Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 98–108.
https://doi.org/10.23941/ejpe.v13i2.501

Do Kantians Drive Others to Extinction?

JEAN-FRANÇOIS LASLIER
CNRS, Paris School of Economics

I. INTRODUCTION

From Thomas Malthus and Pierre Verhulst to Alfred Lotka and Vito Volterra, theoretical biology has studied the dynamics of living species (see Berryman 1992 for an account of this history). The interaction between theoretical biology and game theory (Smith 1982) has also been fruitful, and—as a result of this interaction—a whole discipline of evolutionary game theory has emerged (Weibull 1995).

Of particular interest in evolutionary game theory is the explanation of cooperative behavior and altruism based on evolutionary arguments. This is because the existence of cooperation may at first sight seem to be in contradiction with the 'individual selection' paradigm in biology. But, as John Roemer recalls in his book, *How We Cooperate: A Theory of Kantian Optimization* (2019), men (and animals too) routinely behave in a cooperative manner, sometimes even at their own expense. This explains why the question of cooperation has been a non-trivial puzzle in evolutionary biology (see the work of Hamilton 1963, 1964, and, more recently, Nowak and Sigmund 2005, or Alger and Weibull 2013). It is also a central question in economic theory, all the more after issues of incentives and selfish behavior became prevalent in mainstream economics. In the eighth chapter of *How We Cooperate*, Roemer applies the solution concept of Kantian optimization to coordination games in order to offer an evolutionary view of this concept. This kind of Kantian optimization is to be contrasted with what Roemer calls 'Nash behavior'.

Coordination requires giving some attention to what others do, and this is of course one element of cooperation. Games where individuals may settle on a low-quality outcome while coordination on a better one is also possible are therefore interesting case studies for the study of co-

AUTHOR'S NOTE: I wish to warmly thank Vaios Koliofotis and Marina Uzunova who spotted and fixed an important mistake in the first version of this paper. All remaining errors are mine.

operative behavior, even if cooperation should not be reduced to efficient coordination. The point was introduced by Jean-Jacques Rousseau in his *Discourse on the Origin and Basis of Inequality among Men* ([1755] 1923). The *Discourse* presents an evolutionary perspective on cooperative behavior, albeit a pre-Darwinian and non-modern one. The well-known Stag Hunt game is discussed by Rousseau as an illustration of the fact that behavioral coordination requires a signaling device: some kind of 'language', but a 'language' that can be restricted to specific goals.

Roemer's analysis of the Stag Hunt game incorporates one important idea of modern evolutionary theory, namely the concept of evolutionary stable equilibrium (ESE). It also alludes to possible dynamics that sustain the stability of these equilibria. There are no explicit dynamics in Roemer's construction, but they are a key feature of evolutionary theories, as the word 'evolutionary' itself suggests. Such dynamics make no reference to Rousseau's idea of conceiving communication as a means to coordination but instead model some Darwinian selection process of the fittest.

In the modeling exercise Roemer performs in chapter eight of *How We Cooperate*, the non-Kantian agents are called "Nashers" (117). The word refers to an equilibrium concept, the Nash equilibrium, which, unlike ESE, has no associated dynamic process that would ensure some form of stability. The analysis in chapter eight, which compares the fate of Nashers and Kantians along their evolutionary dynamics, therefore needs to be spelled out more precisely. Additional insights can be provided by a double dynamics model that takes into account both the dynamics of selfish optimization, which might sustain Nash behavior, and the dynamics of selection or survival of Kantians. Below I explore in greater detail the differences between Roemer's model and the double dynamics model.

II. Roemer's Model

Although Roemer discusses symmetric coordination games in some generality, I will concentrate on one example: the so-called Stag Hunt game.

