as to take the best-response strategy with regard to the current population state, but only a fraction α of them will actually succeed in doing so, — because of the fundamental inertia inherent to any population dynamic.

Geometrically, the BR-dynamic requires trajectories that consist of straight line sections pointing to the extremal points of Δ. E.g. if H is the best reply to some current state p', then the flow will go along the straight line connecting p' with $\lambda_H = (1, 0, 0)$ as far as the path is still in the area where H is the best reply. This implies also that the proportional relation between the non-best reply strategies remains the same along that piece of path. The idea here is that, whenever one has no reason why one sub-population should shrink faster than another, the status quo ratio is used as a proxy. An illustration can be found in Figure 3.

This dynamic is obviously compatible, but not necessarily order-compatible. Apparently, for $p \in \partial\Delta$ trajectories are kept staying in Δ, since $p_i = 0$ implies $F_i^{\beta}(p) \geq 0$. Moreover, the dynamic is not continuous on the line sections b^+M, a^-M and c^-M. For the 1-FP case, it can be easily seen that the conditions of the Appendix are satisfied for F to be admissible.[17]

Note now that there is an inherent kinship between D and K types in our model, as both are cooperative by conviction. Thus, it is natural to assume that D is more likely to become a K if both K and H have similar fitness advantages, for the sake of affinity. Imagine the mental burden a honest person has to overcome to take on the cheater behavior she has detested a life long! Besides, compared to sophistication, cooperation or not is naturally deeper-seeded in one's conscience. Consequently, it is the last to be changed and there should be a more intensive population exchange between K and D, everything else being equal.

With the above discussion in mind, we introduce the notion of ***kinship-modified*** best-response dynamic as

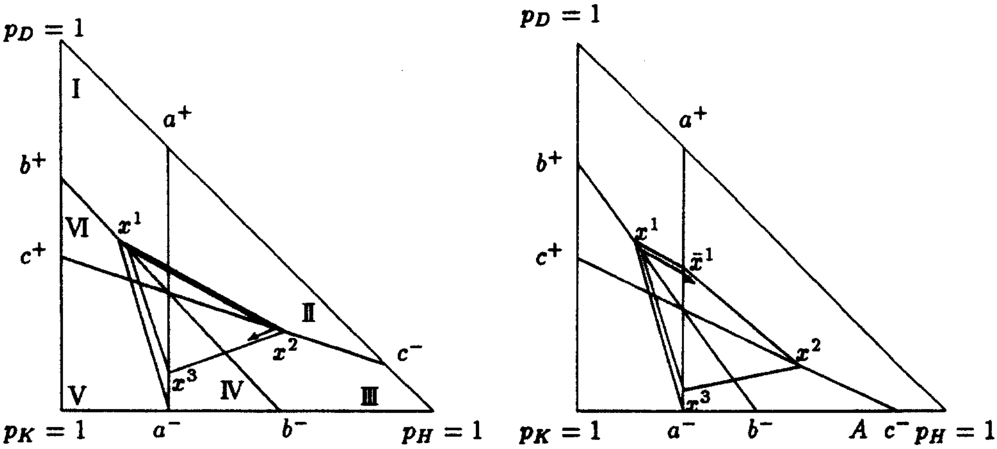
$$F^{\eta}(p)=\begin{cases}\eta\begin{pmatrix}0\\-\alpha\\\alpha\end{pmatrix}+(1-\eta)F^{\beta}(p) & \text{if } f_H>f_K>f_D\\ F^{\beta} & \text{otherwise}\end{cases}$$

where $\eta\in[0,1)$. If $\eta=0$, we have the original BR dynamic by Gilboa and Matsui. If $\eta=1$, we have a specific dynamic where in case of $f_H>\mathrm{f}_K>\mathrm{f}_D$ (i.e., in area (II) of Figure 2) the proportion of H remains constant while Ds are busily converted into Ks. In between, we have a class of dynamics that correspond to different degrees of kinship distortion. F^{η} is apparently compatible for all η as $(0,-\alpha,\alpha)$ is compatible in $f_H>\mathrm{f}_K>\mathrm{f}_D$. We can now show:

