Robustness in Signaling Games

Simon M. Huttegger[†][‡]

The spontaneous emergence of signaling has already been studied in terms of standard evolutionary dynamics of signaling games. Standard evolutionary dynamics is given by the replicator equations. Thus, it is not clear whether the results for standard evolutionary dynamics depend crucially on the functional form of the replicator equations. In this paper I show that the basic results for the replicator dynamics of signaling games carry over to a number of other evolutionary dynamics.

1. Introduction. Various kinds of social behavior have been explained by evolutionary game theoretic models. Such models usually remain agnostic about many details of the phenomenon under consideration, for example, the mechanisms that might produce a specific behavior. Not specifying details leaves evolutionary game theoretic models open to a number of criticisms. In particular, one might argue that including some of these details might result in a dynamic which is considerably different from the dynamic of the less detailed model.

Appealing to robustness may sometimes help to escape such criticisms. A result of some evolutionary game theoretic model, like the emergence of a certain social behavior, is robust relative to particular changes if it continues to hold in models which resemble the original one except in those changes. If some result is robust in this sense, then certain details of the original model don't matter.

At one level, this paper may be regarded as an exercise in providing mathematically sound arguments for a specific kind of robustness: robustness with respect to qualitative changes in the evolutionary dynamics. Such changes generate different classes of dynamical systems which share certain features. At another level, this paper may be seen as a contribution to research on the evolution of simple communication systems. The game

[†]To contact the author, please write to: Konrad Lorenz Institute for Evolution and Cognition Research, Adolf Lorenz Gasse 2, A-3422 Altenberg, Austria; e-mail: simon.huttegger@kli.ac.at.

[‡]This research was supported by the Konrad Lorenz Institute for Evolution and Cognition Research.

Philosophy of Science, 74 (December 2007) pp. 839–847. 0031-8248/2007/7405-0024$10.00

https://doi.org/10.1086/525625 Published online by Cambridge University Press

I will study is a model of social communication which was introduced by Lewis (1969). I will first review the results on the standard evolutionary dynamics of this game (Section 2). These results show that standard evolutionary dynamics is quite likely to lead to states of partial communication. But it does not always lead to states of perfect communication. In Sections 3 and 4 I will show that the results for the replicator dynamics basically carry over to some general classes of evolutionary dynamics.

2. Signaling Games and Standard Evolutionary Dynamics. A Lewis signaling game is based on three sets with n elements, where n is an arbitrary finite number: a set of world states $S = \{\sigma_1, \ldots, \sigma_n\}$, a set of messages $M = \{m_1, \ldots, m_n\}$ and a set of possible acts $A = \{\alpha_1, \ldots, \alpha_n\}$. For any i, act α_i is the right response to state σ_i. It is the wrong response to any other state. Moreover, it is assumed that there are two players. A sender observes the state of the world and may choose one of n messages from the set M. A receiver, who is, for whatever reasons, incapable of observing the state of the world, may choose an act after she has received the sender's signal. If we assume that the players get the same payoff for each outcome,[1] then a simple signaling game Σ_n may be defined as a triple $\langle I, \{S_i\}_{i \in I}, \{u_i\}_{i \in I}\rangle$. $I = \{1, 2\}$ is the set of players: the sender, 1, and the receiver, 2. $S_1 = \{s_k | s_k$ is a function from S to $M\}$ is the set of sender strategies. $S_2 = \{a_l | a_l$ is a function from M to $A\}$ is the set of receiver strategies. And $u_i : S_1 \times S_2 \to \mathbb{R}$ are the payoff functions. Let $u_i = u$ and

$$u(s_k, r_l) = \sum_{j=1}^{n} \mathbb{P}(\sigma_j) \cdot u^*(\sigma_j, (r_l \circ s_k)(\sigma_j)).$$

Here, $\mathbb{P}$ is a probability distribution over S, $\circ$ is the operation of function composition and $u^* : S \times A \to \{0, 1\}$ such that $u(\sigma_i, \alpha_j) = \delta_{ij}$ (δ_{ij} being the Kronecker symbol: $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$). The computation of the players' payoffs implements the assumption of complete common interest between the players. If some state of the world obtains, they both get a positive payoff just in case the right act is chosen by the receiver. Figure 1 shows an extensive form representation of a simple signaling game.

