1.211.4

# Re-evolving Re-identification

**Philip Jones**

Submitted for the degree of M.Phil

University of Sussex

April, 2002

# Declaration

I hereby declare that this thesis has not been submitted, either in the same or different form, to this or any other university for a degree.

Signature:

# Acknowledgements

This thesis would not have been possible without the educational and emotional infrastructure provided by my fellow students, friends and family. This work supervenes heavily on the following people ...

Argiri Arfani, Gisel Carriconde Azevedo, Guillaume Barreau, Tom Beamont, Hilan Bensusan, Jo Brook, Seth Bullock, Daniel Cain, Ron Chrisley, Julie Coultas, Clive Cox, John and Kathleen Dege, Manuel De Pinedo Garcia, Stephen Dunn, Stephen Eglen, Peter Elliot, Berkan Eskakaya, John and Susan Jones, Matthew Knight, Ronald Lemmen, Murali Ramachandran, Jason Noble, Nadja Rosental, Oliver Sharp, Darius Sokolov, Adrian Thomson.

Further advice and stimulating discussion was provided by Harry Barrow, Ezequiel Di Paolo, Joe Faith, Inman Harvey, Phil Husbands, Giles Mayley, Mike Scaife, Blay Whitby, Michael Wheeler.

Shouts also to ... everyone at Runtime Collective, Aharon, Angela, Sara, Richard and Charlotte, Sinnet and Jordan, John and Jane, Christy, Shirley, Adalene, Victoria Real and John H. (who got me into this A.I. stuff to begin with)

Peace to anyone I forgot ...

This thesis is dedicated to the memory of John Jones and John Dege.

# Re-evolving Re-identification

**Philip Jones**


## Summary

An individual is someone you meet, part with, and, upon meeting again, carry on the same relationship with. To live in a world of individuals, one must have the cognitive faculty of individual recognition. But what situations encourage this to evolve?

In this thesis I look at one answer to that question. That individual recognition evolves to allow us to form reciprocally altruistic societies. I therefore look at external encouragements and constraints on the evolution of individual recognition in players of the iterated prisoner's dilemma.


Submitted for the degree of M.Phil

University of Sussex

April, 2002

# Chapter 1

# Introduction

---

### 1.0.1 Introduction to the introduction

*In a world of individual success and failure, how were the seeds of cooperation nurtured? One answer to this depends upon being able to recognise another animal as an individual - and being able to remember if he took advantage of your cooperation during your last encounter with him ... Obviously, a brain good at recognising faces would be a better brain to benefit from the virtues of cooperation ... We tend to take memory for granted without enquiring into the evolutionary pressures that that might shape memory capabilities... One candidate is the individual recognition needed for cooperation strategies.*
William H. Calvin[8]

There we have, in a nut-shell, the intuition behind this thesis, whose purpose, put as succinctly as possible, is *to investigate the evolution of individual recognition as a means to enabling reciprocal altruism.*

However, although simple to state, this characterisation is ambiguous and confusing. Here are just some of the pitfalls that await when making sense of the above statement.

- What is an investigation into the evolution of something? Is it a research leading to a claim about actual historical happenings that occurred in the phylogeny of a particular species? If so, how could this be done, as my work is, *in silico*, by experimenting on virtual creatures? If not, what else could it be?
- What exactly is *individual recognition*? Is it an observable behaviour, one of several which ethologists have defined, or is it a cognitive capacity which is inferred from that behaviour? If the former, which behaviours? And if the latter, in what sense *can* we experiment on a cognitive capacity?[1]

[1] A particularly key question, if we think of it (individual recognition) as a cognitive capacity which is inferred from behavioural range *B*, when modelling the capacity in a virtual creature, is producing behaviour *B* a) necessary, and b) sufficient to say we are modelling such capacity?

- What is *reciprocal altruism*? How can it be differentiated from other altruistic or even apparently altruistic behaviour?
- And anyway, why do I talk about the evolution of a trait as a means to enabling reciprocal altruism. Does this mean that other reasons for the evolution of the trait are ignored? What if the trait doesn't, in fact, enable reciprocal altruism? Is this "enabling" subject to empirical testing, or is it an assumption?

Now, while I hope the above list alerts the reader to the fact that I am aware of these pitfalls, I will not attempt to answer these questions, systematically, here. By the end of the thesis, I do believe that a coherent position will have been taken and defended on all these points; and that the work can be seen as consistent with this position.

Here I do want to stress one issue. These pitfalls and confusions are exacerbated, I believe, because the work is influenced by multiple disciplines, each of which flavours the understanding of certain terms in their own way. Hence the idea of using *in silico* evolution of virtual creatures is typical of the field known as *Artificial Life*. However, much artificial life is focused on analysing the interactions of behaviours; whereas my notion of *individual recognition* is far closer to a cognitive capacity. Even more atypically, it is a notion more inspired by philosophy than science. This would be far less unusual in a good old fashioned *Artificial Intelligence* context where cognitive philosophers have traditionally had an influence on AI research, but seems almost uncouth in the ALife world where philosophers' ideals of rationality are rejected in favour of behavioural kludges inspired by biology.

Another area of confusion : I borrow the *iterated prisoners dilemma* - one of the most famous *games* in game theoretic literature, and one which is used throughout evolutionary, economic and social application of game theory. Yet I am not very interested in co-operation or why it should exist. I believe that the prisoner's dilemma is a suitable abstract and plausible model of a social environment; and that it is a game where co-operation can be sustained due to individual recognition. It can also be sustained without individual recognition.

## 1.1 Outline of the following chapters

- Chapter 2 : Altruism and Evolutionary Theory. This chapter discusses the problem of altruism for evolutionary theory and introduces the ways that it has traditionally been dealt with. Including the theories of mutualism, group selection, kin selection and reciprocal altruism.
- Chapter 3 : Biological Approaches To Individual Recognition. This chapter surveys ethological literature on individual recognition. Where it is found and what it is used for.
- Chapter 4 : Computer Based Research of the Prisoner's Dilemma. This chapter looks at some notable computer based research I have drawn inspiration from. Particularly that of Robert Axelrod.
- Chapter 5 : Introduction to the Experimental Work. This chapter gives the aims of my experimental work and shows some results of pilot studies with some simple models. The chapter introduces three research topics under which the experimental work is organised :

*The Evolution of Individual Recognition*, *The Role of Kin Oriented Altruism in the evolution of TFT* and *Games on a Spatial Grid*

- Chapter 6 : The Complex Model : Framework 3. Framework 3 is the name for the main software model used in my experiments. This chapter explains how it works and what values are recorded from it.
- Chapter 7 : The Evolution of Individual Recognition. This chapter gives results of experiments into the evolutionary dynamics of individual recognition.
- Chapter 8 : The Role of Kin Oriented Altruism in the evolution of TFT. Robert Axelrod suggested that a pre-existing kin oriented altruism might be a precursor to the evolution of reciprocal strategy like TFT. What can the model tell us about this?
- Chapter 9 : Games on a Spatial Grid. An alternative to the idea that TFT was helped by kin oriented strategy comes from an observation that in restricted environments such as grid-worlds or viscous populations, pockets of indiscriminate co-operation can survive due to group-selection. On a simple grid-world variation to the model, we investigate whether this effect encourages or discourages individual oriented strategy.
- Chapter 10 : Discussions. On the results, on the choice of this notion of individual recognition, on ALife's scientific respectability and on future work.

# Chapter 2

# Altruism and Evolutionary Theory

## 2.1 Introduction

The next three chapters cover some background issue to the research.

- A discussion of the problem of altruism, in evolutionary theory.
- A survey of experimental approaches to the prisoner's dilemma,
- An overview of biological literature on individual recognition.

## 2.2 The history of evolution

The tradition of evolutionary research that concerns us can be sketched as starting with Darwin[11] whose essential ideas were these.

- That the existence of animal species might not be the result of special creation by a divine designer, but instead, that two different species could be descended, with divergent characteristics, from a common ancestor.
- That this might be generalised for all species. Possibly all life on earth has evolved from a single species. This therefore explains the existence of the species and their particular characteristics.
- That the changes are the result of a directed, selection process analogous to the selective breeding programs that lead to different breeds of domestic animals.
- That unlike artificial selection, this selection is intrinsic to nature and the result of the struggle for survival between individual animals. Those animals which are best suited to take advantage of their situation are most likely to leave children who inherit similar traits.

## 2.3 The Problem of Altruism

The focus of this thesis is individual recognition as a means to sustaining reciprocal altruism. Here I introduce the "problem" of altruism for traditional evolutionary theory. Before we can see why there is a problem, let us briefly be clear on another question : *What is evolutionary theory for?*

One answer to this question is that it is to *explain* the existence of biological things. Questions of the form "why are there Xs?" can be answered using evolutionary theory. The form the answer takes is : *there are Xs because Xs evolved*, and what that implies is something like the following :

- Obviously there was a time before there were Xs. Nevertheless, there were ancestors who were or had very primitive Xness, or proto-Xs, and these tended to do better than those who didn't, and so prospered.
- The result was that soon everyone had some Xness. But further, the more advanced the Xness was towards modern Xness, the more it helped those individuals thrive and reproduce relative to other members of the species.
- And so, over many generations, Xness became more and more pronounced in the population until it reached the position you see today.

There is a great deal wrapped up in "tending to do better". Typically we recognise that there are two clear ways of "doing better". One is to live more, to live better, to acquire resources and so leave more children. The other is to increasing the number or quality of one's offspring through mating with more or better members of the opposite sex. The first of these notions of "doing better" is known as survival. And sometimes we talk about *fitness* as the measure of it. The second is referred to when commentators talk about "sexual selection".

Often the distinction is not interesting. But occasionally one may be faced with a trait T which confers no apparent survival value and conclude that its purpose is purely to attract mates. Very naive critics of evolutionary theory - who suppose that it is only about "survival" - sometimes appeal to such traits to argue against the theory as a whole.

This is a broad brushed account, but it will do[1]. I am not in the business of defending evolutionary theory from radical detractors here. But to understand the issues related to altruism, it is worth putting oneself into those detractors' shoes for a second.

There *is* a more problematic class of traits which challenges the above explanatory strategy : those traits which seem to directly sacrifice one's fitness, or promote the fitness of rival members of the same species. These are the apparently *altruistic* traits.

It is hard to see how these traits should have evolved according to the story above, because, by definition, they don't improve their owner's chance of doing better than rivals. In fact, they help the rivals, and so make the owner of the trait relatively less fit. This is a trouble for evolutionary theory

[1]It should be noted that I use "evolutionary theory" almost interchangeably with a position sometimes known as the *adaptationist program* or even *panglossian paradigm*, famously criticised by Gould and Lewontin[16] who point out that not everything that exists can or should be explained in terms of its fitness contribution or function. Whilst this is undoubtedly true, I nevertheless hold the adaptationist program to be the core of evolutionary theory and related ethological and ecological research. As I argue later in the majority of biological "kinds" of the sort which are amicable to scientific investigation are defined in terms of their adapted function.

because one of its great strengths is its universality. Darwinism seems to solve every problem of the form "why are there Xs?". If it doesn't, particularly if it leaves mysteries, such as altruism, for which we must appeal to a rival explanation, then the suspicion must arise that this alternative theory might be a more suitable candidate to explain everything.

Altruism, therefore, is a problem for the naive Darwinian theory sketched previously, but we have four explanations offered for it that *are* compatible with the Darwinian world-view.

- Mutualism
- Group Selection
- Gene Selection or Inclusive Fitness
- Reciprocal Altruism

### 2.3.1 Mutualism

Mutualism is the case where animal, X, behaves in such a way as to increase the fitness of animal, Y. But in fact, X also increases her fitness. Flocking for protection from predators, huddling together for warmth, symbiotic relationships between members of different species, are all examples of mutualism. Cases where there is some conflict, because the benefits are uneven, can still be cases of mutualism, as long as both parties ultimately gain from their respective behaviours. Any apparent altruism that also increases one's own fitness can be seen as an example of mutualism.

### 2.3.2 Group Selection

An early assumption about some altruistic behaviours, was that they were due to the phenomenon we call *group selection*. A behaviour of excessive altruism (such as risking or sacrificing one's own life) might be for the "good of the species". One can see why, humans, who sometimes sacrifice their lives for the good of the cause or the country, would have little trouble appreciating the idea. But from the point of view of evolutionary theory it is more problematic.

For "group selection" to work on evolutionary grounds the benefit to the group, of the altruistic individual, must have greater evolutionary impact than the deficit to the individual. For example, a something like this should be the case :

- an animal, X, must be a member of a sub-population of a larger population.
- Within this sub-population, the trait of self-sacrifice for the good of the group (the sub-population) must have arisen.
- Despite some individuals sacrificing themselves, this trait must survive within the sub-population.
- Because of the benefits to the overall sub-population, from self-sacrificing individuals, this sub-population must itself thrive relative to the other sub-populations of the species, so allowing it to expand and take over their resources.