Proposition 3 *In the 1-FP case, F^{η} is admissible and M is contained in an asymptotically stable invariant set for all $\eta\in[0,1)$. In the 3-FP case, there is some $\bar{\eta}\in(0,1)$ such that M is contained in an asymptotically stable invariant set for any F^{η} with $1>\eta\geq\bar{\eta}$. Moreover, F^{η} is admissible for all $\eta\in(\bar{\eta},1)$.*

Proof: Consider the 1-FP case. Look at Figure 3(a). As only the best response counts, Δ can be partitioned in three relevant regions: (I)+(II), (III)+(IV), and (V)+(VI). Let us start at point a^-, the BR dynamic heading straight on the point $p_D=1$ leads to point x^1 on the curve section b^+M. From there on, the trajectory changes its direction straight towards $p_H=1$ hitting

[17] Warning: This is not the case if #NE=3, since the flow is leaving away from the curve c^-M. As discussed in the Appendix, any conceivable solution φ has then the problem of discontinuity in the states on c^-M. Fortunately, the relevant modified BR-dynamics below do not have this problem.



(a) (b)

Figure 3. Kinship-modified BR-dynamics

c^-M on some point x^2. Connecting x^2 with the point $p_K=1$ via straight line yields the point x^3 on a^-M. Following the BR dynamic, we get sequences $(x^{3n-2})_{n\in IN}$, $(x^{3n-1})_{n\in IN}$ respectively $(x^{3n})_{n\in IN}$ on the curve sections b^+M, c^-M, respectively a^-M that are monotone ordered towards M. It is obvious that the triangle T_n generated by the points $\{x^{3n}, x^{3n-1}, x^{3n-2},\}$ for arbitrary n is an invariant set under the BR dynamic. Moreover, $T_{n+1}\subset \mathring{T}_n$ for all n. Hence, $L_\omega(a^-)$ is asymptotically stable. Notice, $L_\omega(a^-)$ can be a non-trivial limit cycle as discussed by Gilboa and Matsui (1991). The proof for the 3-FP case is analogous as illustrated in Figure 3(b). If $\eta=1$, the solution orbit in region (II) will flow perpendicular to the boundary $p_D=0$ till it hits the c^-M line, whenever the H-share in the starting point is not too high. Due to continuity of F^η in η, there exists some $\bar{\eta}\in(0,1)$ such that the orbit will hit c^-M at some point x^2 for all $\eta\in(\bar{\eta},1)$.

A closer look at the proof reveals that there is a rich class of dynamics admitting an asymptotically stable set around M. For $\eta\geq\bar{\eta}$, consider $\bar{x}^1$ on a^+M and $x^2(\eta)$ on c^-M as shown in Figure 3(b). Suppose there is some y on c^-M such that $p(t(y),y)$ intersects the line a^+M at some interior point at the time $t(y)$. From continuity of $p(\cdot,\cdot)$ there is some $\eta'>\bar{\eta}$ such that $x^2(\eta')$ lies on the line yM and $p(t(x^2(\eta')), x^2(\eta'))$ lies on $\bar{x}^1M$. This however implies that there exists an attracting set around M. Consider some C^1-approximation $F^{\eta,\tau}$ of any F^η in $\mathring{\Delta}$ under Proposition 3. At least for $\eta\in(\bar{\eta},1)$ large enough $F^{\eta,\tau}$ can be chosen to be admissible. Hence the claims of Propositions 1 and 2 are not empty.

Issues of Robustness

As argued previously, both the specific form of the PD game and the assumption of perfect detection by K are used for simplicity which are not essential for the results about persistence of cooperation in the present paper. In particular, an imperfect detection can be roughly interpreted as a situation of perfect detection but with then higher cost of sophistication δ. Only if δ is too high, K, and therefore cooperation, has no chance to survive. Notice however that the introduction of the K-mutant here is meant to show that there are a large range of situations favorable for cooperation to prevail in a more unconscious way, as alternatives to reciprocal approaches or other formal institutions in the sense of North (1991).

A very important question is however how robust the persistence outcome in our now (H, D, K)-society is against further conceivable mutants. The answer depends crucially on the nature of information, in particular when there are new types. Without an explicitly specified dynamic we can hardly go further beyond the following verbal account of it.[18]

Consider a "knowledgeable hawk" who has the same abiltiy as to information gathering as a K but plays H unconditionally. If he is to be identified by K as a noncooperative type, his fitness is dominated by that of the simple H. This nasty type would not survive.