Some combinations of sender strategies and receiver strategies allow perfect communication. They are called *signaling systems* in Lewis 1969. A strategy combination (s_k, r_l) is a signaling system if the composition $s_k \circ r_l$ maps σ_i on α_i, for each i. Equivalently one may say that a signaling system guarantees the maximum payoff of 1 to both players regardless of the state of the world. A signaling system determines the meaning of

[1] This assumption expresses complete common interest between the players.

https://doi.org/10.1086/525625 Published online by Cambridge University Press

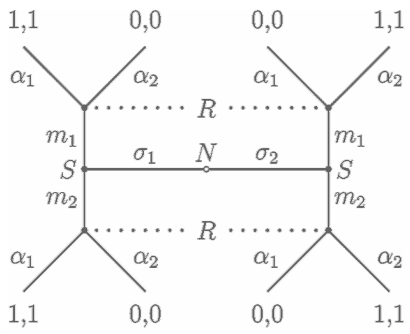

Figure 1. Extensive form representation of a simple signaling game. There are two states, two acts and two messages. Nature, N, decides which state occurs. The sender, S, chooses between sending message m_1 or sending message m_2. The receiver, R, does not know which state has occurred (indicated by the dotted line). R chooses act α_1 or act α_2.

signals. That is to say, in a signaling system players use the signals in such a way as to allow information to be transmitted. Since, for $n \geq 2$, there is always more than one signaling system, meaning is conventional.

Signaling games have already been studied in terms of evolutionary game theory by looking at the replicator dynamics (see Skyrms 1996, 2000 and Huttegger 2007). Let Σ_n^r be the two-player role conditioned game based on the signaling game Σ_n. (See Cressman 2003 for details on role conditioned games.) That is, a strategy of Σ_n^r is a pair of strategies (s, r) where s is a sender strategy and r is a receiver strategy of Σ_n. It is assumed that each player of Σ_n^r is sender (receiver) with probability 1/2. This guarantees that the payoff matrix of Σ_n^r is symmetric. The payoff for each player, and each outcome, of the role conditioned game may then be obtained by computing expected values. A signaling system type s is a pair of strategies which constitute a signaling system of Σ_n. The $\phi(n) = n^{2n}$ strategies of Σ_n^r may be thought of as types of individuals in a population. If $\Delta^{\phi(n)}$ denotes the simplex in $\mathbb{R}^{\phi(n)}$,[2] then the state of the population may be described by the proportion of those types. The replicator dynamics determines the growth rate of each type i given the current population state $\mathbf{x}$ in terms of success relative to the current population average:

$$\dot{x}_i = x_i(u(x_i, \mathbf{x}) - u(\mathbf{x}, \mathbf{x})), \quad i = 1, \ldots, k, \tag{1}$$

where $u(x_i, \mathbf{x})$ is the expected payoff for type i and $u(\mathbf{x}, \mathbf{x})$ is the average

[2] The simplex in $\mathbb{R}^k$ is the $k-1$-dimensional manifold given by $\Delta^k = \{\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k : \sum_i x_i = 1)\}$.

https://doi.org/10.1086/525625 Published online by Cambridge University Press

payoff in the population. (1) may be thought of as a model for cultural evolution or as a model for biological evolution. Before we proceed, recall the following concepts from the theory of dynamical systems. A point $\mathbf{x} \in \Delta^k$ is a rest point if $\dot{\mathbf{x}} = 0$. That is, the population is at a rest point when its configuration does not change anymore. A point $\mathbf{x} \in \Delta^k$ is stable if solutions starting near $\mathbf{x}$ stay nearby. It is asymptotically stable if there is a neighborhood U of $\mathbf{x}$ such that solutions starting at $\mathbf{y} \in U$ converge to $\mathbf{x}$. $\mathbf{x}$ is unstable if it is not stable.