II.I. The Stag Hunt Game

The Stag Hunt game is commonly traced back to Rousseau who tells a story about how men gradually came to acquire the concept of mutual commitment:

> In this manner, men may have insensibly acquired some gross ideas of mutual undertakings, and of the advantages of fulfilling them: that is, just so far as their present and apparent interest was concerned: for they were perfect strangers to foresight, and were so far from trou-

> bling themselves about the distant future, that they hardly thought of the morrow. If a deer was to be taken, every one saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs.
>
> It is easy to understand that such intercourse would not require a language much more refined than that of rooks or monkeys, who associate together for much the same purpose. (Rousseau [1755] 1923, 209–210)[1]

A standard version of this game is the one described by Roemer on pages 29 and 128 of his book. It goes as follows. There are two hunters who belong to the same population. If they hunt separately, they will grab hares, and this is worth the normalized payoff 0. If they hunt the stag together, they will each earn the highest payoff: say, 1. Now, if only one hunts the stag, he will catch neither the stag, nor a hare, and will therefore earn a negative payoff: say, -1. Meanwhile, the other hunter, who is chasing hares alone, will catch more hares than he would if both players were hunting hares, and this yields a payoff 0.5. This is Roemer's (a, b) game (29) with $a = -1$ and $b = 0.5$ (see Table 1). This game is a classic of game theory; for instance, Brian Skyrms (2004) sees it as a fable that describes the key feature of the social contract, and Ken Binmore uses it to defend his claim that "fairness evolved as Nature's answer to the equilibrium selection problem in the human game of life" (2006, 11).

II.II. Kantians

It is clear that hunting the stag is the thing to do for efficiency reasons. That is to say, the outcome that obtains when both players hunt the stag is the unique Pareto optimal outcome. In the Stag Hunt game, under the

[1] In the original French, the exact quote reads:

> Voilà comment les hommes purent insensiblement acquérir quelque idée grossière des engagements mutuels, et de l'avantage de les remplir, mais seulement autant que pouvait l'exiger l'intérêt présent et sensible; car la prévoyance n'était rien pour eux, et loin de s'occuper d'un avenir éloigné, ils ne songeaient pas même au lendemain. S'agissait-il de prendre un cerf, chacun sentait bien qu'il devait pour cela garder fidèlement son poste; mais si un lièvre venait à passer à la portée de l'un d'eux, il ne faut pas douter qu'il ne le poursuivît sans scrupule, et qu'ayant atteint sa proie il ne se souciât fort peu de faire manquer la leur à ses compagnons.
>
> Il est aisé de comprendre qu'un pareil commerce n'exigeait pas un langage beaucoup plus raffiné que celui des corneilles ou des singes, qui s'attroupent à peu près de même.

	Stag	Hare
Stag	$(1, 1)$	$(-1, 0.5)$
Hare	$(0.5, -1)$	$(0, 0)$

Table 1: The Stag Hunt game in normal form.

unique Kantian equilibrium—the strategy profile where strategies answer the question “what is the strategy I would like both of us to play?” (12)—both players hunt the stag because each player is better off when both hunt the stag than when both hunt hares. The key concept of Roemer’s analysis gives a very clear verdict in this game: any Kantian player hunts the stag.

II.III. Nashers

The game has two pure strategy Nash equilibria:

(1) Both players hunting the stag is a strict Nash equilibrium because, in that case, hunting the stag yields a payoff of 1 while chasing hares alone yields a payoff of only 0.5.

(2) Both players hunting hares is a strict Nash equilibrium because, in that case, hunting hares yields a payoff of 0 while chasing a stag alone yields a payoff of only -1.

Following Harsanyi and Selten (1988), the stag-hunting equilibrium is called the *payoff dominant* equilibrium while the hare-hunting equilibrium is called the *risk dominant* equilibrium. The game also has a mixed strategy Nash equilibrium:

(3) Both players deciding at random and independently to hunt the stag with a probability of 2/3 and to hunt hares with a probability of 1/3 is a Nash equilibrium because if one player uses this mixed strategy, then the other player’s average payoff remains the same whatever he does. (The second player thus has no strict incentive to choose one strategy rather than the other, so, under the usual hypothesis about choice under risk, he can as well randomize in any way, so they might as well randomize in the same way.)

Roemer’s definition of a ‘Nasher’ is not perfectly clear as a general definition. The word is used as a short-hand for the expression “Nash optimizer” (117), but what it means to be a ‘Nash optimizer’ depends upon a given Nash equilibrium—it is not determined by the game itself or the players’ strategies. Roemer writes that “If there are several Nash equilibria, a Nasher randomizes among them” (118). This is difficult to follow:

it is unclear whether this means that different Nashers end up playing different strategies, or that they manage to correlate their randomization so that everyone plays the same (randomly chosen) pure strategy. Moreover, in some instances, as in the game above, each one of the two pure strategies is played in a Nash equilibrium, but choosing at random does not, in general, result in an equilibrium. In fact, except for a very specific randomization scheme, mixed strategies are almost never best responses and are therefore almost never chosen by an optimizer.