If K's knowledge comes from decoding the Frankian signals, one can think of some "deceiving hawk" who might invest to imitate signals of cooperation but always play H. If the cost of imitation is low, this type can decimate the population of K seriously. However, if p_K is too low, simple H will dominate the costly deceiving hawk. Hence, there is likely a cycle with all types present. However, low cost of imitation is hardly conceivable in the signaling setting of Frank (1988), as nobody would waste energy in such useless signals which would not survive the natural selection anyway. In the more reasonable case of high cost of imitation, it is difficult for the deceivers to get enough fitness compensation for their imitation costs. Put differently, these parasites need enough K s in the population in order to

[18] In a subsequent paper by Amann and Yang (1994) where replicator dynamic is applied to a generalization of the $c(\cdot)=2p_H$ case, some of them can be substantiated.

survive — persistence of K and D is hence not in danger.

Now, consider the Robson-type of mutation. Suppose they have to bear some costs compared to simple Hs. Let M be the long-run stable state here. If the invading population of Robson-mutants is very small, they have no chance to survive in the sense of the ESS by Maynard-Smith (1982), similarly as in the hawk-only equilibrium of the former (H, D)-society.[19] In addition, R-mutants are likely to be identified as noncooperative type by K which makes their survival more difficult.

The general lesson from Robson (1990) and Banerjee and Weibull (1993) is that costless mutation in PD-like situations can only be some transient phenomena and costly ones may entail substantial changes which is consistent with findings both of Frank (1987, 1988) and of the present paper. Hence, persistence of cooperation is generally to be expected even if more complex, derivative mutation types are considered, as their wellbeing relies essentially on the presence of the basic types H, D and K via parasitic or symbiotic relationships.

Concluding Remarks

There are different ways in which people manage to realize gains from trade in PD-like situations. Building reputations, making binding contracts and constructing legal systems to enforce these contracts are examples of those devices, which technically make use of the structure of repeated games. Embracing the idea of bounded rationality,[20] the emergence and persistence of the knowledgeable cooperator discussed in the present paper can be understood as some nature-established device solving this cooperation problem. Our approach in fact deals with what Mailath (1992, conclusion) describes as a structurally new mutation which not only tests the stability of current population but also changes the game being played. Basicly,

[19] Only in degenerate case of costless mutation can the red-spot mutant invade the hawk equilibrium independent of the invading population. If the population is too small, their fitness advantage from intragroup trading can be outweighed by the costs.

[20] Frank (1988) argues that the neurological capacity of humans is a scarce good which is to be economized on. It is less problematic to think that this economization happens in some evolutionary mode than to think that it be a calculating and optimizing consequence of this limited capacity itself.

we introduced a minimally sophisticated mutant and analysed its impact. We found that there is a nonempty class of reasonable dynamics, under which the cooperative behavior persists in the long run. Along the way, we discovered that the natural kinship between the naive and the sophisticated cooperators is instrumental for our result. Besides, in non-exhausitive way, we argued that our result of persistent cooperation is robust against further conceivable structural mutants, avoiding the weakness of Robson (1990). All in all, we showed that cooperation can persist in a simple, natural, and robust evolutionary model.

From a comparative static point of view, formal institutions can be devised to further facilitate cooperation. E.g. establishing the law merchant on Champagne fairs (see Milgrom, North and Weingast (1990)) can reduce the information costs for K and therefore increase their fitness advantage. This causes a shift of the fixed point M so that the population of hawks will decrease and the average fitness in equilibrium increase. Punishing hawk behavior via laws has however ambivalent effect. On the one hand, hawk's share will decrease. Simultaneously, sophistication becomes *less* attractive for doves. If there is some structural catastrophe as in time of political unrest, war etc., such that the previous legal system is no longer in use, then the dynamic may push the society to the hawks-only equilibrium — homo homini lupus. This supports the assertion by North (1991) that informal institutional constraints like customs or morals are more substantial for stability in a society than formal ones like laws.[21]

Appendix: On admissible dynamics

Our goal here is to characterize a class of the local law (vector field, dynamic, ODE) $F:\Delta\rightarrow T\Delta$ which have a unique solution path (dynamic system) $\varphi:[0,\infty)\times\Delta\rightarrow\Delta$ that is continuous and satisfies

$$\varphi_{s+t}(x)=\varphi_s(\varphi_t(x)) \qquad \forall x\in\Delta, \forall s, t\geq 0$$

and $\dot{\varphi}_t^{+}(x)|_{t=0}=F(x)$. For this we need the following classical result of ODE:[22]

[21] Consequently, as far as sophistication incurs costs such that naive simple behavior patterns always have a chance to survive, evolution of informal institutional constraints should be an essential part of any transaction cost analysis.