The replicator dynamics of signaling games has been studied in Huttegger 2007. The main results are summarized in the following theorem. For a number of additional results, see Pawlowitsch 2006. Before stating Theorem 1, let me explain two concepts used in its statement. The interior of $\Delta^{\phi(n)}$ is the part of $\Delta^{\phi(n)}$ where all types have positive relative frequency. The boundary of $\Delta^{\phi(n)}$ is the part of $\Delta^{\phi(n)}$ where at least one type has zero relative frequency.

Theorem 1. *Let* Σ_n^r be a symmetrized simple signaling game. Then the following statements are true:

1. Denote the set of points in the interior of $\Delta^{\phi(n)}$ which do not converge to the boundary of $\Delta^{\phi(n)}$ by S. Then S has Lebesgue measure zero.
2. A state $\mathbf{p}^* \in \Delta^{\phi(n)}$ is asymptotically stable if and only if $\mathbf{p}^*$ is a signaling system type.
3. Denote by W the set of solutions which do not converge to a signaling system. Then W has Lebesgue measure zero if and only if $n = 2$ and $\mathbb{P}(\sigma_1) = \mathbb{P}(\sigma_2)$.

Suppose we are given a probability distribution over $\mathbb{R}^{\phi(n)}$ which is absolutely continuous with respect to Lebesgue measure for $\mathbb{R}^{\phi(n)}$. Theorem 1 tells us that the replicator dynamics will with probability 1 carry the population to some state where not all types are present. Thus, some degree of coherence for communication is achieved almost surely. But, although signaling systems are the only asymptotically stable states, there is a positive probability of not reaching them. This is expressed by the third part of Theorem 1. If at least one of the conditions of this statement fails to hold, then there exist connected components of rest points on the boundary which attract a set of positive measure from the interior of state space. It can be shown that these connected sets of rest points are not attractors. This means that there exists no neighborhood U such that all states in U converge to the connected set of rest points. Some states on the boundary of these connected components are unstable. Hence, signaling systems are the only states which are stable relative to selection and relative to neutral drift.

These results leave open a number of interesting questions. For in-

https://doi.org/10.1086/525625 Published online by Cambridge University Press

stance, does the evolution of perfect, or nearly perfect, communication systems get more likely if we add certain features to the replicator dynamics (such as mutation or correlated encounters between individuals)? If we assume a reasonably high number of signals n, the evolutionary dynamics might spend a very long time in states of partial communication which are far from optimal (even if we take into account neutral drift). Numerical simulations suggest that these states are not observed under replicator-mutation dynamics.

The robustness of the results stated in Theorem 1 is a related issue. Do these results depend on the specific functional form of the replicator equations (1)? To answer this question, we will first look at a quite large class of evolutionary dynamics called *payoff monotonic*, which contains the replicator dynamics.[3] We can get still more general results when we study *adjustment dynamics*. This is a class of games which contains all payoff monotonic dynamics. (See Weibull 1995, 144–148, and Hofbauer and Sigmund 1998, Section 8, for more information on payoff monotonic and adjustment dynamics).[4]

3. Payoff Monotonic Dynamics. Consider a dynamics of a simple signaling game Σ_n^r on the simplex $\Delta^{\phi(n)}$ given by the system of differential equations

$$\dot{x}_i = x_i g_i(\mathbf{x}). \tag{2}$$

The functions $g_i : \Delta^{\phi(n)} \to \mathbb{R}$ are assumed to be continuously differentiable. This guarantees the existence and uniqueness of solutions (see, e.g., Hirsch and Smale 1974). Moreover, it is assumed that $\sum_i x_i g_i(\mathbf{x}) = 0$. This implies that the overall growth rate of the frequencies x_i is constant. The frequency of one type can only increase if the frequency of other types decreases. As a consequence, $\Delta^{\phi(n)}$ and its boundary faces are invariant.