Therefore, in order to understand what a 'Nasher' is, one has to look closely at how the concept is used.

II.IV. Roemer's Evolutionary Argument

The argument that leads to the conclusion that "Kantians drive Nashers to extinction" (125) is provided in the proof of Proposition 8.4 (121–122). The proof first fixes the Nash equilibrium under consideration and denotes the associated strategy by q^*. (Thus Nashers do not randomize among different equilibria.) A Nasher hunts the stag with probability q^* and hares with the complement probability $1 - q^*$. Roemer considers two cases: (1) $q^* = 0$ (hunting hares), and (2) the mixed equilibrium $q^* = q_1^*$ that has, in this game, the player hunting the stag with probability $q_1^* = 2/3$ and hares with probability $1/3$, yielding an average payoff of $1/3$. According to Roemer, the third case (Nashers hunting the stag) is not to be considered because in that case Nashers and Kantians cannot be distinguished.

Following the standard evolutionary model, one imagines that individuals are randomly matched in pairs (with no 'assortative matching'). Let ν be the proportion of Kantians in the whole population. Then, the average payoff of a Kantian ($V^K(\nu)$) and of a Nasher ($V^N(\nu)$) in the two cases above can be computed as follows:

(1) Nashers hunt hares ($q^* = 0$):

$$\begin{array}{rcccccc} V^K(\nu) & = & \nu \cdot 1 & + & (1-\nu)\cdot(-1) & = & 2\nu - 1 \\ V^N(\nu) & = & \nu \cdot \frac{1}{2} & + & (1-\nu)\cdot 0 & = & \frac{\nu}{2} \end{array}$$

(2) Nashers use the mixed strategy $q^* = q_1^* = 2/3$:

$$\begin{array}{rcccccc} V^K(\nu) & = & \nu \cdot 1 & + & (1-\nu)\cdot\left(\frac{2}{3} - \frac{1}{3}\right) & = & \frac{2+4\nu}{6} \\ V^N(\nu) & = & \nu \cdot (\frac{2}{3} + \frac{1}{3}\cdot\frac{1}{2}) & + & (1-\nu)\cdot\frac{1}{3} & = & \frac{2+3\nu}{6} \end{array}$$

Now, it is clear that, in the case where Nashers use the mixed strategy q_1^*, Kantians have an evolutionary advantage because their payoff is larger than the payoff of Nashers: $V^K(v) = {}^{(2+4v)}/_6 \geq {}^{(2+3v)}/_6 = V^N(v)$.

When Nashers hunt hares, the advantage will be on the side of Nashers or Kantians depending on the value of v: Kantians have an advantage if and only if v is large enough ($v \geq {}^2/_3$). In other words, if Nashers hunt hares but most players hunt the stag, Nashers earn less, on average, than the other agents.

This is Roemer's argument for the claim that Kantians drive Nashers to extinction.

III. A Double Dynamics Model

In the argument above, there is only a sketch of the evolutionary analysis that is necessary for a convincing evolutionary argument. In particular, 'Nashers' are neither optimizing agents (as they should be in an economic model of rational behavior), nor adapting agents (as they should be in a behavioral model of learning), nor evolving agents (as they should be in a biological model of Darwinian selection)—they are just stubborn hare hunters in one case, and (quite strangely) stubborn users of a specific mixed strategy in the other case. I now propose a standard model of a replicator-dynamics type in order to study, for this game, the evolution of a population that consists of Kantian individuals (in Roemer's sense) and of adaptive individuals. I will simply call the non-Kantian agents 'selfish', although one could think of many names for them.

III.I. Replicator Dynamics

The following explication uses the most standard mathematical model of evolution called the replicator dynamics. So I begin with a very brief presentation of this model.

First, the idea of *fitness* defines the number of offspring of some replicating unit as a function of its environment, so that a population of size n characterized by a fitness per individual f grows at the rate f. In discrete time, the population of size n will be of size $f \cdot n$ at the next generation, and, in continuous time, the time derivative of n is $\dot{n} = {}^{\mathrm{d}n}/_{\mathrm{d}t} = f \cdot n$ with f being now a replication rate by unit of time.