[22] See Arnold (1973) or Hirsch and Smale (1974) for a proof.

Fundamental Theorem *For any open Euclidean set D and $F{:}D \to TD$ being* C^1, there is a unique continuous $\varphi{:}(-\infty, \infty) \times D \to D$ such that $\dot{\varphi}_t(x)|_{t=0} = F(x) \forall x \in D$.

As discussed previously, economicly motivated dynamics are frequently discontinuous on a submanifold in $\mathring{\Delta}$ and require a proper extension to $\partial\Delta$ as well. (E.g. best-response and linear dynamics mentioned.) Our main insight here is that we still can obtain a unique economically meaningful dynamic system as a solution if the underlying dynamic is C^1 upto a nowhere dense set of similar structure of submanifold and does not behave too unfavourably in its neighbourhood.

Consider a finite partition of Δ, $P=\{D_i, M_j, E_h, i \in I_D, j \in I_M, h \in I_E\}$, where D_i open in IR^2, M_j a connected C^1 1-manifold with corners contained in $\bar{M} \backslash M = E = \{E_h, h \in I_E\}$, $M := \cup_{j \in I_M} M_j$ and $\partial\Delta \subset \bar{M}$. Let $\partial M = \bar{M} \backslash M$, ∂M_j denote the closures relative to the 1-manifolds. Alternaitvely, think of $\cup_{i \in I_D} \bar{D}_i = \Delta$ such that ∂D_i consists of finite pieces of C^1 manifolds with corners. Hence, every point $x \in M_j$ is some regular one such that there are either exactly two sets D_{j_1}, D_{j_2} with $x \in \partial D_{j_1} \cap \partial D_{j_2}$ or one such D_{j_1} in case $x \in \partial\Delta$ and the Gauss map $g^i{:}\partial D_i \to TIR^2$ is well-defined for every such regular point, with $g^i(x)$ being the unit length normal vector perpendicular to TD_i at x pointing outward relative to D_i.

Consider a given $F{:}\Delta \to T\Delta$. Suppose its restriction $F|_{Di}$ is a C^1 map for all i. For all $x \in \partial D_i$, define $F^i(x) := \lim_{\substack{y \to x \\ y \in \bar{D}_i}} F(y)$. W.r.t. D_i, F^i at $x \in \partial D_i \cap M$ is called *inward pointing, outward pointing* or *neutral*, if $\langle F^i(x), g^i(x) \rangle < 0, > 0$ or $= 0$. At any $x \in \partial M \cap \partial D_i$, the Gauss map is not defined as there exists exactly two sets M_{i_1}, $\mathrm{M}_{i_2} \in P$ with $x \in \partial M_{i_1} \cap \partial M_{i_2} \in \partial D_i$, but the limits $g^i_l(x) = \lim_{\substack{y \to x \\ y \in Mi_1}} g^i(y)$ and $g^i_r(x) = \lim_{\substack{y \to x \\ y \in Mi_2}} g^i(y)$ can be found. Moreover, x is a regular boundary point to the 1-manifolds M_i, $i = i_1, i_2$, so that the Gauss map $g^{Mi}{:}\partial M_i \to TM_i$ at x pointing outward w.r.t. M_i is well-defined, that satisfies $\langle g^{Mi_1}(x), g^i_l(x) \rangle = \langle g^{Mi_2}(x), g^i_r(x) \rangle = 0$, (i.e. they point to one direction in the normal space of M_i.) Therefore, any vector field $f{:}\tilde{M} \to T\tilde{M}$ for arbitrary C^1 1-manifold $\tilde{M}$ can be analogously defined to be *inward, outward* or *neutral* at $x \in \partial\tilde{M}$, where neutrality implies $f(x) = 0$.