A game dynamics (2) is said to be *payoff monotonic* if and only if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \Longleftrightarrow u(x_i, \mathbf{x}) > u(x_j, \mathbf{x}). \tag{3}$$

Thus, a payoff monotonic dynamics is characterized by the property that the proportion of types with a higher payoff grows at a higher rate than the proportion of types with a lower payoff. This is a reasonable assumption for any dynamics for which the relative payoffs are assumed to influence the evolution of types.

The replicator dynamics is clearly payoff monotonic. In this case we

[3] Similar classes of dynamics have already been studied for a signaling mini-game in Skyrms 2000.

[4] Another kind of robustness concerns structural stability, i.e., small perturbations of the differential equations in function space. See D'Arms et al. 1998 and Skyrms 2000.

https://doi.org/10.1086/525625 Published online by Cambridge University Press

even have $g_i(\mathbf{x}) - g_j(\mathbf{x}) = u(x_i, \mathbf{x}) - u(x_j, \mathbf{x})$. Other important examples of payoff monotonic dynamics include different kinds of imitation dynamics (see Hofbauer and Sigmund 1998 for more). There is a close relationship between payoff monotonic dynamics and the replicator equations (for a proof see, e.g., Weibull 1995, 147).

Theorem 2. $\mathbf{p}^*$ is a rest point for (1) if and only if $\mathbf{p}^*$ is a rest point of a payoff monotonic dynamics (2).

This does not imply, however, that the stability properties of $\mathbf{p}^*$ will be the same under any payoff monotonic dynamics. The next proposition implies that, for simple signaling games, some stability results for rest points of the replicator dynamics indeed carry over to payoff monotonic dynamics. It shows that the average payoff $u(\mathbf{x}, \mathbf{x})$ is a global strict Liapunov function for the systems under consideration. This means that $u(\mathbf{x}, \mathbf{x})$ is strictly increasing along non-stationary solutions and constant on connected components of rest points. The significance of this result lies in the fact that $u(\mathbf{x}, \mathbf{x})$ is also a strict Liapunov function for the replicator dynamics of signaling games. (Indeed, it is even a potential for the replicator dynamics of signaling games [Huttegger 2007]. For more information on Liapunov functions and potential functions see Hirsch and Smale 1974.)

Theorem 3. $u(\mathbf{x}, \mathbf{x})$ is monotonically increasing along every nonstationary solution and is constant on every connected set of stationary states for any payoff monotonic dynamics (2) of Σ_n^r.

Proof. The payoff matrix A for Σ_n^r is symmetric. This is shown, for example, in Huttegger 2007. The average payoff in the population is $u(\mathbf{x}, \mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$ (where $\cdot$ denotes the dot-product). The symmetry of A yields

$$\dot{u}(\mathbf{x}, \mathbf{x}) = \dot{\mathbf{x}} \cdot A\mathbf{x} + \mathbf{x} \cdot A\dot{\mathbf{x}} = 2\dot{\mathbf{x}} \cdot A\mathbf{x} = 2\sum_i \dot{x}_i u(x_i, \mathbf{x}).$$

Inserting (2), we get

$$\dot{u}(\mathbf{x}, \mathbf{x}) = 2\sum_i x_i g_i(\mathbf{x}) u(x_i, \mathbf{x}).$$

Suppose the solution starting at $\mathbf{x}$ is not stationary. Then, since the stationary states for payoff monotonic dynamics coincide with the stationary states for the replicator dynamics, there exists a j such that $u(x_j, \mathbf{x}) \leq u(x_k, \mathbf{x})$ for all k with a strict inequality holding for at least one k. Hence

$$\dot{u}(\mathbf{x}, \mathbf{x}) = 2\sum_i x_i g_i(\mathbf{x}) u(x_i, \mathbf{x}) > 2u(x_j, \mathbf{x}) \sum_i x_i g_i(\mathbf{x}) = 0.$$

https://doi.org/10.1086/525625 Published online by Cambridge University Press

The last equality follows from the second constraint on payoff monotonic dynamics. Thus, $u(\mathbf{x}, \mathbf{x})$ is monotonically increasing along every nonstationary solution. If $\mathbf{x}$ is a stationary state, then $x_i g_i(\mathbf{x}) = 0$ for all i. Hence $\dot{u}(\mathbf{x}, \mathbf{x}) = 0$ and u is constant. ■