This scenario was successfully criticised by Williams[48] who pointed out that, the competition between individuals within the sub-population would happen at a higher frequency than competition between the sub-populations. As the altruistic behaviour is purely detrimental to the individual who carries it, you would expect it to be driven extinct *within* the sub-population, long before its benefits to the sub-population as a whole had a noticeable effect.

More recently there has been a revival in group selection ideas.[49] Before considering them (in 2.4.2), we must touch upon the successful alternative which displaced it.

### 2.3.3 Gene or Kin Selection

Hamilton[19] provided an alternative to group selection that proved effective in solving one of the most widely observed cases of altruism. This solution focused on the *gene*, rather than than the group or individual, as the *unit of selection*. The gene was typically understood as the chemical recipe which carries a description of a trait from parent to child.

If one switches one's perspective to that of the gene, it is possible to see that it is genes, rather than individuals, which are in competition for survival, and to which a notion of fitness can be attached. How does this help explain altruistic behaviour? In the paradigm case, thinking in terms of gene selection can explain perfectly why animals should co-operate with their relatives. If X shares half of the same genetic material with her sister, any behaviour she has which increases the sister's fitness will increase the fitness of the genes that they share. Hence X's genes have an incentive to promote co-operation with X's kin, even at a cost to X herself, as long as the net gain to the genes is greater than the benefit of X remaining a selfish individual.

This is dramatically demonstrated in eusocial[2] insects where the majority of females remain sterile in order to support the queen (their mother) in raising further siblings. The reason is due to a genetic quirk. The majority of these social insects are *haploidiploid* which implies that they share more genetic material with their sisters than their offspring. Consequently supporting their mother produce more sisters, promotes their genes more successfully than producing their own children[3].

Because gene selection so obviously explains co-operation with relatives, we sometimes label this kind of explanation using the term *kin selection*.

### 2.3.4 Reciprocal Altruism

The final explanation of altruistic behaviour is *reciprocal altruism*, famously discussed by Robert Trivers[46].

According to this explanation, X will perform a favour for Y on the assumption that Y will later perform a similarly valued favour back for X. Such bargains needn't be explicitly recognised as such by the animals. Co-operation can be unreflective and instinctive. But for reciprocal altruism

[2] Societies "traditionally characterised by reproductive division of labour, an overlap of generations, and co-operative care of the breeders' young".[41]

[3] Not all eusocial species are haploidiploid. Bees, wasps and ants are, but similarly social termites are not.

to be evolutionarily viable there must be plenty of opportunity for both X and Y to behave altruistically, and the benefits of mutual co-operation must outweigh the value of non-co-operation. Overall both parties must benefit from their co-operative behaviours.

How is this case different from mutualism, as described above? As Trivers stresses, the distinguishing feature of reciprocal altruism is that there *is* an opportunity for one (or both) partners to renege on their reciprocation; to benefit themselves greatly by accepting the altruism from the other, but to choose *not* to reciprocate it. Theorists have long modelled this kind of situation using the *prisoner's dilemma*.

I will take the relationship to be *definitional*. When I write of *reciprocally altruistic situations*, I am writing *only* of those situations which the prisoner's dilemma is a valid model of[4]. Hence I will not use reciprocal altruism when I mean mutual co-operation where there is either no possibility of one party *defecting*[5], or no incentive for a party to do so.

As already mentioned, that animals engage in reciprocally altruistic behaviour, doesn't *necessarily* imply anything about their cognitive capacities. But the possibility of cheating makes new demands on the explanation we are offering. Typically, in the reciprocal altruism case, the question, "why does this altruistic behaviour exist?" is accompanied by the implicit sub-text. "Sure, mutual co-operation is obviously a good thing. But how is honesty maintained? Why, given the superior benefits to the individuals, of accepting co-operation without reciprocating, doesn't the cheating behaviour dominate and actually drive altruism extinct?" This, then, is what is at the core of a reciprocal altruism explanation for altruistic behaviour : an account of how cheating is prevented.

*Avoiding cheats*

Recognising cheats might be possible through innate traits, perhaps even social markings such as a public criminal record. But innate markings are likely to be disguised by mimics, and public branding would seem to require a sophisticated public cultural mechanism to work. Trivers, at least, does not offer it as a possible solution. The solution which interests us is for individuals to remember the previous behaviour of opponents and to be able to re-identify those individuals. Thus, where reciprocally altruistic behaviour is apparent, *and where there is no other mechanism for preserving honesty*, we would expect to find individual recognition alongside a strategy of aiding only those known to reciprocate.[6]

[4]Note : this is to distinguish reciprocal altruism from mutualism. As we will see later, some situations that really are prisoner's dilemmas, nevertheless allow sub-populations of indiscriminate altruists to survive due to certain other population characteristics. I believe all these cases are covered by, what I will call neo-group selection theory.

[5]The opposite behaviour to co-operating in prisoner's dilemma parlance.

[6]Trivers is not interested in individual recognition. His first aim is to convince the reader that reciprocated altruism can be a viable and robust behaviour, given certain costs, benefits and the demography of the population. One of the main examples given are shrimps that provide a "cleaning service" to fish by eating parasites from inside the fish's mouth and gills. As Trivers points out, the client fish being larger, and having the cleaner already in its mouth, could "cheat" by swallowing the cleaner, thus gaining a free meal. Trivers considers not swallowing to be a reciprocating co-operative behaviour performed by the fish. Such fish-to-shrimp co-operation even includes the host risking its own life to allow the shrimp to leave before fleeing from a predator. However in the case of these fish, recognition appears to be due to recognition of geographical features. Shrimp spend their entire lives in one location and fish learn to come to this location to be serviced by the same cleaner.

### 2.3.5 Trivers on social factors varying with reciprocal altruism

Trivers identifies the observable population characteristics that could affect whether reciprocally altruistic behaviour will be selected for. These are :

- length of life,
- dispersal rate,
- degree of mutual dependence,
- parental care,
- dominance, and
- aid in combat.

He notes that long-lived creatures are likely to find themselves in a greater number of potentially *altruistic situations*, which he defines as those situations where an individual can "dispense a benefit to a second greater than the cost of the act to himself". A low dispersal rate will increase the likelihood of these situations being with the same individuals. Species with these characteristics are good places to look for reciprocal behaviour. A high degree of mutual dependence in an activity such as foraging, will also increase the number of possible repeat altruistic situations between conspecifics.

On the other hand, according to Trivers dominance hierarchies (see 3.3.2) discourage reciprocal altruism. He bases this assumption on the grounds that a dominant individual can demand aid from another without reciprocating. Trivers here considers the aid as giving up or sharing food or equivalently valuable resource. When dominant individuals need the aid of underlings - in a situation such as a dominance challenge to a rival - then a more symmetrical relationship might be possible, and thus reciprocation can appear.

Finally, family relationships confuse things for the observer. Individuals might behave altruistically towards their relatives due to kin selection. But reciprocally altruistic behaviours could still be selected for on top of the already altruistic behaviour.

So, the picture given by Trivers is that we are likely to observe reciprocal altruism when lifetimes are long and repeat altruistic situations are high. However, dominance will likely override it, unless allies need to be recruited from among the underlings.

## 2.4 Recent work on altruism challenging the traditional explanations

In recent years, the four traditional explanations for altruism, have been subject to some hard scrutiny and criticism. One historical cause of this has been the increase in computer based numerical simulations which have allowed us to study more detailed and specific situations.

I would characterise the new critique as falling into two and a half camps.

We might be concerned whether this as a case of reciprocal altruism as I have defined it. Does the shrimp chose which fish to co-operate with, or just service any that go past? Does the shrimp have an opportunity to "cheat" on a co-operating fish? If the game is so one sided, it may be closer to the case of disease virulence (in 2.4.2) or a tragedy of the commons among the fish. The fish, may co-operate with each other, not to destroy the resource of cleaner shrimps.

- Emergentism
- Neo-group selection.

The half-camp being halfway between the two : a combination of both emergentism and neo-group selection.

### 2.4.1 Emergentism

Emergentists have a project of criticising the atomicity of traditional models of altruism. Altruistic behaviours, they argue, shouldn't be seen as single actions, but as the results of multiple micro-behaviours.

For example, Barbera Hemelrijk who is on a crusade to debunk many of the naive attributions of cognitive faculties to animals in dominance hierarchies, has built several models to examine the micro-behaviours underlying one apparent form of altruism. In primate dominance hierarchies, competitions between two conspecifics, are sometimes interrupted by a third primate, who is seen as helping one or other of the protagonists. Later, the helped individual, is seen reciprocating the favour. This has lead observers to the conclusion that these primates have a sophisticated model of the social world in their heads.; and that they keep tally of favours owed.

As the form of the support appears to be aggressive approaching, Hemelrijk has constructed a model[20] where agents, wandering in a space, are drawn towards others within their field of immediate vision, and will be drawn into aggressive approaching when too close to another conspecific. In her models, the agents have no conception or recollection of third party behaviours *at all*. Nor, is there any objective, innate notion of superiority.

What happens, to establish the dominance hierarchy, is that agents who win a dominance competition, re-enforce their own aggressiveness. And more aggressive agents are both more likely to engage in further competitions, and win those they get into. Dominant agents, therefore, start on their road to dominance by being fortunate winners at a time when all are roughly evenly matched; then, drunk on the wine of their own success, are likely to pick, and win, more fights. Dominants, remain dominant, because they find themselves as the nexus or hub of multiple dominance showdowns.

In early experiments, this seems to have been literally by being in the spatial centre of the group[7]. Helping, by third parties then, is just a side effect of those third parties becoming caught up in a nearby challenge. As the more dominant individuals seem to spend the most time brawling in the town centre, it is not surprising that they also get tangled up with each other, and help each other out more often, while those, avoiding fights on the periphery, are also less likely to pick up local support.

While Hemelrijk's models are compelling, one can still feel that they have somehow missed the point in the discussion of altruism. Particularly so when she criticises the use of the Prisoner's Dilemma as the method of investigating altruism. Firstly, Hemelrijk's is research is firmly anti-adaptationist. She doesn't model evolution, and she says nothing about the value of the behaviours

[7] Though later Hemelrijk claims even physical space is not necessary[21].

(whether micro or emergent.) She can be seen as giving a proximal explanation of altruistic behaviour, ("how it works") without any kind of "why". Or, to phrase it another way, the only explanation of "why altruism" is that altruism just emerges from those micro-behaviours.

But nothing seems to be at stake in this supposed altruism. The helpers don't donate fitness to the helped. Whilst we see altruism emerge from the micro-behaviours, there is no discussion of cheating or defecting. Why wouldn't this behaviour be invaded by non-co-operating creatures? Because the micro-behaviours that would add up to defection can't exist? Because, in fact, this is a case of mutualism where there is no temptation to defect?

In conclusion, Hemelrijk's emergentism is a salutary warning to those who observe animals, not to jump to cognitive conclusions. But actually it is no answer to the question "how can altruism exist given Darwinian presumptions?", unless we are radically willing to redefine altruism, not in terms of its fitness consequences, but simply in terms of a range of observed behaviours.

### 2.4.2 Neo-Group Selection

Neo-group selection *is* still within the adaptationist program. Its revival involves two ideas. A discovery that there *are* cases where Williams's compelling argument against group selection, that the selective force on individuals, operates more swiftly than selection on groups, can be wrong.

An intuitive example is virulence of diseases. More virulent diseases, that reproduce faster in the host, will also kill the host faster; often before there is a chance to infect the next victim. For this reason, while the fastest breeding, most virulent strain, ought, by individual selection, invade the sub-population of bacteria within the host, it is those sub-populations, uninvaded by such mutants, who will conquer the next victim.

Neo-group selection therefore compares the speed at which deleterious mutations arise within sub-populations, with the rate at which new sub-populations are spawned, and the exchange of individuals between them. Sub-populations don't need to be as extremely separated as those of bacteria within host organisms. Any degree of *viscosity* - that is tendency of agents to spend their lives within one locality, interacting with the same conspecifics - can lead to pockets of greater altruism. Sometimes, the benefits of this are attributed purely to kin oriented effects, where agents find themselves mainly in the company of family members. However, Di Paolo has experimentally demonstrated that viscosity can encourage co-operation even when kin selection theory would predict against it[33].

But neo-group selection can also threaten to absorb both kin selection and examples of reciprocal altruism - within a unified theory, where families and pairs of reciprocating co-operators become simply special cases of sub-populations for who mutual co-operation is fitter, overall, than mutual defection.