Then F^i is called *inward pointing* at $x \in \partial D_i \cap \partial M$ w.r.t. D_i if

$$\begin{cases} \langle F^i(x), g^{M_j}(x) \rangle < 0 & \text{for both } j = i_1, i_2 & \text{if } \langle g^{M_j}(x), g^i_l(x) \rangle > 0 \\ \langle F^i(x), g^{M_j}(x) \rangle < 0 & \text{for at least one } j = i_1, i_2 & \text{if } \langle g^{M_j}(x), g^i_l(x) \rangle < 0 \end{cases}$$

If $\langle g^{M_{i_j}}(x), g^i_j(x) \rangle = 0$, then x is either a regular point on ∂D_i, i.e. there is some $j, x\in M_j$, or $g^{M_{i_1}}(x)=g^{M_{i_2}}(x)$ and F^i is *inward* pointing iff $F^i(x)\neq -g^{M_{i_1}}(x)$. A graphical illustration can be found in Figure 4

Let us further partition the set M into relevant subclasses: $M=\cup_s M^s=\cup_s \cup_{i\in I_s} M_i^s$ described in formula (5) below.

$$x\in\begin{cases} M^1\Leftrightarrow F^i(x)\in T_x\partial\Delta & \text{if } x\in\partial\Delta \\ M^2\Leftrightarrow F^i(x) \text{ inward} & \text{if } x\in\partial\Delta \\ M^3\Leftrightarrow F^i(x) \text{ outard} & \text{if } x\in\partial\Delta \\ M^4\Leftrightarrow F^i(x)=F^j(x) & \text{if } x\notin\partial\Delta \\ M^5\Leftrightarrow F^i(x) \text{ inward, } F^j(x) \text{ outard, } F^i(x)\neq F^j(x) & \text{if } x\notin\partial\Delta \\ M^6\Leftrightarrow F^i(x) \text{ neutral, } F^j(x) \text{ outward} & \text{if } x\notin\partial\Delta \\ M^7\Leftrightarrow F^i(x), F^j(x) \text{ outward} & \text{if } x\notin\partial\Delta \\ M^8\Leftrightarrow F^i(x) \text{ inward, } F^j(x) \text{ not outward} & \text{if } x\notin\partial\Delta \\ M^9\Leftrightarrow F^i(x)\cdot F^j(x) \text{ neutral but } F^i(x)\neq F^j(x) & \text{if } x\notin\partial\Delta \end{cases} \tag{5}$$

D_i, D_j are the relevant sets at an arbitrary $x\in M$ here.

Now, we are about to construct an extension $\tilde{F}:\bar{M}\to T\Delta$ such that $F|_{\Delta\setminus\bar{M}}\cup\tilde{F}:\Delta\to T\Delta$ yields us the desired dynamic system as solution. A trivial extension by setting $\tilde{F}=0$, though gives us a dynamic system, has the unpleasant property to have lots of unmotivated additional FPs. Our concern is to find an extension that provides no additional FP of F which is defined as some $x\in\bar{D}_i$ such that $F^i(x)=0$. We assume F is continuous at all FPs, i. e. $F^i(x)=0$ implies $F^j(x)=0$ whenever $x\in\partial D_i\cap\partial D_j$.

Let us assume that $M^s\subset M$ can be partitioned into finite maximal connected subsets M_i^s, $i\in I_s$ for all s. For $x\in M^s$, $s\in\{1, 2, 4, 5, 6\}$, $\tilde{F}:=F^i(x)$ where $F^i(x)$ is taken from (5). For $s\in\{3, 7, 8, 9\}$, it is obvious that no FP is contained in M_i^s as any FP is a continuous point of F. Moreover, $\partial M_i^s=:\{x_{ik}^s, k=l, r\}\subset\partial M\cup M^1\cup M^4\cup M^6$ due to continuity of $F|_{\Delta\setminus\bar{M}}$. If $\partial M_i^s\cap\partial M=\emptyset$, then $\tilde{F}|_{\partial M_i^s}(x):=\lim_{\substack{y\to x,\\ y\in M^{S'}}}\tilde{F}(y)$ for some $s'\in\{1, 4, 6\}$ and we hence have

