The Evolution of Society

Jeff Inman
MIT Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139

August 5, 1991

AI-Working Paper-333

Abstract

We re-examine the *evolutionary stability* of the **tit-for-tat** (**tft**) strategy in the context of the *iterated prisoner's dilemma*, as introduced by Axelrod and Hamilton. This environment involves a mixture of populations of "organisms" which interact with each other according to the rules of the *prisoner's dilemma*, from game theory. The **tft** strategy is *nice*, *retaliatory* and *forgiving*, and these properties contributed to the success of the strategy in the earlier experiments. However, it turns out that the property of being *nice* represents a weakness, when competing with an **insular** strategy. A large population of **tfts** can resist incursion by a small number of **insulars**, but the reverse is also true, which means that **tft** is not an *evolutionarily stable strategy*. In fact, **insular** strategies prove to be better at resisting incursion. Finally, we consider the implications of this result, in terms of naturally occurring societies.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

Working Paper No. 333 August, 1991

The Evolution of Society

Jeff Iman

Abstract

We re-examine the *evolutionary stability* of the **tit-for-tat (tft)** strategy in the the context of the *iterated prisoner's dilemma*, as introduced by Axelrod and Hamilton. This environment invovles a mixture of populations of "organisms" which interact with each other according to the rules of the *prisoner's dilemma*, from game theory. The **tft** strategy is *nice, retaliatory* and *forgiving*, and these properties contributed to the success of the strategy in the earlier experiments. However, it turns out that the property of being *nice* represents a weakness, when competing with an insular strategy. A large population of the **tfts** can resist incursion by a small number of **insulars**, but the reverse is also true, which means that **tft** is not an *evolutionarily stable strategy*. In fact, **insular** strategies prove to be better at resisting incursion. Finally, we consider the implications of this result, in terms of naturally occurring societies.

1 Introduction

In 1981, Axelrod and Hamilton presented a paper in which they described a simplified environment in which to study the stability of various strategies in competition [1]. In their environment, competition is studied by modeling organism interactions in the context of the *prisoner's dilemma*, from game theory. The advantage of this context (compared with some contemporary evolutionary simulations) is that it has an easily computed metric for survivability; namely, the number of points an organism scores, relative to the rest of the population. Based on the stability of the cooperative strategy **tit-for-tat (tft)**, they argued that cooperation had evolutionary stability. This paper does not dispute that claim, but presents new strategies, which seem intuitively more natural and which prove more durable than **tft**. The new strategies are based on the **insular** behavior, which cooperates only with its own kind.

In fact, these simulations can not properly be called "evolutionary". They model the rise and fall, and hence the relative "survivability", of various configurations of populations of different species. However, they do **not** model the evolutionary machinations which might have allowed such distributions of populations to arise in the first place. These experiments show only how some forces may be exerted in a specified (simplified!) ecosystem, but they fail to address the more interesting systemic questions raised by evolutionary theory. In a real ecosystem, for example, the lions do not "win" if they consume all the antelope, yet an individual lion does win if it is a very successful hunter. In addition, a slightly advantageous mutation in the lions will select for corresponding mutations in the antelope, or vice-versa. Dawkins refers to this type of evolutionary feedback loop as an *arms race* [2]. Certainly, there are other types of feedback loops in which some species or individuals appear to have a symbiotic relationship.[1] Our experiments allow a limited, microscopic view of some of the forces that may be at work in a real ecosystem, but real ecosystems evolve as complete systems. In real ecosystems, not only do mutations occur, but a mutation in one individual has impact on its own and other species in the ecosystem. The analyses presented here, and in the earlier paper, proceed by dissecting the behavior of each species and studying its performance in different mixes of populations. Thus, we are making some very broad simplifications.

2 The Experiment

We simulate a mixed population of different "species" of organisms, competing against each other for survival. When any two organisms encounter each other they have only two options, as prescribed by the rules for the *Prisoner's Dilemma* [3]. Specifically, when encountering another creature, an organism may either *cooperate*, or *defect*. So, with two organisms each choosing between two actions, there are four possible interactions as shown in figure 1.