At the beginning of this section I, jokingly, described the new critiques as falling into two and a half camps. The half is actually the large amount of work that combines these neo-group selection with a simulation work. As Di Paolo points out, here you can find demonstrations that pockets of altruism survive because of the discrete nature of the models. Continuous models would predict infinitesimal amounts of defection to emerge, which would steadily invade the population.

But with discrete, agent based models, such infinitesimal amounts might never be translated into actual behaviours. Discrete and continuous models of my own described in 5.3 show similar effects.

### 2.4.3 Conclusion

To sum up this chapter. We use evolutionary theory to explain the existence of biological phenomena including behaviours. The standard explanation being to show that the behaviour increases the fitness, defined in terms of number or quality of offspring, of the animal performing it.

However, a problem to this general theory is posed by the general class of *altruistic* behaviours, which donate fitness to another animal, at some cost to oneself. Traditional evolutionary theory has four explanations for this :

- *mutualism*, where both parties gain in absolute terms from the apparently altruistic behaviour;
- *kin selection* where an animal donates fitness to a relative to ensure the survival of the genetic material shared with that relative;
- *group selection* where some sub-population containing altruists does better overall, than rival sub-populations of purely selfish individuals; and
- *reciprocal altruism*, a special class of mutualism, where conspecifics could cheat and take advantage of each other's altruistic acts, but don't due to some mechanism that identifies and discriminates against cheating.

Another type of investigation, *emergentist* exemplified by Barbera Hemelrijk was noted, although I raised doubts as to whether it really addressed the same problem; in that it was concerned with a different notion of altruism. Where emergentist investigations do address the fitness definition of altruism, and where it is combined with a sophisticated understanding of groups and population characteristics it can explain altruistic behaviour.

We have not yet addressed ourselves to considering our real quarry : *individual recognition*. In the next chapter we will do so. There we will also see that an understanding of the problem of altruism is valuable background.

# Chapter 3

# Biological Approaches To Individual Recognition

## 3.1 Introduction

This chapter looks at some examples of the discussion of individual recognition within the biological literature. We will see that recognition is usually identified with some sort of observable discriminatory behaviour rather than as a cognitive faculty. We will also see that actual examples of individual recognition are highly specific to a a particular activity or situation.

## 3.2 Types of Recognition

Asking what it is to recognise something is tantamount to asking what it is to know something : a problem which goes beyond the scope of this thesis. In the biological literature, recognition is normally equated with discriminatory behaviour. This is the line taken in a useful overview by Paul Sherman, Hudson Reeve, and David Pfennig[42]. they state that "although [discrimination and recognition] are not synonymous when recognition refers to an internal neural process that underlies, but can occur without, detectable behavioural discrimination ... if discrimination never occurred, recognition ... would be an empty concept."

Sherman et al distinguish types of recognition by the object of discrimination. Hence "*kin recognition* is differential treatment of conspecifics (including self) differing in relatedness." and a slew of other recognitions : *species recognition*, *sex recognition*, *mate-quality recognition* and *mate-resource* recognition - which add up to *mate recognition* which they conceive of, not as mate re-identification but as identification of a potentially good mate.

Given this behavioural categorisation of recognitions, it is hardly surprising that the more problematic *individual recognition* isn't considered in the article. But it has been discussed elsewhere within the biological literature. In a letter to the Journal of Theoretical Biology, Michael Breed and Marc Bekoff[5] define individual recognition thus :

"Individual recognition involves the ability to perceive, to process, and later to use characteristics of another member of a population to discriminate that individual from other members of

the population."

They continue with some thoughts on constraints : "Since learning is often a function of the period of time and context in which the item to be learned is perceived, repeated inter-individual interactions might be expected to occur before complete recognition is possible. In an infinitely large population an infinite number of recognition cues would theoretically be required for each individual."

Researchers in the wild must be careful that the mechanism they suppose underlies individual recognition must be capable of supporting a large enough number of discriminable states, to represent the different individuals. For example, researchers into individual recognition in birds will provide acoustic analyses to show that the calls can contain sufficient information.

"Thus individual recognition, like other population phenomena must be thought of in probabilistic terms; the probability of discriminating one individual from any other individual is determined by (1) the discriminating individual's knowledge of the individuals that it is encountering and (2) the number of available recognition characters that can be used and the mean number of states that can be discriminated. It should be noted that by using this definition, individual recognition can be discriminated from group recognition ... [given] the probabilistic nature of the system, we would expect in practice that animals may make *mistakes*, because individuals will overlap their characteristics."

Note particularly the last part of the quote. That true *individual recognition* can be distinguished form mere *group recognition*, largely by whether the medium of *recognition cues* is sufficient to discriminate every individual.

This definition is given in the middle of a debate between Breed and Bekoff, and C. J. Barnard and Theodore Burk[3]. In the view of Barnard and Burke, animals are evolved to react to particular *classes* such as the class of dominant or subordinate conspecific, the class of kin, etc. The recognition of membership of all these classes is pattern matching a set of cues to some greater or lesser extent. But, they stress that animals will recognise only those classes, which they term *assessment units*, that are ecologically significant. If it is a good idea to recognise kin, animals will do so.

In their scheme, individual recognition should be no different. Individuals are a particular assessment unit, or class, to be distinguished. There is a continuity between individual and class recognitions; and perhaps even an evolutionary trajectory through increasingly refined and accurate discrimination. Barnard and Burk are led to their position by contemplating the question of individual recognition in dominance hierarchies. If an increasing refinement of recognising dominant quality within dominance hierarchies, leads to a fine tuned perception of exactly where one's own position is, this might look to an outside observer like individual recognition.

But Breed and Bekoff want to deny this continuum, to deny that individual recognition is just a highly refined perception of this sort. They do so by bringing examples that are either *not* from dominance hierarchies, or if they are, break the proposed scheme. These include

- male sweat bees that remember whether they have mated with a particular female or not and

mate less frequently with her as familiarity grows (though remaining equally active with new female bees);

- rejection, by workers, of a new queen replacing the old queen;
- cockroaches, preferentially orienting to a familiar, dominant conspecific, but not to an unfamiliar one;
- partner recognition among lemmings.

However, in their response. Barnard and Burk simply define kin as another sort of class and re-iterate the claim. There is a general sort of thing which is an assessment unit. When it is ecologically significant to discriminate who has membership of the assessment class, we can predict that that animal will have that behaviour.

This idea of *unit of assessment* is the valuable one to take away from this discussion. This is another way of stating Sherman et al's definition of the type of recognition by the class of things discriminated. But it enriches it. What is recognised is not just what we observe being discriminated, but from an ecological perspective, what would it be fit to discriminate."

The other issue raised is whether individual recognition *can* be thought of as a refinement of a class recognition. For Breed and Bekoff the question as to the distinctness of individual from class recognition is merely one of the capacity of the discriminatory mechanism. If it is of high enough resolution to distinguish particular conspecifics, the animal has individual recognition; if not, it has merely recognition of particular classes. In contrast, by tying the notion to ecological significance, Barnard and Burk allow that two animals could have identical discriminatory mechanisms but one, merely for the sake of group recognition while the other uses it for individual recognition. Barnard and Burk bring purpose into the discussion. They also, seem to imply that there is a continuum of ecological significance between class and individual recognition.

Does this idea make sense? Or rather, from a Darwinian perspective, we imagine that, of course, the *perceptual mechanism* of individual recognition, for example, hearing and processing of audio signals, has evolved through a process of continuous refinements from some proto-hearing capability. Is this the same as a continuum of ecological significance? I would guess not, if an ear that was part perfected to detect and distinguish predators, later became used to distinguish offspring in the nest. Hence it could just be argued that by definition, when the unit of ecological significance stopped being a class and became an individual the continuum changed.

## 3.3 Where individual recognition might occur

- Mate, parent and offspring recognition in birds
- Dominance hierarchies from lobsters to primates
- Reciprocal altruism in mammals

### 3.3.1 Mate, parent and offspring recognition in birds

Research into mate, parent and offspring recognition has covered a wide number of species. Individual recognition is often found to be the result of highly specific circumstances. For example, many colonial sea birds have finely adapted sensitivity to individual calls [23][43] where there is a need for returning, foraging parents to identify mates and chicks in densely crowded nesting sites. On the other hand, in non-colonial species, where parents can reliably know their offspring as being those in the home nest, parental recognition of offspring can be weak or non-existent[26].

### 3.3.2 Dominance Hierarchies

The behavioural notion of dominance has been investigated since Schjelderupp-Ebb described the peck order in the early part of the century. A ranking of individuals, known as a *dominance hierarchy* is established through aggressive display and competition between conspecifics. Higher ranking individuals receive a larger share of scarce resources. Ranking is established by contests between pairs, although Wilson[50] notes that individuals are sometimes supported by allies in these contests.

Often the contests take the form of aggressive signals. There is also an apparent difference between those contests that establish the initial rank order and further contests that maintain it. In other words, individuals seem to remember who has won previous contests, and who is dominant.

If this observation is correct then there is a possible function for individual recognition in maintaining these dominance hierarchies. Wilson again suggests that the social wasp Polistes might use individual cognition for this purpose.

In the literature there is some controversy about whether dominance requires or implies individual recognition. With some authors presuming that it does, while others suggest that during the first stand-off, an animal might simply learn its own status, or to better correlate an opponents aggressive signals with the likelihood that it will engage in a fight.

Answering this question is confused by the fact that there are disagreements about what dominance actually is. Carlos Drews [12] has made a survey of the rival positions, taxonomized them and made an attempt to distill a reasonable definition. On the issue of individual re-identification Drews has this to say : “Memory and individual recognition are implicit in the “peck-order” definition as proximate mechanisms to explain the deference behaviour. These mechanisms may apply in some cases but need not be necessary for the consistent deference behaviour to be observable. Mechanisms and function should not form part of the definition ... [but be][1] ... used instead within hypotheses concerning the causation of dominance relationships.” Hence, dominance should be defined as a kind of behaviour of dyads, making no claims about mental innards criteria for either agent.

In other words, it isn’t part of the definition of a dominance hierarchy that it is enabled by individual recognition. That leaves it an empirically open question as to whether a particular dominance hierarchy is enabled by individual recognition.

[1] The paper says “...part of the definition and used instead ...” which I take to be a misprint.

Work such as that by Christa Karavanich and Jelle Atema with lobsters [24] exemplifies how the question can be tackled experimentally. They find that newly introduced lobsters will fight for dominance, after which the loser will defer, by backing away, to the winer. The loser seems to have no tendency to similarly defer to unknown lobsters, regardless of other observable characteristics, suggesting that the loser has learnt to recognise the individual opponent. Two further interesting results were obtained form this work. The first is that the ability to remember opponents was not disrupted by encounters with other lobsters, suggesting that the lobsters were not overwhelmed by having to remember multiple partners (though the authors admit that this was not an explicit result and that further experiment was necessary.)

The other, was that memory faded over a period of 1 to 2 weeks of two lobsters being separated. After one week separation, 7 out of the 10 experimental subjects seem to have forgotten previous competition, and will challenge. 3 of the experimental subject retained their subservience. After a two week separation, previous losers showed no signs of deferring and all pairs fought again. In no case did the the previous loser now triumph, suggesting that the renewed fighting was not due to any perceived change in status by the lobsters. Instead, we can conclude that the dominance ranking had simply been forgotten.

Recent work, again by Barbera Hemelrijk, with Christoff Goessman and [15] demonstrates that certain kinds of dominance hierarchy in crayfish can be sustained without individual recognition. It seems that her model doesn't capture the notion of a time of memory, which is so suggestive in Karavanich and Atema. As with her earlier work, Hemelrijk's model does feature some memory, that of personal aggressiveness. It is plausible, though not tested, that were she to introduce forgetting of personal aggression level into her model, she might be able to reproduce the all the behaviours of Karavanich and Atema's lobsters.

### 3.3.3 Reciprocal altruism in mammals

Although Trivers wrote the classic paper that introduced reciprocation to the altruism debate his examples, such as fish-cleaning shrimp, are not obvious places to look for a more full blown individual recognition. A paper by Laela Sayigh, Peter L. Tyack, Randall S. Wells and others[38], explicitly suggests that true individual recognition underlies reciprocal altruism in monkeys (citing Seyfarth and Cheney[37]), bats (citing Wilkinson's famous research on food sharing.[47]) and dolphins.

Other mammalian individual recognisers include Rattus norvegicus in laboratory conditions to recognise individuals by their odour[14]. Burda[7] describes experiments with eusocial mole-rats, where individual recognition, rather than some kind of parental oppression seems to enable incest avoidance between siblings.