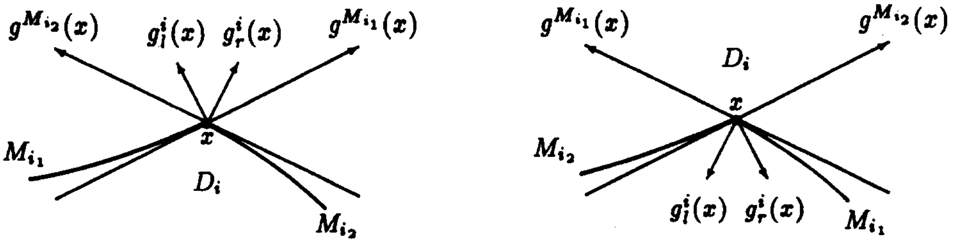


Figure 4. Directions at corners

$\tilde{F}(x_{ik}^s)\in T_{x_{ik}^s}M_i^s$, $k=l, r$. $\forall x\in\mathring{M}_i^s$, the relative interior to M_i^s, $\tilde{F}\in T_xM_i^s$ is defined to be the linear connection between $\tilde{F}(x_{ik}^s)$, $k=l, r$, w.r.t. the manifold $\mathring{M}_i^s$.[23] If $\partial M_i^s\cup\partial M\neq\emptyset$, $\tilde{F}|_{\mathring{M}_i^s}$ can be defined arbitrarily. For all s, i, an arbitrary C^1-extension $\tilde{F}:M_i^s\rightarrow TM_i^s$ is called *consistently directed* if $0\notin\tilde{F}(\mathring{M}_i^s)$ and there is no more than one $k\in\{l, r\}$ such that $\tilde{F}(x_{ik}^s)\in T_{x_{ik}^s}M_i^s$ is outward respectively inward pointing w.r.t. M_i^s.

Let $P(x)=\{\beta\in P : x\in\bar{\beta}\}$ denote the relevant sets at x. For any $x\in\partial M$, if there is a unique $\alpha\in P(x)\backslash\partial M$ for which $F^\alpha(x)$ is inward pointing, then $\tilde{F}(x) := F^\alpha(x)$. Otherwise $\tilde{F}(x)$:0. Finally, a C^1 dynamic F defined on $\Delta\backslash\bar{M}$ with only continuous FPs is called to be *(zero free) consistently extendable* to $\bar{M}$ if there is a $\tilde{F}:\bar{M}\rightarrow T\Delta$ as above such that (a) $\tilde{F}|_{M_i^s}$ is consistently directed for all s and i and (b) there is exactly one $\alpha\in P(x)\backslash\partial M$ at any $x\in\partial M$ which is no FP such that $F^\alpha(x)$ is inward pointing.

Main Theorem *Suppose P is a partition of Δ as characterized above with $M^8=M^9=\emptyset$. Suppose $F:\Delta\rightarrow T\Delta$ with F being C^1 on open sets D_i, $i\in I_D$, and $F|_{\cup_i D_i}$ is consistently extendable to $\Delta\backslash\cup_i D_i$. If $F|_{\Delta\backslash\cup_i D_i}=\tilde{F}$ which is one consistent extension as characterized above, then F is an admissible dynamic, i.e. there is a solution $\varphi : [0, \infty)\times\Delta\rightarrow\Delta$ such that*

$$\varphi_{s+t}(x)=\varphi_s(\varphi_t(x)) \quad \forall x\in\Delta, \forall s, t\geq 0$$

and φ is continuous in (t, x) and right differentiable in t with $\dot{\varphi}_t^+(x)=F(\varphi_t(x))\forall x\in\Delta$, $t\geq 0$.

Proof: From the Fundamental Theorem we know that there is a unique maximal solution $\varphi.(x):(a_x, b_x)\rightarrow D_i$ for each $x\in D_i$, $i\in I_D$ with $a_x<0$, $b_x>0$ such that φ is a C^1 dynamic system with $\dot{\varphi}_t(x)=F(\varphi_t(x))$, $\forall x\in D_i$ and $t\in(a_x, b_x)$.