An initial population of organisms is generated. Then the simulation steps through the entire population, taking each organism and having it interact with another organism from the environment, chosen at random.[2] When two organisms interact, their strategies decide

[1] We say these relationships "appear" symbiotic (connoting mutualism), because it is impossible to assess the real meaning of a fragment of an ecosystem, especially when only viewed over a fragment of evolutionary history. For example, it could turn out that a relationship between A and B ultimately contributed to the extinction of C, which might later have saved the pair from being overrun by D, a late-arriving competitor.

[2] As a determinist, I think the words "at random" should always appear in quotes.

		B's action: *cooperate*	B's action: *defect*
A's action	*cooperate*	A 3, B 3	A 0, B 5
	defect	A 5, B 0	A 1, B 1

Figure 1: The distribution of points in the 4 possible scenarios where A meets B in the Prisoner's Dilemma game.

to *cooperate* or *defect*, without knowing the choice of the other. Points are then assigned to each of the organisms, based on the table in figure 1. The *lifespan* is the number of times the simulation steps through the entire population, before deciding to stop and produce the next generation. The lifespan is really an average over the population, because some organisms may interact slightly more than others, if they are more often chosen as the random partner for other organisms.

After a lifespan, a number of individuals of each species is produced for the next generation, in proportion to the relative number of points scored by that species. For example, if **tft** individuals account for 20% of the total points in the fifth generation, then 20% of the individuals in the sixth generation will be **tft**. The total population is kept fixed (at 50) and the absolute number of points scored is not important. If p is the total population and a species accounts for less than $1/p$ of the total points, then that species will not appear in the next generation[3]. The simulation continues until one species prevails, or a specified number of generations have elapsed.

[3]However, for any species scoring more than $1/p$ points, there may be some fractional point remainder. For example, if **tft** accounts for $1.999/p$ of the total score, then it would generate one individual in the next generation, but would have lost "credit" for the additional $0.999/p$ points it scored. So, this remainder is accumulated and awarded to the species most hurt by "round off". Thus, **tft** would be counted for $2/p$ points, producing 2 offspring instead of just one.

3 Some Strategies

3.1 DEFECTOR

The prisoner's dilemma is a game-theoretic model which is intended to capture an intuition about the nature of cooperation in a hostile setting. The points are distributed so as to produce a dilemma. If you could always defect while your opponent always cooperated, you would get the most possible points. Also, if your opponents might sometimes defect, then at least you would never get "taken". This simplest of strategies could be called the **defector**:

- always defect

Unfortunately, if everyone uses this strategy, nobody cooperates and everyone is stuck making 1 point per move. It is a safe strategy, and is arguably the best strategy to take if there is a low likelihood of two organisms ever encountering each other twice. An organism that is too trusting will provide an easy living for the "con men" (i.e. **defectors**). This is the root of the dilemma.

3.2 TIT-FOR-TAT

The strategy called **tit-for-tat** (henceforth, **tft**) can introduce cooperative behavior without being repeatedly "taken" by other organisms it encounters. **Tft** behaves as follows with each individual it meets:

- on the 1^{st} encounter,
  cooperate
- on the n^{th} encounter,
  do what the other guy did
  on the $(n\text{-}1)^{th}$ encounter.

Tft is successful because it does not continue to cooperate with others once they have defected (*retaliation*), yet it will resume cooperating after the other cooperates (*forgiveness*). It also initially cooperates with everyone it meets (*niceness*), which is understood as promoting cooperation. In the long run, this strategy will approach break-even, even with the most malicious opponent, because **tft** always gains back what it loses before cooperating.[4] Also, if two **tft** organisms meet, they will rack up the maximum number of (collective) points because they will never have reason to defect.[5] This is the basic idea behind Axelrod and Hamilton's assertion that cooperation is an *Evolutionarily Stable Strategy* (ESS). They showed both that a few individuals with this strategy are capable of overrunning a larger population of **defectors**, and that a large population of **tfts** are capable of resisting incursion by **defectors**.