### 3.3.4 Conclusion

The intuitions that underlie the biologist's notion of *individual recognition* are not mine. Ethologists must decide on their balance between defining activities in terms of observables, and where,

*hidden terms*, or *cognitive innards* may be inferred. The biological literature such as Breed and Bekoff's discussion of individual recognition and Drews's discussion of dominance reveals that biologists are often in disagreement over such attributions.

What is also clear is that nature provides some striking examples of the circumstance specificity of apparent individual recognition. Even closely related species such as bank and barn swallows can have different behaviours, one which involves apparent individual recognition, the other of which doesn't. Faced by this, it is plausible to assume that what might appear as an example of a more general cognitive capacity, is in fact an extremely activity specific mechanism which will fail in any abnormal circumstance.

After such a survey, reciprocal altruism starts to look like the best bet (or last resort) for finding a general individual recognition that corresponds more to the notion indicated by Munitz in the introduction.

Individual recognition in dominance hierarchies is also problematic. Hemelrijk has raised serious doubts. In the previous chapter I found her critique of *reciprocal altruism* in dominance hierarchies missed the question about altruism. But, her more general demonstrations that dominance hierarchies can be maintained by a learned, internalised status of aggression rather than individual recognition, are compelling. We can not rule out the possibility that Karavanich and Atema's lobsters genuinely recognise and remember each other. But Hemelrijk produces sufficient evidence that similar hierarchies can exist without it that we should be suspicious,

The discussion between Breed, Bekoff, Barnard and Burk seems to indicate deeper conceptual issues. To expect animals to have a more general notion of individual, they must live in a world for which that notion of individual has ecological significance. The reciprocally altruistic animal, involved in social contracts and reciprocal relationships seems the closest to one living in a world where individuals have ecological significance. But I am forced to confess that the picture is more dependent on the abstractions from game theory and social science than drawn from the biological evidence discussed. There are many cases of applying the prisoner's dilemma to natural examples, but increasingly biologists are taking a more sceptical view of the Trivers's story and considering the group selection and mutualism alternatives more seriously.[13]

A more positive way to look at the situation is that the paucity of biological material is also an opportunity for researchers with a more cognitive perspective. We know from our own experience that *at some point* the notion of re-identifiable individual appeared; and if biologists findings have thrown no light on the subject then we must seek alternative inspiration.

# Chapter 4

# Computer Based Research of the Prisoner's Dilemma

## 4.1 Introduction

In the previous two chapters we looked at the suggested explanations for altruistic behaviour and recent biological work on individual recognition. I hope by now the reader will have followed me to the point where it is plausible that one should go on to research individual recognition in the context of reciprocal altruism and hence a prisoner's dilemma type situation. That is not to say that there aren't other circumstances where the category of *individual* might not be ecologically significant. But the other cases considered from the biological literature, all seem to lead to less interesting or more controversial notions of individuals. Nor is it to say, as we will particularly see in this chapter, that a strategy for playing the IPD that involves individual recognition is either optimal or the only one that leads to co-operation.

Reciprocal altruism and the prisoner's dilemma game have been widely studied. A famous and seminal work is that of Robert Axelrod, described in his book "The Evolution of Co-operation"[1], with which I'll begin this chapter. Axelrod is famous for both running automated tournaments; and for running a simple evolutionary simulation where successful strategies spread to dominate an initially mixed population.

Having discussed Axelrod, I will then briefly review some other categories of work which have been influential on my experimental designs, in particular :

- models with memory,
- viscous and spatialised populations, and
- models with kin.

## 4.2 Axelrod demonstrates the fitness of tit-for-tat

For his series of famous experiments, Robert Axelrod invited researchers from several disciplines to submit computer programs for playing the IPD; each pair of strategies was matched 200 times,

sequentially with the pay-offs below.

| * | C | D |
|---|---|---|
| C | 3 | 0 |
| D | 5 | 1 |

Table 4.1: Axelrod's Scoring Matrix

The winning entry was the standard Tit-For-Tat(TFT) strategy submitted in both tournaments by Anatol Rapoport. The TFT strategy consists of co-operate on the first move, then doing back to the opponent what that opponent did in the previous match.

In a second tournament, where programmers of strategies knew the results of the previous tournament, and where there was an indefinite number of matches between pairs, TFT also won.

TFT is not guaranteed to be the highest scoring player against any particular opponent. Always defect (ALLD), for example, will defeat it as it gets one defect against a co-operate on the first move. However, when both TFT and ALLD play against a third strategy, TFT often gets into a virtuous circle of mutual co-operation and scores higher in total.

From the results of the experiments, Axelrod diagnoses several properties that tend to mark out high scoring strategies in the tournament situation and shows that TFT has all of them.

These properties, which I will call the *Axelrod diagnostics* are

- *niceness*
- *provokability*
- *forgiveness*

A *nice* strategy is one which does not defect first. It may defect, once it has been betrayed, but will not initiate defection.

A *provokable* or *retaliatory* strategy is one which is willing to defect, once defected against.

A *forgiving* strategy is one which can return to co-operating with an opponent, once that opponent has signalled it wants to return to co-operation and possibly paid a requisite forfeit.

TFT has all these properties and Axelrod sometimes *explains* the success of TFT in terms of them.

### 4.2.1 Explaining the success of tit-for-tat

Attempts to improve on TFT fall into two classes :

- More forgiving
- Taking advantage of opponent modelling

At the end of the first tournament, it was noted that sometimes TFT would meet a strategy which attempted to get away with a defection every now and then; having been punished by TFT;

would revert to playing TFT itself, but out of synchronisation with TFT, hence leading to an "unnecessary" vicious circle of defection. It was demonstrated that a more forgiving strategy such as Tit-for-two-tats (TF2T) - which waits until an opponent defects twice before retaliating - would actually score better than TFT in an environment of these. TF2T thus reduces provokability. However, overall, such strategies ultimately did worse in the second tournament. Any extra leniency they provided was itself open to exploitation by other strategies.

Another attempt to improve on TFT are strategies which try to use some analysis of the character of the opponent to make predictions about what that opponent will do next. Unfortunately for such strategies, they are often defeated by players that co-operate just over 50% of the time, thus establishing the appearance of a co-operating profile; while getting in several defects.

Axelrod posited that the right balance of the properties of *niceness*, *forgiveness* and *provokability* make for a strategy which is, in Axelrod's terminology, *robust*. Such a strategy will be good against a range of different players including itself. A strategy which scores highly with itself, makes for a fairly stable ecology if it finds itself forming one.

Axelrod followed his two tournaments with a simulation of, as he put it, future tournaments. Assuming that those strategies which did the best in the second tournament would be resubmitted in larger quantities, he produced a model that he carefully makes clear is *ecological* but not *evolutionary*. "This simulation provides an ecological perspective because there are no new rules of behaviour introduced. It differs from an evolutionary perspective, which would allow mutations to introduce new strategies into the environment." [2]

What *can* happen in an ecological simulation is that the proportion of strategies in the population changes; and some strategies are driven into extinction. In this simulation Axelrod manages to show that the one non *nice* strategy which had been reasonably successful in the second tournament, by exploiting more forgiving strategies, first thrived, but then drove its prey to extinction, before becoming extinct itself in the harsher environment.

From this, and the fact that the prisoner's dilemma is not a zero sum game, he concludes that part of the success of TFT is that it does not try to beat its partner. Any player it plays with does either as well as, or better than itself. However, beating the opponent is not the key to success in the long run.

## 4.3 Axelrod and Hamilton : an *evolutionary* discussion

Axelrod's book contains one chapter co-written with the evolutionary biologist William Hamilton. In it they point out how the research contributes to the evolutionary use of game theory.

- "In a biological context, the model is novel in its probabilistic treatment of the possibility that two individuals may interact again."
- "The analysis of the evolution of co-operation considers not only the final stability of a given strategy, but the initial viability of a strategy..."

Axelrod and Hamilton are clearly making a claim for the advantage of bottom up, individual oriented investigation of the dynamics of evolving a behaviour, compared with top down, analytic

mathematical models traditionally used by theoretical biologists. One particular way their notion of evolutionary model goes beyond mathematical analysis.

The usual game theoretic analyses concentrate on finding the *evolutionarily stable strategies* (ESS)[44]. The evolutionarily stable strategy is described by John Maynard Smith as one where "if all the members of a population adopt it, then no mutant strategy [can] invade the population under the influence of natural selection".

Axelrod and Hamilton's evolutionary investigation identifies three qualities in a strategy, when considering it in an evolutionary perspective.

- robustness,
- stability, and
- initial viability

Here *robustness* is success in a variegated environment; *stability* is success once established in the face of new 'mutations; and *initial viability* is success when rare.

### 4.3.1 Binmore's criticism of *niceness*, *provokability* and *forgiveness*

Ken Binmore[4] has criticised Axelrod for leading too many people into the easy assumptions that the iterated prisoner's dilemma is the correct way to model the evolution of co-operation. He also has some specific attacks on Axelrod's conclusions.

According to Binmore, among the the 63 strategies originally tested by Axelrod, TFT can not be an ESS because it is not a Nash equilibrium[1]

In work cited by Binmore, all possible 1 or 2 state strategies have been considered and the society converges on the non-forgiving GRIM - which starts as TFT but once defected against becomes ALLD - which is played by half the population. GRIM has no forgiveness. Yet in a deterministic world of ALLD and TFT it plays exactly as TFT and can form a highly co-operative society. Forgiveness is not necessary for co-operative societies, nor must the punishment fit the crime.

The success of the strategy Tat-for-tit - aka PAVLOV and a simple example of what Binmore calls a mean machine - challenges the assumption of the necessity of *niceness*. Tat-for-tit starts by defecting, and then, changes whenever the opponent defects. When playing itself, it starts with mutual defection, before settling into mutual co-operation. Therefore niceness is also unnecessary to a society negotiating its way to co-operation.

Note also that Tat-for-tit can beat tit-for-tat. If the pair play an odd number of matches, PAVLOV is one match up against TFT. In an even number of matches, TFT will equalise. This means TFT can not invade a society of PAVLOV. Does this mean that, in Axelrod's terminology, TFT is not initially viable? Theoretically, a society of PAVLOV can't be invaded by TFT. However PAVLOV itself is *not* initially viable against a population of TFT players. Although it will beat every TFT player it meets, because of the $2R > T + S$ constraint, its score against TFT is

[1] Where no player can improve its score by changing strategy.

lower than the score between two TFTs. In a population dominated by TFTs, these players will be scoring more highly with each other. PAVLOV is also not robust in the variegated environment. It is soundly beaten by ALLD.

What should we learn from this? I think the lesson is the distinction between what Axelrod calls *stability* and *robustness*. It seems as though *stability* is very close to *evolutionarily stable strategy*, And this second term has a formal definition. Binmore's strict application of that definition shows that neither *niceness* nor *forgiveness* are necessary. But *robustness* is a different matter. It is an idea that goes beyond ESS. While the ESS can not be invaded by a single mutant strategy, it can be brought down by the right combination of opponent strategies, *robustness* is a statistical quality, a resistance to many combinations.

### 4.3.2 How might TFT have begun to invade a stable ALLD society?

Axelrod considers that the two tournaments and his ecological simulation have demonstrated that TFT is particularly *robust*. It is also a *stable* strategy though not the only one as Binmore has shown. What about *initial viability*?

We can take Axelrod's diagnostics as sub-behaviours out which TFT can be constructed. Therefore the evolution of TFT requires the following traits to be evolved.

- The actual co-operating behaviour itself
- Individual Recognition
- Niceness
- Provokability
- Forgiveness

I will say little about the co-operating behaviour itself. This is something which will differ from case to case. An emergentist study, such as those of Barbera Hemelrijk that we looked at in 2.4.1, might focus on how the earliest co-operating behaviour got started; and could maybe suggest reasons *why* this forms of altruism rather than that. But from our perspective, such behaviours only become interesting when they have fitness consequences that fit the pattern of the prisoner's dilemma.

For Axelrod, individual recognition is also not explicitly considered. From our perspective, the growth of individual recognition, from an infinitesimal proto-individual recognition, to the full capacity is interesting. I will discuss this issue fully in 5.3.3.

This leaves the niceness, provokability and forgiveness.

Two hypotheses are introduced by Axelrod and Hamilton. One, flagged in Trivers' original paper, is kin oriented altruism; the other is through mutation within a small, tightly interacting sub-population. In other words a form of group selection.

### 4.3.3 Kin oriented altruism

One suggestion is that kin selection created the background of co-operation against which reciprocal altruism can get started. Once that existed, slight tweaks to the response of the player could have various effects.

Axelrod and Williams suggest that *niceness* would be the default behaviour towards those recognised as kin.