With two groups $i = 1, 2$ of size n_i and fitness f_i each, writing $n = n_1 + n_2$ and $x_i = {}^{n_i}/_n$, one gets in full generality:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{n_i}{n}\right) = \frac{\dot{n}_i n_j - \dot{n}_j n_i}{n^2}$$

That is:

$$\dot{x}_i = x_i x_j \cdot (f_i - f_j)$$

Interestingly, this differential equation, which generates replicator dynamics, appears in like form in different models that describe (i) population genetics, (ii) social imitation, or (iii) individual adaptive learning (see Laslier, Umbhauer, and Walliser 2006). We will now apply this idea to obtain an evolutionary model for Roemer's argument.[2]

In the absence of Kantians, the standard evolutionary analysis of the family of Stag Hunt games indicates that a pure strategy equilibrium is reached in the long run: if the initial composition of the population contains a sufficient number of individuals of one type (be it stag hunters or hare hunters), coordination on this type will occur in the long run. The mixed strategy equilibrium is, on the contrary, unstable and is not reached. To introduce Kantian players, I propose the following model.

Let $v(t)$ be the proportion of Kantians in the population at time t. This proportion will vary with time. Following the evolutionary paradigm, non-Kantians will not be assumed to jump directly to some optimal or 'Nash' strategy, but they will adjust their strategies gradually with time. So let $x(t)$ denote the proportion of stag hunters among the non-Kantians. In the whole population the proportion of stag hunters is therefore:

$$y = v + (1 - v)x$$

Two processes of evolution are coupled: within non-Kantians for their choice of strategy, and between Kantians and non-Kantians. The two processes may occur at different speeds.[3] For instance, one may wish to study the case where selfish individuals can change strategy relatively quickly while it is only at a slow pace that selfish individuals become Kantians or Kantians become selfish. This is rather natural: it means that selfish individuals adjust their behavior by choosing a best response to the circumstances, if not instantly, at least relatively quickly. Roemer's definition of a Nasher does not presuppose a dynamic adjustment process—the underlying assumption is that Nashers find best responses instantly. Therefore, in order to relax this assumption, I will consider the case where this pro-

[2] Note that, in evolutionary game theory, an important literature exists, which deals with the stag-hunt problem and with extensions of the basic game (Kandori, Mailath, and Rob 1993; Samuelson 1997). The main focus in this literature is on the question of communication: does language help to coordinate on the payoff dominant equilibrium? Can one even explain the emergence of language as a coordinating device that allows forward induction in stag-hunt situations? See chapter eight in Samuelson (1997).

[3] The same idea of a two-level dynamic process is used in Laslier and Öztürk Göktuna (2016), and in Öztürk Göktuna (2019).

cess is not instantaneous but relatively faster than the transition from selfish to non-selfish behavior. In my model, individuals can be viewed from the perspective of two time scales. In the long-run or evolutionary time, it takes many generations to converge on a long-run equilibrium in accord with the dynamic process of natural selection (a slow process of change). Being a Kantian or a selfish optimizer is determined by this long-term evolution. In the short-run or decision-making time, rational individuals make choices or acquire new strategies via social learning (a fast process of change). A selfish optimizer chooses her strategy in this shorter term. Hence, the overall evolutionary process has the complex structure of a slow evolutionary and a fast decision-making time horizon.

Among selfish individuals, the difference in payoff between those who hunt the stag and those who hunt hares is:

$$\delta^1 = [y - (1-y)] - \left[\frac{y}{2}\right] = \frac{3y}{2} - 1$$

Let s be the adaptive speed of the selfish individuals. The replicator dynamics within this group is described by the following differential equation (where $\dot{x}$ denotes the time derivative ${}^{\mathrm{d}x}/_{\mathrm{d}t}$):

$$\dot{x} = s \cdot x(1-x) \cdot \delta^1 \tag{1}$$

At the level of the whole population, the difference in payoff between Kantians and selfish individuals is obtained as follows: for the Kantians, since they all play the same strategy (stag), the average payoff is simply the average payoff of the stag strategy, that is: $2y - 1$. For the non-Kantian group, one should think of them as carriers of a ‘selfish’ gene whose fitness is the average fitness of the individuals who carry it. Therefore, the relevant payoff for the evolution of the selfish population is the average payoff in this population. That is:

$$x \cdot (2y-1) + (1-x) \cdot \frac{y}{2}$$

Hence:

$$\delta^2 = [2y-1] - \left[x(2y-1) + (1-x)\frac{y}{2}\right] = (1-x)\left(\frac{3y}{2} - 1\right)$$