In figure 2, the effect of initial population distribution is shown. Each datapoint represents an average, over 3 different runs, of the number of generations required for one species to prevail, given some starting distribution. We refer to this as the *latency*, where a latency of 3 means that the third generation consists of only one species. The type of marker indicates the species that finally prevailed, which in this case was nearly always **tft**. The bars around the data points indicate the range of the actual data. The lifespan was set at 200. In the case where the initial distribution was 10% **tft** and 90% **defector**, the simulation ran until it was called a draw at 50 generations, on all three runs. Examination of the population after such a draw shows that the distribution is still 10% **tft**, so this particular equilibrium seems very stable.

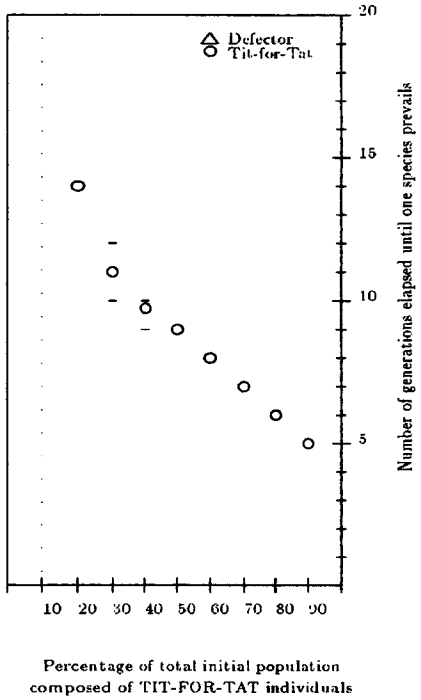

Figure 2: Initial population distribution affects the "latency" of the prevailing species

[4] Axelrod and Hamilton note that **tft** depends on a high likelihood of meeting any individual more than once.

[5] It is a significant property of the scoring that the total score for two cooperating organisms (3 + 3) is greater than the total score during a "con" (5 + 0).

3.3 INSULAR

Observe that **tft** is a rather egalitarian strategy. It suggests a social philosophy in which individuals participate equally in the formation of cooperatives. *Tft* suggests the power of a collection of "nice" individuals, which "retaliate" when attacked and which "forgive" when appeased. Another possible strategy, which we call **insular**, places more dependence on the collective by favoring other individuals of its own kind. Here is **insular**'s strategy:

> • IF the other guy is also an **insular**,
> THEN cooperate
> ELSE defect

This strategy, distasteful as it may seem, trounces **tft** in a wide range of settings. The trick is that every **insular** can con every **tft** once. After that, every previously acquainted different-species pair will mutually defect, while every same-species pair will cooperate. The "edge" gotten in the first encounter, is the essential advantage for **insular**. The rest is a matter of the weight of numbers of the two opposing populations. It was noted in [1] that **tft** depends on a high likelihood of meeting any organism more than once. We preserve this condition by using a lifespan of 200. We could elaborate the model by supposing that different species might inhabit physical niches, increasing the relative likelihood of same-species contacts, or by supposing that an **insular** individual could tell when it was at a disadvantage and would then begin cooperating more, or by supposing **insular** was more motile, or shorter-lived, or by increasing the payoff for a "con", and so forth.

The **insular** strategy, as the name suggests, is better at resisting incursion than **tft** is. This is the main result on which this paper is based. Figure 3 shows latency over initial population