But, if a player mistakenly co-operated with a non-relative, that other, gifted perhaps with better eyesight, would defect. If the first player failed to treat this as a corrective lesson in the unrelatedness of the two individuals; he would be penalised. Thus provokability would be selected for.

Once niceness and provokability were up and running, the scaffolding would be in place for improvement in re-identification; perhaps as a refinement of the discrimination mechanism used to distinguish kin from non-kin, or perhaps as a refinement of another perceptual mechanism.

The final stage would be reached when, unrelated individuals meeting each other repeatedly, but now imbued with the virtues of niceness and retaliation could also take advantage of the virtuous circle of co-operation. Hence, there would be the possibility of relaxing the need for strategy to refer to kin.

This seems quite a specific prediction of TFT evolving from a background of kin oriented altruism in four steps :

- A rise in niceness proportional to an error in accurate kin recognition.
- A rise in provokability.
- A rise in the recognition of individuals.
- And finally niceness increasing regardless of kinship.

### 4.3.4 Small groups in close proximity

As a rival to the kin-selection theory we have some form of group-selection theory. As pointed out previously2.4.2, in spatialised or viscous populations mutant co-operators can survive. This requires at least two co-operating mutants, whose existence is fortuitous.

## 4.4 Recognition and Memory in Axelrod's research

Axelrod is certainly aware of the importance of recognition; as a means to increasing co-operation, he suggests trying to improve recognition of both individuals and acts; and points out the failure of co-operation that results from the failure of recognition. He also notes that birds recognise songs, as one of the skills they have to maintain territoriality. Interestingly he points out that this "allows them to develop co-operative relationships - or at least avoid conflicting ones". Axelrod does not

discuss this further but certainly considers that the prisoner's dilemma game is abstract enough to model co-operative behaviour such avoiding aggression.[2]

Apart from stressing the virtue of recognition, Axelrod's work implies no theory of what recognition is or how it works. It seems likely - though Axelrod doesn't explicitly state it - that all matches between a pair were played in one session. In other words, player X met player Y; played all the matches; then broke off; never to meet again. Such players would never have to remember a previous partner from previous games.

Under such conditions, one notion of "degree of recognition capability" that makes sense, is the length of the historical sequence of moves handled by the program'. The contributers to Axelrod's tournaments were able to submit programs with any sort of architecture. These programs received accurate information about the match that had just transpired; and could store as much history as required. For example, NYDEGGER, is a strategy using a three step record; and DOWNING tried to learn the character of its opponent through building up a probabilistic model based on all of the previous interactions.

## 4.5 Kristian Lindgren : Evolutionary Phenomena in Simple Dynamics

In Artificial Life 2, Kristian Lindgren presented a paper entitled Evolutionary Phenomena in Simple Dynamics[27], describing the evolution of a population of IPD players. The focus of interest was the dynamics of *open ended* evolution; where, in contrast to Axelrod's ecological model, novel strategies could arise due to mutations in a variable length genotype, which coded for ever increasingly complex strategies.

Although not ostensibly a research into the evolution of the ability to re-identify, the dimension of complexity that was allowed to evolve was the length of memory of the sequence of interactions. In the model, the behaviour of a player is determined by the previous $n1$ moves by the same opponent; and $n2$ moves by that player against that opponent. As the genotype grows, so does the number of previous moves taken into account. Consequently, there is no fixed set of possible strategies. The number of possible strategies, like the genotype, is potentially infinite.

The possible mutations in Lindgren's genotype include point mutations at a particular locus, and both doubling and halving the length and content. For example, doubling a genotype 10 gives $1010$[3].

Another feature of Lindgren's model is that the players evolve in a noisy environment. In such environments, Tit-for-tat has a problem. Whenever it plays another TFT player, one accidental defection can knock the pair into a vicious cycle of recrimination from which a more forgiving strategy might recover.

Strategies with one step memories have no escape from this. But a two stage memory strategy such as the less provokable tit-for-two-tats, which only defects against opponents who have

[2]This *is* actually explicitly spelled out in a discussion of the "live-and-let-live" behaviour that arose between the English and German armies in the trenches in the first world warfare. Nevertheless, as shown in the previous chapter territoriality is actually modelled differently.

[3]Thus the length of genotype grows exponentially from 1 to 2 to 4 to 8 steps described in the paper

defected twice in a row, will survive.

### 4.5.1 Results of Lindgren's Model

Lindgren's model started by producing a typical populations of evolving IPD players. The population is, to begin with, taken over by ALLD which takes advantage of indiscriminate co-operators (and Lindgren's other one step strategy, an anti-Tit-for-tat that co-operates with defectors and vice versa). However, TFT, once started, can invade ALLD through being just as ruthless with ALLD but benefiting from co-operating with other TFTs.

But as TFT takes over, the population becomes more benign, allowing always co-operate (ALLC) back. And ALLC, in turn, lets in anti-TFT (ATFT). Once these are established ALLD can return. Consequently there is a cyclic waxing and waning of these strategies. But the oscillations eventually attenuate.

This is not the only possible outcome. Lindgren also discovers a stable mix of TFT and ATFT which cannot be invaded by any single mutation. Eventually, this stable mix *is* brought down by either a multiplicity of mutants or one of several longer memoried strategies.

In fact, over many generations, the apparent message of the model is that memory length continually increases. Or rather, longer memoried mutant strategies appear, destabilise the old order, and then establish their own. Only to be overthrown in turn by even longer memoried strategies. Lindgren follows this process from 1 to 2 to 3 and finally 4 step memories.

There seems to be no simple theory as to why longer strategies should proliferate. The examination of the dominant strategies reveals their success to be due to different factors. The dominant two step strategies are monolithic. But the 3 step memory world is dominated by two mutualistic interdependent strategies that score most highly when they play together.

Lindgren does not offer an explanation for this . But Anil Seth who has done similar experiments[40] suggests that there is a connection between the noise in the environment and the complexity of the evolved strategy. He showed that an increase of noise could overcome the inhibition that a cost placed on long memories.

## 4.6 Crowley et al : Evolving Co-operation : the role of individual recognition

A paper with a large number of authors but principally Philip Crowley, Louis Provencher, Sarah Sloan and Lee Alan Dugatkin[10] explicitly looks at re-identification and the evolution of individual re-identification. Their EvCo[4] model is derived from Axelrod's 1987 model. What is new, is that the memory capacity of the players is variable, and costly.[5]. The authors are primarily interested in the effect of re-identification *on* co-operation, so in most of the experiments, the degree of re-identification is fixed.

The EvCo model is complex and slightly surprising. As in both Lindgren's and Axelrod's later models, the players base their decisions on the preceding moves by both themselves and the

[4] *Ev*olution of a *Co*mbinatorial genome.

[5] Seth also introduces cost, but a year later.

opponent. However there is a discrepancy between how many preceding moves they try to use, and how long their memory is reliable. Memory can be of the full event : ie. the identity of the opponent; and the moves made; or just of the moves made. Where knowledge is not available it is provided by so called "virtual memory" : that is, a default history of moves provided by the initial assumption loci on the genome.

In the *no recognition* game, the player is provided with only the most recent moves made by *any* opponent. If there are not enough, eg. if the player has a memory capacity of 3 but has only played 1 previous game, or because the player strategy is based on 3 previous events, but player memory only has a capacity of 1, then the 2 missing experiences are provided by the initial assumption virtual memory.

In the *strong recognition* game the player is provided with a history of the previous interactions with this opponent. Once again, if demand outstrips capacity, the shortfall is made up from the virtual memory.

In the *weak recognition* game the player looks first for accurate event knowledge of previous encounters with this opponent; if it needs more, it turns to the memory of previous interactions with other opponents.

This may seem strange, but in this model, pairs of players engage in a number of matches sequentially, and have a probability of breaking off to find new opponents. (As opposed to being randomly paired with different opponents each round.) Consequently, there can be long unbroken sequences of interaction with the same opponent. When partners change infrequently, memory of previous moves without memory of previous player identity is the same. When the average pairing is long, partner recognition will only add a little value.

### 4.6.1 The results of EvCo

The EvCo team pick out as their response variables : mean fitness-per-interaction; percentage of pairwise interactions resulting in mutual co-operation, mutual defection; the co-operation-defection combination; and the evolved memory capacity.

Their first result shows a strong correlation between mean mutual co-operation and mean fitness. Co-operate / defect pairings are constant and relatively infrequent across all mean fitnesses. This is to be anticipated. One would not expect any pair to maintain a co-operate / defect relationship for any duration, as the co-operating with defectors strategy is particularly unfit. In evolved populations, such mixed interactions are only likely to occur when a player meets a stranger, or an unrecognised previous partner. C/D pairings occurred more frequently under weak recognition than strong.

The authors noted early in the results that the system seldom converged definitively. Periods of mutual co-operation would give way to periods of defection and vice versa. This fits my own experience of this kind of simulation.

Without recognition, substantial co-operation appeared only for high pairing continuity. In other words, co-operation was beneficial if the players played the same individual for a longish run. This is consistent with most people's intuitions about the iterative prisoner's dilemma.

Increasing memory length increased the amount of mutual co-operation, though there were diminishing returns for each extra unit. When the length of memory was under selection pressure only the most recent interaction seemed to be reliably selected for. This is quite compatible with Axelrod's original findings that a simple tit-for-tat is as good a strategy against a variety of different opponents as many more complex ones.

One surprise is that longer memory capacity was selected for under weak recognition than under strong recognition. Given the details of the EvCo model this might be less surprising. Under strong recognition failures of recognition are made up by default values which appear to be pretty much arbitrary. Under the weak recognition, the short-fall is made up from memory of real encounters with other partners. In the strong recognition model, any increase in the length of memory required will introduce more arbitrariness into the game; which can disrupt many reciprocal strategies. In the weak model, as long as pairing continuity is very short, or there is some similarity of players, the results of using this information will introduce less noise into the world.

The EvCo also puts a cost against memory length. As this is increased, both the focal event memory and the other event memory decrease, at what looks like the same rate.

### 4.6.2 Suggestions from Crowley et al

The most interesting suggestion from the EvCo paper is their general hypothesis about the evolution of what they call *social networks*. They conclude that co-operation will only happen when multiple interactions occur between the same partners. But they also posit that these multiple interactions between each partner must be interspersed with interactions with other partners due to patchy food distribution or predation. Recognition allows sequences of interactions to be broken off, while allowing them to be returned to at a later date. Thus mere pairings disappear in favour of social networks.

Where a pair play all their matches in an uninterrupted sequence, and where recognition of events can be detached from recognition of their perpetrators, evolution may favour event recognition without individual recognition.

This is an interesting dichotomy which could be further explored; especially if one were to investigate the evolution of a learning mechanism where the two might become detached. If there are real biological situations where pairing is continuous over several matches, with little changes of pairing, and some sort of reciprocal altruism takes place; we might consider the possibility of situation memory without individual recognition[6]. Compare also with Trivers' cleaning shrimp.

## 4.7 Conclusion

Each of the writers discusses the evolution of reciprocal altruism between agents and each notes the importance of recognition in forging reciprocal relationships. Lindgren, Seth and Crowley et

[6] A pure speculation. Consider a human tendency that an angry reaction can sometimes be displaced from the legitimate target onto another. This need not be explicitly selected for, but could be a failure of identification that is not worth correcting.

al all include some notion of degree of recognition capability; with Crowley et al going further in giving two degrees of capability. From their discussions we can begin to get some tentative predictions as to how reciprocal altruism evolves and how it relates to the recognition capability.

One striking thing that all have in common is reciprocal altruism's precariousness. Although Axelrod claims that TFT is robust; in the evolutionary simulations we often see a dominant culture of reciprocal altruism crashing back into defection.

### 4.7.1 Predictions

This is a list of predictions made by various commentators discussed.

### 4.7.2 Axelrod and Hamilton

TFT is a robust, stable and initially viable strategy

Reciprocation can invade a society of defectors in one of two ways :

1. Via kin oriented altruism
2. From small groups in spatial proximity

The kin oriented altruism story goes like this.

- The species evolves kin oriented altruism and therefore has the co-operative behaviour
- Mistakes in recognising kin lead to an increase in co-operative behaviours with non-kin; an increase in niceness.
- Non-kin will nevertheless take advantage of this; thus as these mistakes arise, mistaken individuals will be defected against.
- This being defected against will become another sign of non-kinship. So, a player with low kin recognition skill could supplement it with individual recognition to remember being defected against.
- As this occurs, the benefit of getting into virtuous relationships with non-kin defended by retaliation will be an incentive to abandon kin oriented altruism and move towards complete reciprocal altruism.