If F^i be inward pointing at some $x\in\partial D_i$, there is then a unique solution path $\varphi(\tilde{x})$ for some $\tilde{x}\in D_i$ such that $\varphi_{a\tilde{x}}(\tilde{x})$:$\lim_{t\rightarrow a\tilde{x}+}\varphi_t(\tilde{x})=x$ and $\dot{\varphi}_t^+(\tilde{x})|_{t=a\tilde{x}}=F^i(x)$. If there was another $\tilde{\varphi}\neq\varphi$ with $\lim_{t\rightarrow a_{x'}}\tilde{\varphi}(x')=x=\varphi_{a\tilde{x}}(\tilde{x})$ as well, where $\tilde{\varphi}.(x')$ is defined for the maximal interval $(a_{x'}, b_{x'})$, $x'\in D_i$, $a_{x'}>-\infty$, then there is a contradiction to the property of the solution of F on

[23] More precisely,

$$\tilde{F}(x) := \psi_*|_x^{-1}\left(\frac{\psi(x)-a_1}{a_2-a_1}\psi_*|_{x_{il}^s}(\tilde{F}(x_{il}^s))+\frac{a_2-\psi(x)}{a_2-a_1}\psi_*|_{x_{ir}^s}(\tilde{F}(x_{ir}^s))\right)$$

where $\psi : \mathring{M}_i^s\rightarrow(a_1, a_2)$, $\psi_*|_x : T\mathring{M}_i^s\rightarrow T(a_1, a_2)$ are diffeomorphisms. See Arnold (1973) for notations.

D_i being continuous in the initial condition (i.e. φ being continuous in x). Analogously, solutions for any $\tilde{F}|_{\mathring{M}_i^s}: \mathring{M}_i^s \to T\mathring{M}_i^s$ posses similar properties.

We are now about to show that $\varphi_{\cdot}(x): [0, \infty) \to \Delta$ exists uniquely for all $x \in \Delta$, which is continuous and $\dot{\varphi}_t^+(x)|_{t=0} = \mathrm{F}(x)$. In fact, if $b_x = \infty$, this is trivially true. If $b_x < \infty$ for some $x \in \alpha \in P(x) \backslash \partial M$, then $F^\alpha(\varphi_{b_x}(x))$ is necessarily outward pointing at $\varphi_{b_x}(x) \in \bar{M}$ w.r.t. α. By construction, there is exactly one $\beta \in P(x) \backslash \partial M$, $\beta \neq \alpha$, such that F^β is inward pointing w.r.t. β. Define $\varphi_{t+b_x}(x) = \varphi_t^\beta(\varphi_{b_x}(x))$ for all $t \in [0, b)$ where $[a, b)$ is the maximal interval of corresponding solution path in β. If $b = \infty$, the job is finished. Otherwise repeat this procedure.

Now, it remains to prove that the so constructed unique $\varphi : [0, \infty) \times \Delta \to \Delta$ is continuous in x in addition. This is trivially true for $x \in \Delta \backslash \bar{M}$. For $x \in \bar{M}$, which is not a FP, consider $\varphi_t(x_n^\alpha)$, $x_n^\alpha \xrightarrow{n \to \infty} x$, with $x_n^\alpha \in \alpha$ and $\alpha \in \mathrm{P}(x)$. Since only one $\beta \in P(x)$ has the property of F^β being inward while F^α being outward for all $\alpha \in P(x)$, $\alpha \neq \beta$, and since the solution φ is continuous in t, $\varphi_t(x_n^\alpha) \xrightarrow{n \to \infty} \varphi_t(x) \in \beta$ for all $\alpha \in P(x)$. This completes the proof.

The last point made in the proof is also exactly the reason why M^8, M^9 are excluded in the main theorem: with extension method used here, the above constructed solution path φ would no longer be continuous in x. Notice that even BR-dynamic may be not admissible in this sense. An important implication of the main result above is that the Theorem of Poincare-Bendixson can be generalized for admissible dynamics defined on some compact set as well.

Theorem (Generalized Poincare-Bendixson) *Suppose F is C^1 on $\mathring{\Delta}$ and admissible. Then, for any $x \in \Delta$, $L_\omega(x)$ contains no equilibrium implies $L_\omega(x)$ is a cycle.*

Proof: Suppose $\varphi_t(x) \in \mathring{\Delta}$ $\forall t$ large enough regardless whether $\varphi_t(x) \cap \partial\Delta = \emptyset$, then the original P-B-thoerem applies, a proof of which can be found in Hirsch and Smale (1974). If however there is a $y = \varphi_s(x) \in \partial\Delta$ for some s with $y \in \mathrm{L}_\omega$, consider the flow $\varphi_t(y)$ $t \geq 0$. If there is no $\tau > 0$ so that $\varphi_\tau(y) = y$, then $y \notin L_\omega(x)$ as $L_\omega(x) = \mathrm{L}_\omega(y)$. Hence, $L_\omega(x)$ is a cycle of period $\lambda = \min\{\tau : \varphi_\tau(y) = y\}$. If there is no such $y \in \varphi(x) \cap L_\omega(x) \cap \partial\Delta$, then there necessarily exists some $s > 0$ such that $\varphi_{s+t}(x) \in \mathring{\Delta}$ for all $t > 0$.