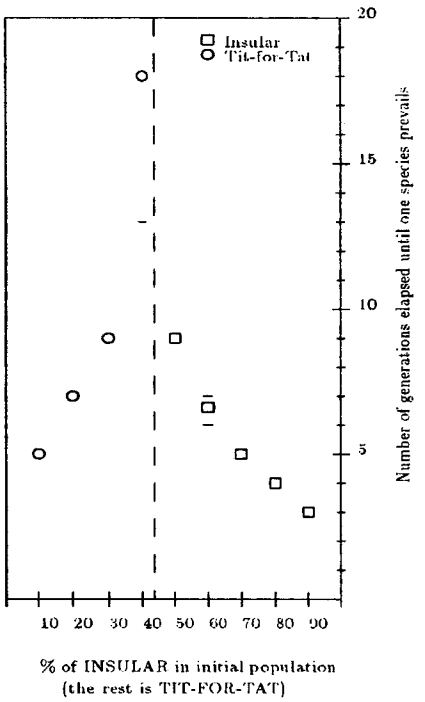

Figure 3: Initial population ratio vs. latency, for **tft** and **insular**

distributions for **insular** versus **tft**. Empirical experiments show that **insular** tends to prevail if given more than 42% of the total initial population (21 out of 50). This equilibrium is not quite as stable as the one shown in figure 2, due perhaps to the fact that our total population is only 50.

3.4 INSULAR2

Insular uses a type-check to recognize others of its kind without ever interacting with them. This might be considered as "cheating", because it involves a facility other than memory to support a strategy. This may also contradict the premise of the prisoner's dilemma itself, because **insular** gets information through channels other than cooperation and defec-

tion. (Given our context, this is not really such an unreasonable technique. We might suppose that selection would tend to produce just such "builtin" capabilities as these. We will return to this thought in section 5.) Anyhow, in order to play more fairly, we developed a version of **insular**, called **insular2**, which recognizes its own kind using a *behavioral protocol* based on a specific sequence of cooperations and defections.[6] **Insular2** uses the simplest signaling and recognition protocol that will allow it to distinguish **insular2**s from **defectors**, **tfts** and **insulars**.

- on the 1^{st} encounter,
  defect
- on the 2^{nd} encounter,
  IF opponent defected on 1^{st} move
  THEN cooperate
  ELSE defect
- on the 3^{rd} encounter,
  IF opponent defected on 1^{st} move,
  AND cooperated on 2^{nd} move
  THEN "recognize" opponent
  (cooperate)
  ELSE defect
- on the n^{th} encounter,
  IF opponent is "recognized"
  THEN cooperate
  ELSE defect

In practice, the implementation of this behavior is simpler than its description here. The point is that this protocol suffices to recognize other **insular2**s and to selectively cooperate only with them. This strategy dominates **tft**, but loses out to **insular** which need never cooperate with any but its compatriots. Note also, that each **insular2** will be conned once by each **defector**, just as each **tft** will be conned once by each **insular2**, but the difference is that the **insular2**s are capable of intra-species cooperation, while the **defectors** are not. This gives **insular2** an advantage over **defector**, which **tft** does not have over **insular2**, as suggested by figure 4. Empirically, we found a transition point (rather than a stable equilibrium) at starting populations of between 43% **insular2** (22 out of 50) and 46% **insular2** (23 out of 50).

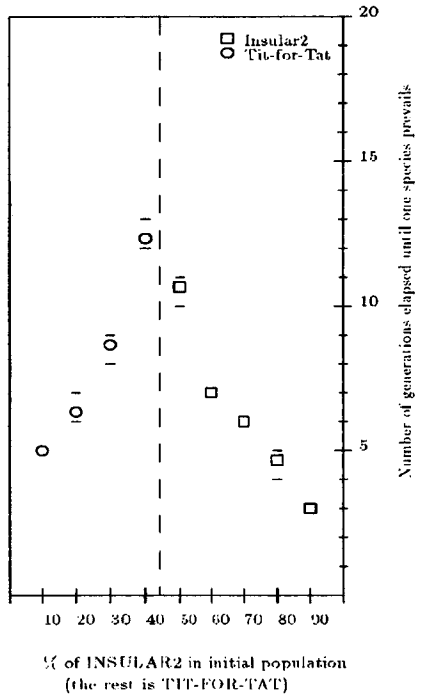

Figure 4: **tft** versus **insular2**

4 Some Analysis

The **insular2** strategy defects in its first interaction with anyone. This means that "nice" **tft** gets conned in its initial dealings with each

[6] We refer to this as the *nudge-nudge wink-wink protocol*

insular2. However, other **insular2**s are not conned, because they mutually defect on their first meeting. In the second encounter with an **insular2**, **tft** will defect, but the **insular2** will also defect, because it will recall that **tft** did not defect in the first encounter and is therefore necessarily *not* a fellow **insular2**. The effects of the first encounter are as follows, showing all possible pairings of **tft**, **insular2** and **defector**:

Strategy A		vs Strategy B	
TFT	= 3	TFT	= 3
TFT	= 0	INSULAR2	= 5
TFT	= 0	DEFECTOR	= 5
INSULAR2	= 1	INSULAR2	= 1
INSULAR2	= 1	DEFECTOR	= 1
DEFECTOR	= 1	DEFECTOR	= 1

In the second encounter between two **insular2**s, both will cooperate. A second encounter between an **insular2** and a **defector** will hurt the **insular2**, but it will never trust the **defector** again. In other words, **insular2** and **tft** score identically against **defector**, for any number of moves greater than 1. However, **insular2** scores better against **tft** than vice-versa, because **insular2** profits from the initial encounter. After two encounters, **insular2**s have recognized each other (using the *nudge-nudge wink-wink protocol*) and will always cooperate in the future. On the second encounter, things go as follows (accumulated totals are in square brackets)

TFT	= 3 [6]	TFT	= 3 [6]
TFT	= 1 [1]	INSULAR2	= 1 [6]
TFT	= 1 [1]	DEFECTOR	= 1 [6]
INSULAR2	= 3 [4]	INSULAR2	= 3 [4]
INSULAR2	= 0 [1]	DEFECTOR	= 5 [6]
DEFECTOR	= 1 [2]	DEFECTOR	= 1 [2]

Every subsequent move now adds a constant value to every individual. The **insular2**s are cooperating with the **insular2**s, the **tft**s are cooperating with the **tft**s, and all other combinations are mutually defecting.

TFT	= 3	TFT	= 3
TFT	= 1	INSULAR2	= 1
TFT	= 1	DEFECTOR	= 1
INSULAR2	= 3	INSULAR2	= 3
INSULAR2	= 1	DEFECTOR	= 1
DEFECTOR	= 1	DEFECTOR	= 1

This means that after the second move, the determining factors are the initial distribution of species and the lifespan. The **tft** population can regain its collective deficit only if there are enough more of them, and enough time, to out-cooperate the **insular2** population. The same argument holds, in the case where no **defector**s are initially present, except that here **tft** is only conned by one species and **insular2** is never conned.

Notice that **insular2** is extremely vulnerable to a new organism that imitates the **insular2** protocol for two moves and then defects. In fact, **insular2** even lacks the ability to "retaliate" because it considers recognition to be final and will always cooperate with other organisms once it has "recognized" them as fellow **insular2**s. Thus, **insular2**'s deep "imprinting" is a risky design, if we allow new organisms to be introduced. This could be fixed by giving **insular2** the ability to change its mind about other **insular2**s which subsequently defect. It would then only be vulnerable once, to being conned by this new organism. However, that one con was sufficient to give **insular2** the edge over **tft**, so it should also be sufficient to give the new organism an edge over **insular2**.

The relative invulnerability of **insular** to deception suggests that a reliable and inexpensive recognition protocol is a valuable component of insular strategies.[7] However, an

[7] An interesting subtlety was discovered when we realized that the original **insular** strategy was vulner-

overly cautious recognition protocol might often make **insular2s** behave more like **defectors**, which would prevent them from developing cooperative relationships.

In the algorithm described earlier, the partner for each organism is selected at random from the environment. Fluctuations in datapoints, in the earlier figures, show where the "latency" of the dominant species is most sensitive to slight fluctuations in the numbers of encounters between and within the various species in the environment. In order to assure that **tft** has a sufficient probability of meeting any opponent at least twice, it is necessary to make the lifespan long enough that most encounters with **tfts** will in fact occur more than twice. It is instructive to look at the results of a non-randomized algorithm, where every pair of organisms would meet some precise number of times. The results in figure 5 were obtained by introducing every individual to every other individual a fixed number of times per generation. The different lines show datapoints for 2, 3, 4 and 10 meetings per generation. This data makes the analysis a little more clear. Where the curves touch the top margin, the populations were at equilibrium. The general trend, as expected, is for the equilibrium point to move towards the 50% distribution level, as the number of encounters per organism per generation increases. This is because the extra points scored by each **insular2** in its initial encounter with each **tft**, account for less of the total score when more interactions happen per generation.