### 4.7.3 Lindgren

There are two glaring take-home messages from Lindgren's work. First, an apparently stable equilibrium can suddenly become unstable, and a period of rapid evolution take place - for purely endogenous reasons. There need be no asteroids, separating continents or the like. Second, he has a convincing demonstration of a ratchet of complexity. Each new period of stability is dominated by a more complex strategy.

But does it follow from Lindgren that all evolutionary histories will end in domination by complex things? Seth shows that noise increases the genome length. It is interesting that, contrary to the expectation raised by Axelrod, the TFT strategy does not re-take the dominant position in the society from the more complex.

### 4.7.4 Crowley et al.

What is new in Crowley et al?

- Two degrees of recognition: the length of memory and the none / weak /strong spectrum.
- Costed degrees of recognition
- The distinction between recognition of events and individuals.

Crowley's team rediscover the necessity of repeat pairings for the emergence of co-operation, with the twist that the pairings need also to be interspersed with others to avoid a strategy of remembering events but not individuals.

Co-operation correlates with re-identification.

They also noted that co-operation dropped as cost increased, though memory capacity fell off faster.

# Chapter 5

# Introduction to the Experimental Work

## 5.1 Introduction

The following chapters describe the experimental work of the thesis. This chapter highlights the aims of the research that steered particular design decisions. Then goes on to cover some preliminary investigations using simple models.

The next chapter describes in detail a more complex model.

I have focused on three topics :

- A look at the dynamics of the evolution of Individual Oriented Strategy
- A comparison of Individual Recognition and an Individual Oriented Strategy, with Kin Oriented Altruism and Kin Recognition
- A look at the effect on these patterns of spatial distribution

## 5.2 Topics of investigation

### 5.2.1 The dynamics of the evolution of Individual Oriented Strategy

Individual recognition is necessary for successful reciprocal altruism. But can we see individual recognition as being in some-way "caused" by the necessity of reciprocal altruism. Talk of such causes is highly problematic, nevertheless what would be interesting would be to show a rise in some other trait or behaviour that corresponded with a rise in individual recognition.

In the previous chapter we looked at some suggestions as to the evolution of co-operation from Axelrod. Such an evolution involves both the putting together the Tit-for-tat strategy from sub-behaviours such as *niceness*, *provokability* and *forgiveness*; and the growth of individual recognition. Because I conceive of individual recognition itself as developing I have been interested in modelling it as a matter of degree. This will be discussed in 5.3.3 below.

Questions :

- Does the rise of Individual Recognition have an interesting character?

- Can we correlate Individual Recognition growth with anything else? Can we find good predictors of Individual Recognition?

### 5.2.2 A comparison of Individual Recognition and an Individual Oriented Strategy, with Kin Oriented Altruism and Kin Recognition

There are several reasons for being interested in Kin Oriented Altruism as part of the individual recognition story. One of these goes back to a suggestion by Axelrod and Hamilton in the previous chapter that Kin Oriented Altruism helps get Reciprocal Altruism going.

One of the most interesting results from a model that includes both Kin Oriented Altruism and TFT would be the light it could throw on this story. Is it possible to show Axelrod's story occurring? That is : errors in Kin Recognition leading to an invasion by TFT which would not otherwise occur. If not, can one at least show that a level of KOA can act as a catalyst for TFT to start?

Three explicit questions will be asked :

- Does a mix of Kin Oriented Altruism improve the chances of TFT taking off?
- Does it do this through failure of kin recognition? item Can we see Axelrod's story in action?

### 5.2.3 The effect on these patterns of spatial distribution

Researchers have long known that co-operation can occur and survive in spatial world or viscous population due to group selection. However, many of these researchers have been focuses on *co-operation* rather than TFT. Hence often the co-operation considered is *indiscriminate*. Axelrod and Hamilton have also considered the possibility that co-operation due to spatialization was the prerequisite. But if such worlds allow indiscriminate co-operation, do they still encourage discrimination (and hence individual oriented strategy) or do they, in fact, discourage it?

My question about space is

- We believe that spatial and similarly restricted versions of the prisoner's dilemma increase co-operation. But does this imply a stronger likelihood of the evolution of individual oriented, individual recognising strategies or does it just imply more indiscriminate co-operation?

## 5.3 Experimental design

### 5.3.1 A note on terminology

In the discussion chapter 10 I give an explanation of my understanding of the scientific status of Artificial Life and the role of computer programs which implement virtual populations. There is no need to preview the discussion here, except to note that in this chapter I will use two terms :

"*model*" and "*simulation*" in a fairly free sense. I will refer interchangeably to "my model" or "my simulation" when referring to one of the programs.[1]

Typically I will use "*run*" or "*experiment*" loosely to refer to one or more executions of the program.

In addition I will use

- *TFT* as shorthand for the tit-for-tat strategy;
- *ALLD* as shorthand for Always Defect;
- *Individual Recognition* or *IRec* to refer to capability of remembering individuals.
- *Individual Oriented Strategy* or *IOS* to refer to any strategy (including TFT but not ALLD) which differentiates between opponents upon the basis of their individuality - and hence requires *individual recognition* (*IRec*) to operate.
- I'll use *Kin Oriented Strategy* (*KOS*) to refer to any strategy that differentiates between opponents on the basis of kinship.
- *Kin Oriented Altruism* (*KOA*) will refer to co-operation that is the result of kin oriented strategy.
- I will use *Kin recognition* (*KRec*) to refer to the capability that allows discrimination of degree of kinship.

Kin oriented strategy requires a reasonably successful KRec to operate correctly. Individual oriented strategy requires reasonably successful IRec to operate successfully. However, neither IOS nor KOS should be taken as *success terms* which imply that these recognition faculties *are* operating correctly. In other words it makes sense to talk of a player using a kin oriented strategy even though it has lousy kin recognition. This means that it uses kinship knowledge to make decisions, even though the knowledge it bases the decisions on is likely to be highly inaccurate.

IOS, KOS, IRec and KRec are *internalist*. They refer to the internal competences of the players rather than their external behaviours. In most cases the distinction is not of any significance to us, but two illustrations should explain the distinction.

The first illustration concerns the appearance of players. The more complex of my models allows players to evolve more or less distinctive *appearances*. When players are looking particularly distinctive, a certain degree of IRec is able to distinguish them. When there is greater similarity of appearance, the same degree of IRec will confuse them more often. Here IRec is an internally defined degree of individual recognition capability which is distinct from the ultimate individual recognition behaviour which depends also on the range of appearances within the population.

A second illustration concerns Axelrod's diagnostics. *Niceness*, *provokability* and *forgiveness* are defined *externally* or behaviourally. This isn't an arbitrary decision. I am obliged to define

[1] As will 'be made clear in the later chapter I take the programs to be neither models nor simulations in any of the technical senses, but *examples*. However, the use of this word would be confusing here, and I hope the reader will find the terms "model" and "simulation" convey the meaning with less cognitive dissonance.

them this way if it to be possible to empirically test Axelrod's story where *niceness* can increase, not due to an internal switch to a different strategy, but due to a failure of KRec.

One final item of terminology. In the next section I introduce *experimental frameworks*. These are simply different programs that describe the area of investigation at different levels of detail.

### 5.3.2 Experimental frameworks

Three distinct experimental frameworks have been used. The first two of which I describe as the *simple models* are used to get quick overviews of the problem space.

The third, which will be described in detail in the next chapter, is a complex and detailed, virtual world, containing players who are typically balancing a mix of strategies, recognition competences and appearance characteristics. Successful players of each generation breed and produce offspring who inherit a mix of all these traits.

- Framework 1 : An analytic model based on the simple game theoretical mathematics. Here I model the proportions of a few basic strategies within the population. Their fitness is calculated as the sum of their scores against each other strategy type, multiplied by the proportion of that strategy within the overall population. It represents an infinite, random-mixing, polymorphic population.
- Framework 2 : This model features a finite population of individually represented players. Each player plays a single strategy. Players can be paired systematically or randomly. Scores for each player are recorded, but the strategy mix of the next generation is
- Framework 3 : This model represents a finite population of players individually. Players' strategies and other personal characteristics are represented in an explicit genotype. Players are matched systematically or randomly. The next generation is bred by crossing the genotypes from two successful parents of the preceding generation.

### 5.3.3 The implementation and degree of individual recognition

It is obvious that the mechanism of recognition will differ from species to species, some animals relying on visual cues, others on olfactory information and still others on sound. Although we can safely assume that some pattern recognition architecture underlies and enables the capability.

This is *not* meant to be research into the evolution of a particular type of mechanism. Still less, is it intended to ask what makes for a *good* pattern recognition mechanism. There is no hypothesis here about what mechanisms animals actually use. Information is simply provided to players as and when they need it.

Having stated this, in practice, it is not possible to be completely innocent of such issues. Although I am not making hypotheses about the mechanisms of recognition, a model still has to *implement* some kind of mechanism. And that necessitates making assumptions about how such a mechanism works and, more problematically, how it *fails*.

When players deserve the information that individual recognition would give them, they get it. But what information should a player without individual recognition should receive? How should this non-recognising player behave?

One intuition, is that players are likely to have the capability to a degree. The quality of the information received by a player, should be variable; its accuracy reflecting the degree to which player has "invested" in the mechanisms that support it. This intuition is also consistent with the usual evolutionary assumption that there must be a continuum between no trait (or a proto-version of a trait) and the trait.

But without an actual re-identification mechanism it is necessary to invent a notion of partial degree of recognition; an idea of what inaccurate recognition information should be like; and a plan of what will happen in the event of misrecognition.

These are *assumptions* of the model. Clearly for any particular animal species, these assumptions might be taken to be a hypothesis about the psychology or cognitive mechanisms it uses; and the results of the simulation could be compared with animal data. This comparative testing is interesting further research that could be done. However, it has not been attempted here. And, again, is not the focus of this work.[2]

### 5.3.4 The notion of kinship

The other big design issue is the implementation of kinship. This is handled differently in each framework.

In Framework 3 players are true descendents of members of the previous generation. Therefore kinship is a relation between players in virtue of shared ancestry.

In framework 1, where players are not explicitly represented, kinship is treated as a simple probability. A certain proportion of games are presumed to be with kin.

In framework 2, although players are individually represented, there is no record of them as children of particular parents. Hence it is not possible to treat kinship as it is in framework 3. Instead, players are arbitrarily located in a 1 dimensional array so that a distance between them, can be calculated. Players within a certain distance are taken to be kin. Players further apart are taken not to be.

Some consequences of this should be noted.

- All players have, roughly, the same number of relatives[3]
- The number of relatives a player has is not dependent on its success. Even if you are the last TFT strategist in the population, you still have as many siblings as the most popular strategy.
- There is no notion of degree of kinship. All players within the kinship distance are equally kin.

[2]Where comparisons of different mechanisms have been made, it has usually been for the purpose of testing the possibility that a particular observation may be an artifact of the mechanism. And the alternative mechanism has been introduced purely to see if it makes the observed effect go away. This tells us nothing about the correctness or value of either mechanism, but is a quick way of seeing whether the observation is an artifact of the inner details of a mechanism or is genuinely something caused by external constraints. Always my interest is in these external constraints.

[3]Note : there is no "wrapping" at the ends of the array. The player at position 0 and the player at position n are separated by a distance of $n - 1$ rather than 1. Hence players at the ends of the array have fewer relatives.

## 5.4 Frameworks 1 and 2

### 5.4.1 Comparing ALLD, TFT and KOA

Frameworks 1 and 2 can be used to look at the interactions between three pure strategies ALLD, TFT and KOA. This is essentially a pilot study to see how the three strategies interact. ALLD and TFT are the usual strategies. The Kin Oriented Altruism strategy is to co-operate with kin, and defect against non-kin.

In the games described here, there is no misrecognition either of individuals, or kin.

The games use the following payoff matrix. (Payment to player on the left.)

| | Co-operate | Defect |
|---|---|---|
| Co-operate | 6 (reward R) | 0 (suckers payoff S) |
| Defect | 8 (temptation T) | 3 (punishment P) |

Table 5.1: Scoring Matrix for the simple model.

which fulfils the two inequalities for the game to be a Prisoner's Dilemma : $S < P < R < T$ and $2R > T + S$.

The score for each pairing is therefore as shown in the table 5.2. From this you can work out that scores for an ALLD, a TFT and a KOA are

$$W_{ALLD} = i(3m+3) + j(3m+8) + k(5pm+3m+5p+3) \tag{5.1}$$

$$W_{TFT} = i(3m) + j(6m+6) + k(6p+3m+3pm) \tag{5.2}$$

$$W_{KOA} = i(3m-3mp+3-3p) + j(8-2p+3mp+3m) + k(3p+3+3mp+3m) \tag{5.3}$$

where $i$ is the frequency of ALLDs in the population, $j$ is the frequency of TFTs in the population and $k$ is the frequency of KOAs in the population, $p$ is the probability that an opponent is kin, and $m+1$ is the average number of matches between pairs of players.