If the adaptive speed of Kantianism is normalized to 1, then the associated replicator dynamics is described by the following differential equation:

$$\dot{v} = v(1-v) \cdot \delta^2 \tag{2}$$

The quantities y, δ^1, and δ^2 depend solely on the variables v and x, so equations (1) and (2) define a system of differential equations in the square $(v, x) \in [0,1] \times [0,1]$.

III.II. Results

Figure 1 collects all results. The left panel of the figure is drawn for a speed ($s = 5$) such that selfish individuals adapt their strategy relatively quickly. This is the most natural assumption. On the horizontal axis is the proportion of Kantians, v, and on the vertical axis is the proportion of stag hunters among the population of selfish individuals, x. The graph represents the flow of the differential system.

The lower left corner, $(0,0)$, corresponds to the situation where there are no Kantians and everyone is hunting hares. The upper right corner, $(1,1)$, corresponds to the situation where the whole population consists of Kantians and everyone hunts stags. The upper segment, where $x = 1$, also describes situations of full cooperation, where everyone hunts stags but some do it because they are Kantians and others do it for selfish reasons.

Following the arrows, one can observe the fate of the system. The flow is divided in two: a lower left region that points to $(0,0)$, and an upper right region that always reaches the upper segment (where $x = 1$). Changing the speed, as depicted in the right panel of Figure 1, confirms this point.

The two basins of attraction are separated by the curve of the equation $\dot{v} = 0$. That is, ${}^{3y}/_{2} = 1$ or, in terms of v and x, $3v + 3(1-v)x = 2$. This is part of the hyperbola $x = {}^{(2-3v)}/_{3(1-v)}$ and is independent of s.

IV. Conclusion

In section 8.2 of *How We Cooperate*, after studying the Stag Hunt game in isolation, Roemer considers several games. He concludes that if Nature chooses at random what kind of a game is played—either a coordination game or a Prisoner's Dilemma—then Nashers and Kantians can co-exist. Instead, this note focused on a single coordination game.

Faced with the claim that "In games of pure coordination, Kantians drive Nashers to extinction" (125), readers of Roemer's book might be tempted to over-interpret the expression 'Nasher'. They may thus conclude that, in coordination games, Kantian optimizers (in the sense of Laffont 1975, and Roemer 2019) have some efficiency advantage that makes them fitter, from an evolutionary point of view, than selfish optimizers. This is not true. In the Stag Hunt game, either Kantians are wiped away by

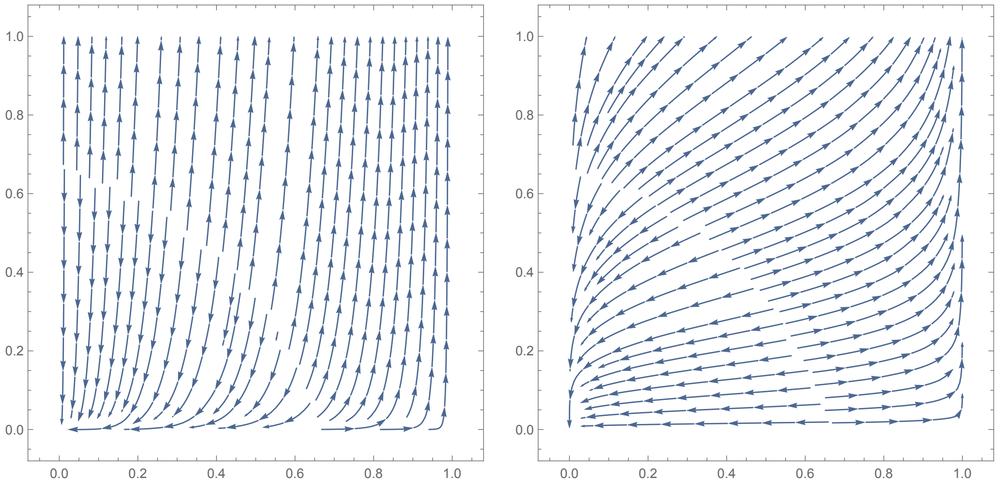

Figure 1: Stag Hunt game. Coupled dynamics for the proportion of Kantians (horizontal axis) and the proportion of cooperators among non-Kantians (vertical axis). Left panel: $s = 5$; right panel: $s = 0.2$.

selfish individuals who do not cooperate, and thus Kantians are 'driven to extinction' by the selfish optimizers, or both remain as some fraction of the population.