Figure 5 also illustrates a loophole in the reproduction algorithm, described in section 2. Notice that the curve for 3 meetings/generation, appears to be at equilibrium at *all* low percentages of **insular2**. In fact, the latency of **tft** *is* decreasing for lower values of **insular2** in that part of the graph. In those experiments, **tft** would tend to dominate, except a "false" equilibrium is reached at 48 **tfts** and 2 **insular2s**. What happens is that those 2 **insular2s** score between 3 and 4% of the total points, enough to generate 1.5+ individuals in the next generation. This is rounded to 1 individual, with a remainder of 0.5+. However, the reproduction algorithm awards the total remainder, after rounding, to the species that was most hurt by "round off error" (see footnote on page 2). Thus, one more **insular2** individual is produced, for a total of 2 **insular2s**, and the cycle repeats indefinitely. This

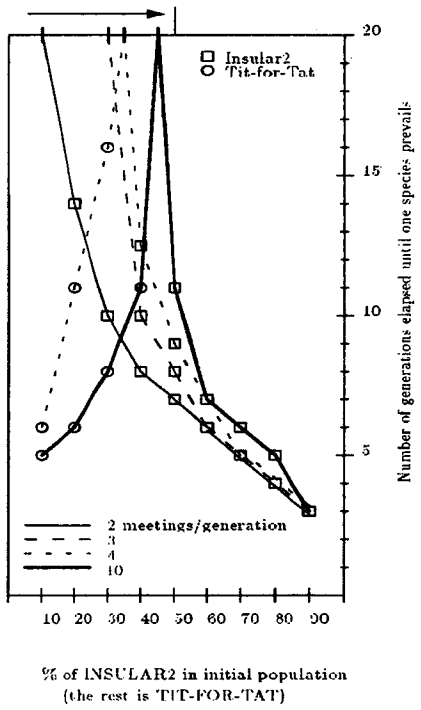

Figure 5: **tft** versus **insular2** using precise numbers of meetings between individuals.

able to "mimicry" in the form of super-types, because it used the Lisp *typep* function to do its recognition. A super-type object, could pass the *typep* test without being specifically of **insular** type. This was easily repaired by making the test more specific, but it demonstrates that even "built-in" recognition protocols may be vulnerable.

round-off crediting may account for slight perturbations in latency values, but the points of equilibria shown in the earlier figures are empirically not of this "false" type. This evidence is also supported visually by the falling away of latency on both sides of our equilibrium points.

5 Reflections

Consider the wolf pack, the school of fish, the flock of birds, and so forth. Though some cooperative social endeavors may be more loosely knit, we can see that "banding together" is common throughout the animal kingdom. We are deeply attached to the values of our own society, so it might be easier to consider the study of social insects. Higher degrees of "sociality" in insects are generally defined to involve various aspects of care for the young, collective food consumption, and so forth [6]. In this way it becomes more clear that a "society", by human definitions, has a *collective self-interest.* Human collectives seem to exist at the level of the family, neighborhood, city, state, nation, and (in the case of some of the more enlightened) the planet or even the macrocosm. It could also be argued that there are collective principles involved in schools, businesses, religions, and political movements. These collectives seem to be natural aspects of human society and yet they may also seem to be the source of much of what we find *destructive* in human society. We sense bigotry and exploitation to be offensive because they suggest the use of cooperative effort to maintain the authority of an "elite", regardless of the competence of the unestablished challengers. Yet it seems pretty clear, for example, that humans consider human lives to be more valuable than the lives of other species. This paper does not claim to resolve these difficult social issues. However, our simple experiments with the Iterated Prisoner's Dilemma *suggest* that there *may* be evolutionary forces that favor self-preserving social orders.