Theorem 3, together with the fact that $u(\mathbf{x}, \mathbf{x})$ is a potential for the replicator dynamics of Σ_n^r, allows us to draw some conclusions about the stability properties of rest points for Σ_n^r under payoff monotonic dynamics. Let us first consider interior rest points. If $\mathbf{p}^*$ is an interior rest point, then, by Theorem 1, every neighborhood contains almost exclusively solutions which tend away from $\mathbf{p}^*$. Since $u(\mathbf{x}, \mathbf{x})$ is increasing along these non-stationary solutions, the same holds for any payoff monotonic dynamics.

Moreover, it is quite obvious that signaling system types s continue to be asymptotically stable for payoff monotonic dynamics. Since s attracts all nearby solutions, s locally maximizes $u(\mathbf{x}, \mathbf{x})$. Hence, s will attract nearby solutions under any payoff monotonic dynamics. That is to say, if a trajectory starting at $\mathbf{x}$ converges to a signaling system type s for the replicator dynamics of Σ_n^r, then the trajectory starting at $\mathbf{x}$ converges to s for any payoff monotonic dynamics (2) of Σ_n^r.

Thus we may conclude that, although trajectories of the replicator dynamics and trajectories of some payoff monotonic dynamics will in general be different, the qualitative behavior of trajectories close to interior rest points and signaling systems of Σ_n^r will be the same for any of these dynamics. The analysis becomes more difficult when we study rest points on the boundary which do not correspond to signaling systems. If $u(\mathbf{x}, \mathbf{x})$ is higher in the interior of the state space close to such rest points on the boundary, then they are unstable under any payoff monotonic dynamics. But, as is shown in Huttegger 2007 and Pawlowitsch 2006, there exist connected components of rest points which attract a set of points with positive measure from the interior. Thus, close to such a connected component W of rest points $u(\mathbf{x}, \mathbf{x})$ is lower than on W. On the other hand, there exist points $\mathbf{p}$ on the boundary of W which are second order unstable. Hence, in every neighborhood of $\mathbf{p}$ there exist $\mathbf{x}, \mathbf{y}$ such that $u(\mathbf{x}, \mathbf{x}) < u(\mathbf{p}, \mathbf{p})$ and $u(\mathbf{y}, \mathbf{y}) > u(\mathbf{p}, \mathbf{p})$. Thus, it is possible for some payoff monotonic dynamics to be such that although orbits tend toward W due to increasing average payoff, they also turn outward toward boundary points, where average payoff is increasing away from W.

4. Adjustment Dynamics. The results presented in the preceding section can be improved by studying another class of dynamics called *adaptive dynamics*. Adaptive dynamics were introduced by Swinkels (1993). See also Hofbauer and Sigmund 1998. Consider the requirement that a pop-

https://doi.org/10.1086/525625 Published online by Cambridge University Press

ulation moves toward a better reply relative to the current state. This means that $\mathbf{x}(t+h) \cdot A\mathbf{x}(t) > \mathbf{x}(t) \cdot A\mathbf{x}(t)$, for h close to 0. A denotes the payoff matrix of some game. Adjustment dynamics are defined by taking the limit $h \to 0$. Accordingly, a dynamics $\dot{x} = \mathbf{f}(\mathbf{x})$ is an adjustment dynamics if and only if $\dot{x} \cdot A\mathbf{x} \geq 0$ and $\dot{x} \cdot A\mathbf{x} > 0$ whenever $\mathbf{x}$ is not a Nash equilibrium or a rest point of the replicator equation.