References

Amann, E. and C.-L. Yang

1994 "Sophistication and the Persistence of Cooperation", forthcoming in *Journal of Economic Behavior and Organization.*

Arnold, V.

1973 *Ordinary Differential Equations*, MIT Press.

Axelrod, R.

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

Banerjee, A. and Weibull, J. W.

1993 "Evolutionary Selection with Discriminating Players", discussion paper No. 375, Industriens Utredningsinstitut.

Binmore, K.

1992 *Fun and Games*, Lexington: DC Heath.

Binmore, K. and L. Samuelson

1992 "Evolutionary Stability in Repeated Games Played by Finite Automata", *Journal of Economic Theory*, 57:278-305.

Ellison, G.

1993 "Cooperation in the Prisoners' Dilemma with Anonymous Random Matching", discussion paper.

Frank, R. H.

1987 "If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience?", *American Economic Review*, 77:593-604.

1988 *Passions within Reasons. The Strategic Role of the Emotions.* New York: Norton.

Friedman, D.

1991 "Evolutionary Games in Economics", *Econometrica*, 59:637-666.

1992 "Economically Applicable Evolutionary Games", CentER discussion paper, No. 9226.

Fudenberg, D. and E. Maskin

1986 "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information", *Econometrica*, 54:533-556.

Gilboa, I. and A. Matsui

1991 "Social Stability and Equilibrium", *Econometrica*, 59:859-867.

Ghosh, P. and D. Ray

1997 "Cooperation in Community Interaction without Information Flow", *Review of Economic Studies.*

Hirsch, M. and S. Smale

1974 *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press.

Holländer, H.

1993 "The Demand for a Fair Share", Disc. paper, univ. Dortmund.

Kandori, M.

1992 "Social Norms and Community Enforcement", *Review of Economic Studies*, 59:63-80.

Kreps, D. M., P. Milgrom, J. Roberts and R. Wilson

1982 "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma", *Journal of Economic Theory*, 27:245-252.

Mailath, G. J.

1992 "Introduction: Symposium on Evolutionary Game Theory", *Journal of Economic Theory*, 57:259-277.

Maynard Smith, J.

1982 "Evolution and the Theory of Games", Cambridge University Press, Cambridge.

Milgrom, P.

1984 "Axelrod's 'The Evolution of Cooperation'", *Rand Journal of Economics*, 15: 305-309.

Milgrom, P., D. North and B. Weingast

1990 "The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs", *Economics and Politics*, 2:1-23.

North, D.

1991 "Institutions", *Journal of Economic Perspectives*, 5:97-112.

Robson, A. J.

1990 "Efficiency in evolutionary games: Darwin, Nash and the Secret Handshake", *Journal of Theoretical Biology*, 144:379-396.

Rosenthal, R.

1981 "Games of imperfect information, predatory pricing and the Chain-Store paradox", *Journal of Economic Theory*, 25:92-100.

1993 "Rules of thumb in games", *J. of Econ. Beh. and Org.*, 22:1-13.

智慧性變種與合作行爲之進化

楊春雷
中央研究院中山人文社會科學研究所

摘　要

經濟理論的一個中心命題是，社會如何解決在“囚徒難局”（prisoners' dilemma）下的合作問題。本篇文章將一個可稱爲“智慧性”（或“小心的”）鴿子的結構型行爲變種引進一基本進化模型，此新型的特異處在於它每次與人交易時都要投入精力以辨別其交易夥伴的本性以避免被剝削。我們可以證明，合作行爲可以以長期穩定的混員均衡點或是穩定周期的形式持續存在，且此結果適用於較大的一組滿足“可允性”及“可容性”條件的動態方程。值得注意的是，我們的結果要求所有合作與不合作方式都持續出現，這與現實觀察頗符。

關鍵詞：囚徒難局，謹慎的“鴿子”，動態下的穩定共生，合作長在