### 5.4.2 Triangle Diagrams

As demonstrated by Maynard-Smith[44], we can plot the proportions of the three strategies as points within an equilateral triangle. Changes in the proportion of the strategies within the population can be graphed as lines through this space. Hence, the diagram shows the, experimentally discovered, channels and attractors of the ALLD-TFT-KOA space.

Compare the following two diagrams. The first, 5.1 shows the results of running the analytic framework 1 based on equations 5.1, 5.2, and 5.3.

We can see that there is no survival for KOA even when it starts in a dominant position (in the bottom right corner of the diagram.) All channels lead eventually to TFT or ALLD.

| Scores for each combination | | |
|---|---|---|
| Pair | 1st round | Subsequent rounds |
| ALLD vs. ALLD | 3 | 3 |
| ALLD vs. TFT | 8 | 3 |
| ALLD vs. KOA | $8p+3(1-p)$ | $8p+3(1-p)$ |
| TFT vs. ALLD | 0 | 3 |
| TFT vs. TFT | 6 | 6 |
| TFT vs. KOA | $6p+0(1-p)$ | $6p+3(1-p)$ |
| KOA vs.ALLD | $0p+3(1-p)$ | $0p+3(1-p)$ |
| KOA vs. TFT | $6p+8(1-p)$ | $6p+3(1-p)$ |
| KOA vs. KOA | $6p+3(1-p)$ | $6p+3(1-p)$ |

Table 5.2: Table showing first and subsequent scores of each pair of strategies. $p$ is the probability that the opponent is kin and $m$ is the no. of matches between each pair.

| Scores for each combination | | | |
|---|---|---|---|
| * | ALLD | TFT | KOA |
| ALLD | $3+3m$ | $8+3m$ | 5p + 3 + 5pm + 3m |
| TFT | $3m$ | $6m+6$ | $6p+3pm+3m$ |
| KOA | $3-3p+3m-3pm$ | $8-2p+3mp+3m$ | $3p+3+3mp+3m$ |

Table 5.3: Scoring Matrix for the ALLD, TFT, KOA game. p is probability of being kin and m + 1 = total number of matches.

In figure 5.2 we see a similar run using framework 2, the population consisted of 50 players. The family constant which represents how many family members the players had, was 10.

### 5.4.3 Interpreting the diagram

Clearly, from both diagrams this is a two strategy game. In populations which start with a critical proportion of TFT players, TFT can invade. Everywhere else the population moves to ALLD. KOA has no purchase at all. This is an issue we will consider in 5.5, below.

By comparing the two diagrams we also see the difference between a smooth, infinite population and a finite population. In the latter, sometimes a population can't move from a point because, although there is a slight imbalance in the scores of each strategy, these differences are too small to change the actual numbers of players of each strategy. In the continuous model, we see movement from these points, but here the population is trapped. In one sense this can be seen ot be an artifact. In another sense, such stability is a feature of real, finite populations. Where real animal populations are typically small, we will likely see this stability.

The end points of lines have open squares when co-operation outweighs defection, and filled

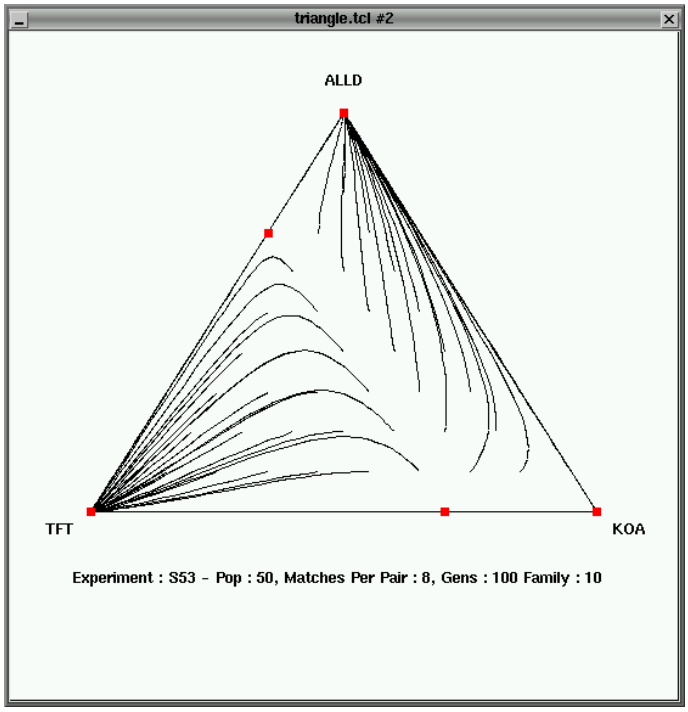


Figure 5.1: Trajectories through a mix of ALLD, TFT and KOA. Framework 1.

squares when defection dominates. It is noticable that defection dominates even where KOA is high. After all, most opponents aren't kin. Also, that defection still dominates some parts of the TFT "cachment area". A TFT strategy can be gaining, even when it is spending the majority of its games punishing defectors. As an aside, this might be interesting for observers of biological populations. A TFT like strategy might be in the process of taking over, even when the average behaviour is still non-altruistic.

### 5.4.4 The shadow of the future

In the iterated prisoners dilemma game, the term "the shadow of the future" is used to refer to the probability that a pair of players will meet and play again. When the number of repeat interactions is high, the virtues of mutual co-operation outweigh the benefit that ALLD gets by defecting against TFT on the first move. An extreme example is figure 5.3 where each pair only meets twice, the shadow of the future is too small to encourage significant reciprocal altruism. In this situation, TFT finds no purchase and all roads lead towards ALLD.

## 5.5 The failure of kin oriented altruism

We might be a little surprised by the total failure of KOA. Surely, we expect KOA to do better than ALLD. In these examples there is no possibility of failure of kin recognition and therefore no mistaken co-operating with non-kin.

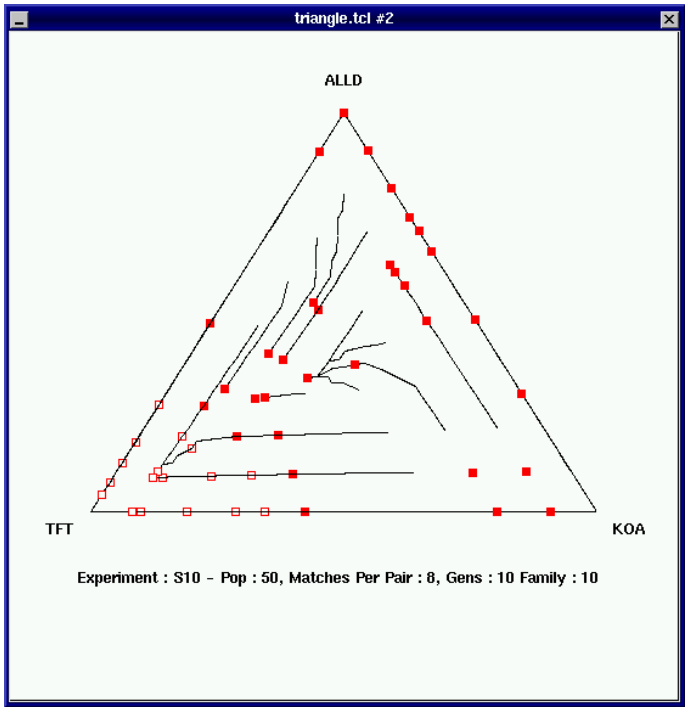


Figure 5.2: Trajectories through a mix of ALLD, TFT and KOA

Of course, the failure is an artifact of the models used. But I believe that this case is worth exploring because the problem is more complex than one might at first believe.

In framework 1, I have two frequency based co-efficients. One for relatedness and one for strategy mix in the population. Hence the score of a KOA player is

$$W_{KOA} = ipE(ALLD(f)) + i(1-p)E(ALLD(n)) + \quad (5.4)$$

$$jpE(TFT(f)) + j(1-p)E(TFT(n)) + \quad (5.5)$$

$$kpE(KOA(f)) + k(1-p)E(KOA(n)) \quad (5.6)$$

where $i$ = probability that opponent is ALLD, $j$ is probability that opponent is TFT and $k$ is probability that oponent is KOA, $p$ is probability that opponent is a family member and $E(STRAT(f))$ is the expected score a KOA player gets from an opponent who plays STRAT with a family member and $E(STRAT(n))$ is the expected score a KOA player gets from an opponent who plays STRAT with a non family member.

But, the two probabilities are independent. Yes, if KOA is a good strategy it will proliferate in the population. *But there seems to be no way that if I do well through KOA, which implies leaving more offspring, I actually increase the number of family members in my population.* The next generation of players, even if they are KOA strategists will be unrelated KOA strategists.

It is the independence of the probabilities which is the problem. Consider a simple situation where there is no TFT. The population starts with different mixes of ALLD and KOA. As shown

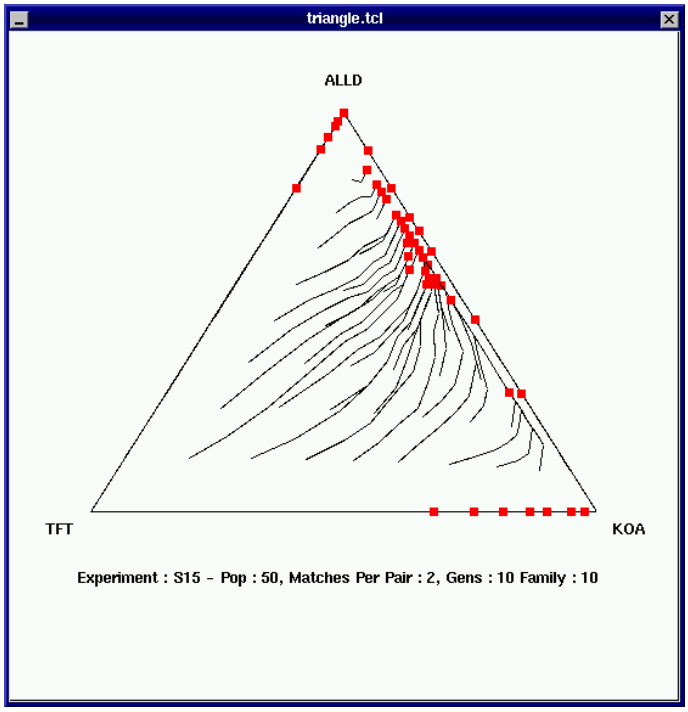


Figure 5.3: Trajectories through a mix of ALLD, TFT and KOA when players only meet twice. Using Framework 2.

by figure 5.4, even when KOA starts with 90% of the population, it is wiped out by ALLD. Why is this?

Looking again at the scores in 5.1. Per match, when KOA plays an ALLD which is a family member it scores 0 (co-operate with a defect). When playing with a non-family member, it breaks even. (3 all.) Against another KOA it scores 6 when a family member and 3 when non family member.

OK so when p is probabililty of being related and q is the frequency of KOA in the population.

$$W_{KOA} = 7pq - 3q - p + 3 \tag{5.7}$$

while

$$W_{ALLD} = 5pq + 3 \tag{5.8}$$

KOA can therefore invade when

$$2pq - 3q - p > 0 \tag{5.9}$$

But remember that p and q are probabilities so $0 < p < 1$ and $0 < q < 1$. So this will never be greater than 0. KOA can not invade ALLD.

Maynard Smith suggests that there are two ways of modelling games between kin :

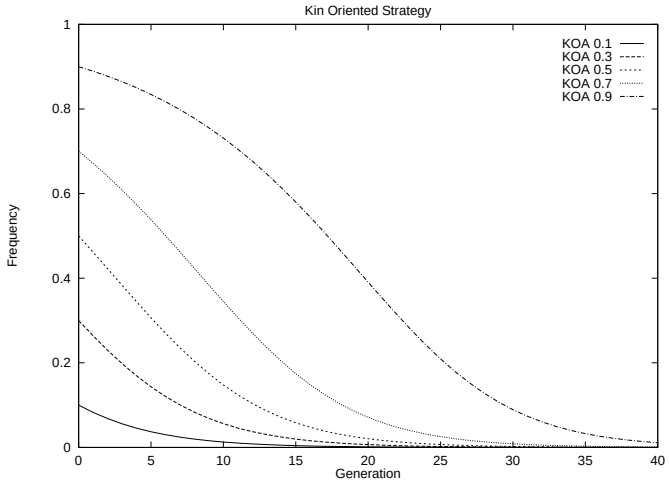


Figure 5.4: KOA driven extinct by ALLD when there is no TFT.