It makes for interesting speculation to consider that some antisocial or exploitive behavior we observe in human society could be viewed as the product of overly restrictive recognition protocols; overly narrow criteria for recognizing one's "own kind". We might also imagine that the much touted discriminatory and manipulative capabilities of humans could sometimes have the unfortunate side-effect of making it harder to appreciate one's relationship to other humans, other species, and to the ecosystem as a whole. Yet, there are many levels of differentiation which individuals may use to identify their "own kind". Could it be that these communities of greater or lesser generality are sensed in relation to specific types of situations? For example, depending on the perceived nature of specific challenges, we may consider ourselves as individuals; as members of a family, city, nation, or species; as holders of a perspective, philosophy or religion; as attributed with gender, heritage, lifestyle, income, disability, traumatic experience, achievement; and so forth. Perhaps it is this flexible group identification that has given *homo sapiens* its (recent) survivability?

The assessment of *niceness* as a "weakness" in **tft** should be qualified. In the oversimplified environment of these experiments, we are missing some of the more complex forces that are necessary to hold a society together when faced with new challenges. I have argued that "banding together" may be an *evolutionarily stable strategy*, but we should be careful about extrapolating this as a justification for bias. Excessive elitism may be just as destabilizing as excessive *niceness.* While there may be social forces (analogous to those demonstrated with our simulated organisms) that promote

"banding together", there might also be complimentary forces that censor those organisms having overly rigid recognition protocols. In fact, as we speculated before, overly rigid recognition protocols may ironically produce "anti-social" behavior. Human societies seem to perceive such behavior as threatening, and threatening things are also often considered to be "alien", so perhaps there is a dynamic tension between the elitist and communal impulses. In this way, the "banding together" impulse may actually be self-regulating. Elitist individuals or sub-societies are perceived as a threat by the rest of society, yet society itself is in some ways an elitist endeavor.

The evidence seems strong that social principles may be genetically "programmed" in a species. At least, most people would not hesitate to recognize this in *other* species. But, what about humans? We'd like to think that our social systems are built on something uniquely superior to genetic influence, namely, the perception of *justice* and our sense of *belonging*. I propose that a sense of *justice* and of *belonging*, and other things, may in fact be part of our genetic heritage. It seems clear that they are part of the "glue" that holds human societies together. Since humans have evolved in societies, it doesn't seem unreasonable that societies and individuals should be so subtly interwoven.

6 Future Work

We still need to explore a more "realistic" context for observing evolution in action. For example, why not allow "mutations" which would produce slight deviations in the recognition protocol of **insular2**s? Such a mutation could produce the briefly-described behavior which tricks **insular2** by mimicking its recognition protocol and then defecting. As recognition and deception enter the picture, species might begin to develop interdependent predator-prey relationships, such that the predators kill, and yet depend on the survival of the prey species that they have evolved to hunt. Perhaps, interspecies symbioses could evolve in such systems, eventually yielding the kind of systemic interdependence which we found lacking in the current experiments.

Unfortunately, the metaphorical context of the prisoner's dilemma is at odds with the systemic perspective, as this paper has implicitly demonstrated. The prisoner's dilemma doesn't offer much flexibility for appreciating the subtle relationships found in real ecosystems. For example, organisms would have to "cooperate" with organisms higher on the food chain, while "defecting" with organisms lower on the food chain (the Sun, at the bottom, cooperates with plankton and does not need to reproduce). But this terminology is awkward and seems to take too microscopic a view of the activity in an ecosystem.

References

[1] Axelrod and Hamilton, "The Evolution of Cooperation", **Science**, vol. 211, March 1981

[2] Dawkins, *The Blind Watchmaker*, W. W. Norton and Company, New York, 1986

[3] A.Rapoport and A.M.Chammah, *Prisoner's Dilemma* University of Michigan Press, Ann Arbor, 1965

[4] Huberman, B.A., *The Ecology of Computation*, North-Holland, New York, 1988

[5] Miller, M.S. and Drexler, K.E., "Comparative Ecologies", *in Huberman*

[6] Wilson, Edward O., *The Insect Societies*, The Belknap Press of Harvard University Press, Cambridge, MA, 1971