Focusing exclusively on 'equilibria' to describe and analyze collective outcomes may be misleading. Following his analysis, Roemer writes that "According to Proposition 8.4, there are no (a, b) games where both Kantian and Nash players exist with positive frequencies in a stable equilibrium" (123). It is not clear what is meant here by 'stable equilibrium' but, as I showed above, the natural process that sustains evolutionary stability leads to, depending on the initial conditions, two possible outcomes. One possibility is that hare hunters (who can be called Nashers) drive Kantian stag hunters to extinction. The other possibility is that Kantians and selfish optimizers (who can also be called Nashers) co-exit, all hunting stags but for different reasons.

References

Alger, Ingela, and Jörgen W. Weibull. 2013. "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching." *Econometrica* 81 (6): 2269–2302.

Berryman, Alan A. 1992. "The Origins and Evolution of Predator-Prey Theory." *Ecology* 73 (5): 1530–1535.

Binmore, Ken. 2006. "The Origins of Fair Play." Papers on Economics and Evolution Working Paper No. 0614. Max Planck Institute of Economics, Jena.

Hamilton, William D. 1963. "The Evolution of Altruistic Behavior." *The American Naturalist* 97 (896): 354–356.

Hamilton, William D. 1964. "The Genetical Evolution of Social Behaviour. I and II." *Journal of Theoretical Biology* 7 (1): 1–52.

Harsanyi, John C., and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games.* Cambridge, MA: The MIT Press.

Kandori, Michihiro, George J. Mailath, and Rafael Rob. 1993. "Learning, Mutation, and Long Run Equilibria in Games." *Econometrica* 61 (1): 29–56.

Laffont, Jean-Jacques. 1975. "Macroeconomic Constraints, Economic Efficiency and Ethics: An Introduction to Kantian Economics." *Economica* 42 (168): 430–437.

Laslier, Jean-François, and Bilge Öztürk Göktuna. 2016. "Opportunist Politicians and the Evolution of Electoral Competition." *Journal of Evolutionary Economics* 26 (2): 381–406.

Laslier, Jean-François, Gisèle Umbhauer, and Bernard Walliser. 2006. "Game Situations." In *Evolutionary Microeconomics*, edited by Jacques Lesourne, André Orléan, and Bernard Walliser, 67–112. Heidelberg: Springer.

Nowak, Martin A., and Karl Sigmund. 2005. "Evolution of Indirect Reciprocity." *Nature* 437 (7063): 1291–1298.

Öztürk Göktuna, Bilge. 2019. "A Dynamic Model of Party Membership and Ideologies." *Journal of Theoretical Politics* 31 (2): 209–243.

Roemer, John E. 2019. *How We Cooperate: A Theory of Kantian Optimization.* New Haven, CT: Yale University Press.

Rousseau, Jean-Jacques. (1755) 1923. "Discourse on the Origin and Basis of Inequality among Men." In *The Social Contract and Discourses by Jean-Jacques Rousseau*, translated by George D. H. Cole, 155–246. London: J. M. Dent & Sons.

Samuelson, Larry. 1997. *Evolutionary Games and Equilibrium Selection.* Cambridge, MA: The MIT Press.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure.* Cambridge: Cambridge University Press.

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

Weibull, Jörgen. 1995. *Evolutionary Game Theory.* Cambridge, MA: The MIT Press.

Jean-François Laslier is a member of the CNRS (French National Centre for Scientific Research) and a professor at the Paris School of Economics. His interests include mathematical economics, games, social choice, and political science. He conducts research on democracy, and, in particular, on electoral systems and voting behavior, from the formal and the experimental points of view. He publishes in the two fields of economics and political science.
Contact e-mail: <jean-francois.laslier@ens.fr>