Every payoff monotonic dynamics is an adjustment dynamics. Moreover, best response dynamics and adaptive dynamics are also adjustment dynamics (see Hofbauer and Sigmund 1998 for details on those dynamics). The rationale of adaptive dynamics is that mutants use strategies close to the current state $\mathbf{x}$ such that the whole population is moving in the most promising direction. Best response dynamics may also be interpreted in terms of a large population model. A small fraction of individuals in a large population revises strategies from time to time by choosing a best reply to the current mean population strategy $\mathbf{x}$. On this interpretation, best response dynamics may be regarded as a boundedly rational dynamics. Payoff monotonic dynamics, best response dynamics and adaptive dynamics do not overlap. Thus, adjustment dynamics is a natural generalization of these three classes of dynamics.

The analogue to Theorem 3 for adjustment dynamics follows easily from the above definition of adjustment dynamics.

> **Theorem 4**. $u(\mathbf{x}, \mathbf{x})$ is monotonically increasing along every nonstationary solution and is constant on every connected set of stationary states for any adjustment dynamics of Σ_n^r.

> **Proof**. If A denotes the payoff matrix of Σ_n^r, then the symmetry of A implies that $\dot{x} \cdot A\mathbf{x} = \mathbf{x} \cdot A\dot{x}$. Thus $\dot{u}(\mathbf{x}, \mathbf{x}) = \dot{x} \cdot A\mathbf{x} + \mathbf{x} \cdot A\dot{x} = 2\dot{x} \cdot A\mathbf{x}$. By definition, the last term is greater than 0 for nonstationary solutions, and it is 0 if and only if $\mathbf{x}$ is a rest point. ■

Thus the average payoff is also a global strict Liapunov function for any adjustment dynamics of Σ_n^r. Since $\dot{x} \cdot A\mathbf{x} > 0$ for all $\mathbf{x}$ which are not rest points of the replicator equation and since Nash equilibria are rest points of the replicator dynamics, adjustment dynamics do not have more rest points than the corresponding replicator dynamics. This allows us to draw the same conclusions concerning the stability of rest points for adjustment dynamics of Σ_n^r as in the case of payoff monotonic dynamics.

5. Conclusion. Often it is difficult to judge whether one model is more realistic than another one. Robustness of a result across a variety of models—each of them being plausible—may be used as a substitute. In this sense, the emergence of states of partial or perfect communication is a robust result relative to changes described by payoff monotonic or, more generally, by adjustment dynamics. The evolution of simple communication systems in these classes of dynamics is at least as likely as it is in the replicator dynamics. This, on the other hand, implies that our results do not show that there exist adjustment dynamics which improve on the replicator dynamics, that is, in which the evolution of signaling systems is more likely than in the replicator dynamics. Results in this direction might be achieved only by studying more specific dynamics.

REFERENCES

Cressman, Ross (2003), *Evolutionary Dynamics and Extensive Form Games*. Cambridge, MA: MIT Press.

D'Arms, Justin, Robert Batterman, and Krzysztof Górny (1998), "Game Theoretic Explanations and the Evolution of Justice", *Philosophy of Science* 65:76–102.

Hirsch, Morris W., and Stephen Smale (1974), *Differential Equations, Dynamical Systems, and Linear Algebra*. Orlando, FL: Academic Press.

Hofbauer, Josef, and Karl Sigmund (1998), *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Huttegger, Simon M. (2007), "Evolution and the Explanation of Meaning", *Philosophy of Science* 74:1–27.

Lewis, David (1969), *Convention. A Philosophical Study*. Cambridge, MA: Harvard University Press.

Pawlowitsch, Christina (2006), "Why Evolution Does not Always Lead to an Optimal Signaling System", working paper, University of Vienna.

Skyrms, Brian (1996), *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

——— (2000), "Stability and Explanatory Significance of Some Simple Evolutionary Models", *Philosophy of Science* 67:94–113.

Swinkels, Jereon (1993), "Adjustment Dynamics and Rational Play in Games", *Games and Economic Behavior* 5:455–484.

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

https://doi.org/10.1086/525625 Published online by Cambridge University Press