- *inclusive fitness*
- *neighbour-modulated* or *personal* fitness

and goes on to say : "[the] personal fitness approach is formally correct, but does not provide a simple way of finding ESSs, whereas the inclusive fitness method does provide a means of finding ESSs but can lead to wrong conclusions."

### 5.5.1 The inclusive fitness approach

How could we capture this inclusive fitness in a model? An explicit way could be to actually make the scores reflect inclusive fitness. As discussed so far, there is no notion of degree of kinship. But now suppose that all kin are in fact full siblings, and therefore share 50% of their genetic material with each other. To make the payoff reflect inclusive fitness, any player playing a sibling, should get an extra 50% of the score that its sibling gets.[4]

This leads to the following equations.

$$\begin{aligned} W_{ALLD} &= i(3m) + j(3m+5) + k(5pm+3m) + \\ &\quad \frac{ip(3m)}{2} + \frac{jp(3m-3)}{2} + \frac{kp(3m-3mp)}{2} \end{aligned} \tag{5.10}$$

‘

[4]This way of modelling Inclusive Fitness is open to question. Without an explicit payoff given to relatives, the basic model is too abstract to capture any notion of IF. But the assumption that all relatives are full siblings is clearly over-simplistic. A future refinement could, perhaps, be made to this model, which derived a more plausible degree of average relatedness of all players. Such a model would nevertheles also require explicit assumptions about the breeding strategies and viscousness of the population.

$$W_{TFT} = i(3m-3)+j(6m)+k(12p+6m-6+6mp)+ \frac{ip(3m+5)}{2}+\frac{jp(6m)}{2}+\frac{kp(3m+3mp+p+5)}{2} \tag{5.11}$$

$$W_{KOA} = i(3m-3mp)+j(3m+3mp+p+5)+k(3m+3mp)+ \frac{ip(5pm+3)}{2}+\frac{jp(12p+6m-6+6mp)}{2}+\frac{kp(3m+3mp)}{2} \tag{5.12}$$

It does work. See figure 5.5 for a comparison of the fate of KOA when one's relative's fitness explicitly counts towards one's own.

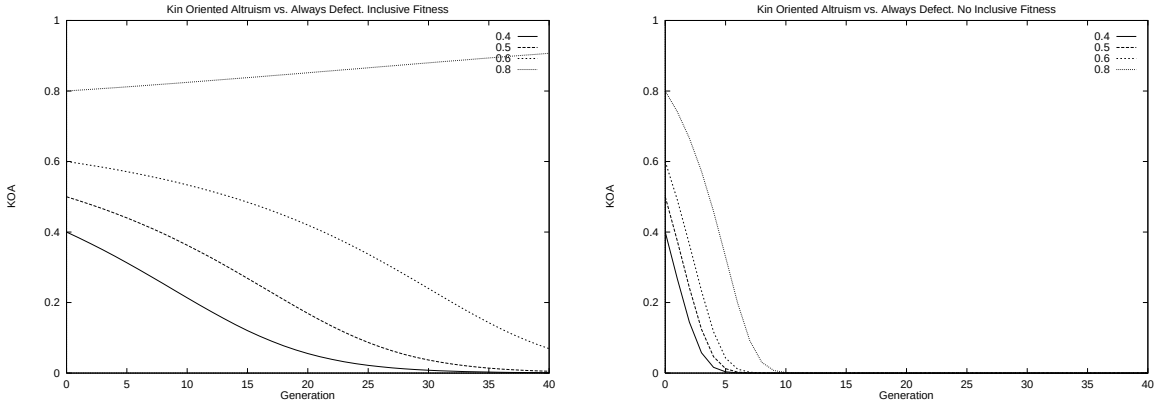


Figure 5.5: Where the scores model *inclusive fitness* (left) KOA can invade, given a high enough frequency in the initial population. Without *inclusive fitness* (right) KOA soon goes extinct.

### 5.5.2 The personal approach

If we model using the personal approach we assume that all relatives play the same strategy. So the fact that an opponent is kin, implies that he plays the same strategy.

It seems intuitively correct. And in figure 5.6 we see that by implementing this rule we now get a movement towards KOA and TFT as we might have hoped to see previously.

Unfortunately, we might have a suspicion that there is something very wrong with the personal approach. Consider a population where the probability of being related is $r$, and ALLD is being driven extinct. There will come a time when an ALLD player has a probability $r$ of being related to an opponent, but where the frequency of ALLDs in the population has dropped below $r$. Assuming that all this player's relatives are ALLD is clearly flawed.

## 5.6 Conclusion

In this chapter I put forward three topics to focus research on.

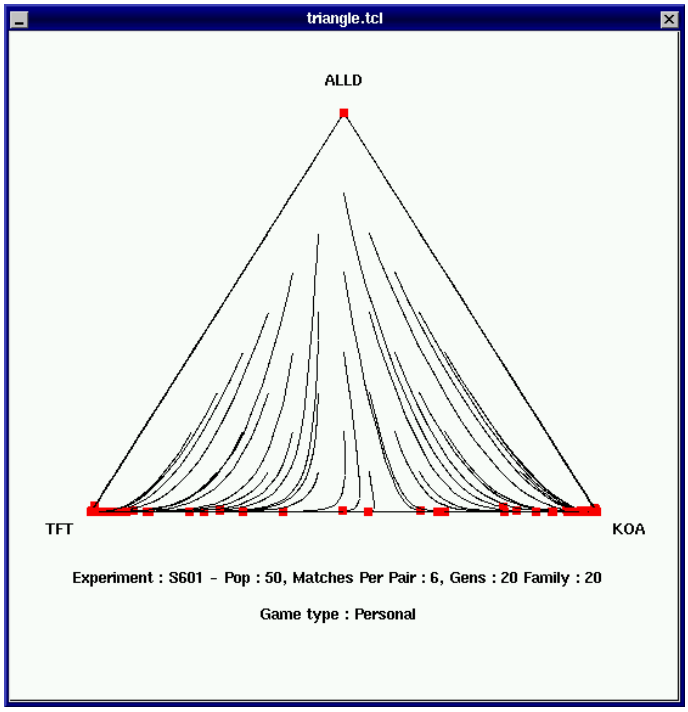


Figure 5.6: Trajectories through a mix of ALLD, TFT and KOA using the personal rule for calculating fitness.

- The trajectory of the evolution of IRec.
- The relationship between IOS and KOS.
- The evolution of IOS in a spatial world.

I also introduced three frameworks, and showed some results using the first two of these. This led us to note that there were problems with the way of thinking about *kinship* in this framework; and I tried using two suggestions from Maynard Smith, as to how to represent kinship.

Neither of these is entirely satisfactory. My feeling is now that we must move to another of these frameworks for a better model of kinship.

In this chapter, as well as exploring kinship in some detail, I also raised the issue that decisions have to made about the nature of misrecognition.

# Chapter 6

# The Complex Model : Framework 3

---

## 6.1 Introduction

The complex program which I am describing as framework 3, is a detailed computer simulation within which virtual agents play the prisoner's dilemma against conspecifics in their population.

This chapter describes the program.

## 6.2 Structure of the Players

Players are prisoners dilemma players and ultimately chose one of two moves : cooperate or defect. The decision as to which move to make is based on their bias towards using either a kin oriented strategy or an individual oriented strategy; and their perception of the opponent. The facts they perceive are these:

- Whether the opponent facing them has previously been met or not.
- If so:
  - What move that opponent played against them last time.
  - What move they played against the opponent last time.
- How similar the opponent appears to themselves.

A record storing this information is known as a piece of *knowledge*. It is a mere design detail of the program that the information about the previous interactions between the players is managed by another part of the program, rather than the player class itself. Accurate, objective knowledge of the history of interactions is recorded, but players only receive *subjective* knowledge; that is knowledge that may be distorted by their fallibility of recognition capabilities.

The players themselves are structured into

- Strategy Tables

- Capability Parameters
- Appearance
- Experience

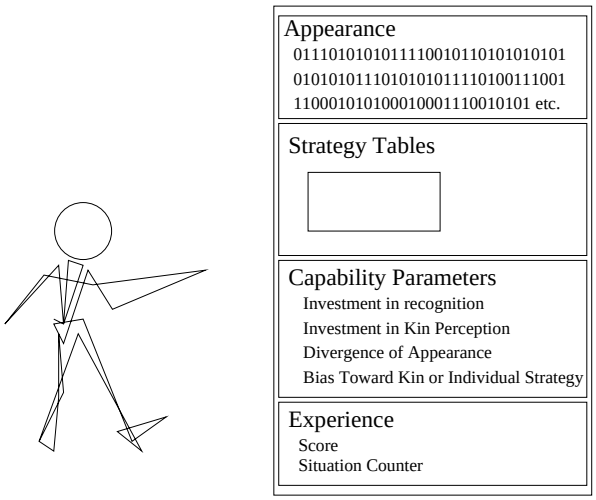


Figure 6.1: The Internal Structure of a Player

## 6.3 Genetic Algorithm

The population structure is defined by four parameters :

- $p$, the number of players in the population
- $g$, the number of breeders
- $s$, the number of high scoring, elite individuals
- $f$, the number of strangers joining the population per generation

.

All members of the group of high scoring players have the same probability of being a parent. No players who fall outside this group will be parents.

However, the model allows the possibility that generation $n+1$ is not entirely composed of the children of generation $n$. There can be *elitism*, meaning that the genotypes of a small group of the fittest individuals of the previous generation can be copied across into the following one. Also, a number of random *strangers* can be added to the population each generation. This number is meant to represent the porousness of the society, the degree to which it can be joined by strangers from other populations.

In all populations studied $s + f < g < p$

### 6.3.1 The Breeding algorithm

```
Let b be an array of the g highest scoring members of the population
Let w be an empty array for the next generation, with length p
Into the first s elements of w, copy s members of b chosen at random
Into the next f elements of w, put f new players with random genotypes
Into all the remaining elements of w, put a child of x and y,
        where x and y are players drawn randomly from b
```

## 6.4 Strategy Tables

The players have two strategy tables.

One, for *individual oriented strategy* (see 6.1), which specifies actions as a response to previous situations; including one action for when there has been no previous interaction with this opponent.

| I Co-oped, He Co-oped | I Defected, He Co-oped | I Co-oped, He Defected | I Defected, He Defected | Who's This? |
|---|---|---|---|---|
| C | D | D | C | D |

Table 6.1: Individual Oriented Strategy of a player. An example table coding for the player's action depending on previous encounter.

The other (table 6.2) consists of probabilities of co-operating with an opponent given the *similarity of appearance* between that opponent and oneself. The appearance of a player is described fully below in section 6.5.4. Within this model the similarity of appearance is usually suggestive of the degree of relatedness between the two players. Hence this information might be used as the basis of a strategy of kin oriented altruism and this table is known as the *kin oriented strategy* table.

| Similarity | chance of Co-operating |
|---|---|
| less than 0.5 | 0.503 |
| 0.5-0.55 | 0.510 |
| 0.55-0.6 | 0.53 |
| 0.6-0.65 | 0.55 |
| 0.65-0.7 | 0.58 |
| 0.7-0.75 | 0.52 |
| 0.75-0.8 | 0.61 |
| 0.8-0.85 | 0.67 |
| 0.85-0.9 | 0.71 |
| 0.9-0.95 | 0.69 |
| 0.95-1.0 | 0.87 |

Table 6.2: Kin Oriented Strategy of a player. An example table coding for the player's action depending on similarity of appearance. (Note this is an actual evolved player.)

The Kin Oriented Strategy Table see 6.2 lists probabilities of co-operating with opponents depending on their similarity. Note that a similarity of 50% is the average genetic similarity of any two arbitrary individuals, so it was considered that distinguishing between values in the range 0%

to 50% would be of little value. Hence one table entry for 0 to 0.5. The rest of the table contains strategies ranging from 50% to 100% similarity in steps of 5%.

## 6.5 Capability Parameters

In addition to the above tables, players are defined by several further, genetically determined, parameters.

These are

- a bias towards using individual or kin oriented strategy,
- an investment in recognition,
- an investment in kin (similarity) perception, and
- a divergence of phenotypical appearance from genotypical appearance

### 6.5.1 Recognition Bias

This is a real number between -1 and 1 which influences which strategy a player will choose. The gross algorithm for the player's decision making goes like this.

```
Let b be a bias towards playing a kin oriented strategy between -1 and 1
Let r be a random number between -1 and 1
if r >= b then
     this player shall play an individual oriented strategy,
otherwise
     this player shall play a kin oriented strategy
```

### 6.5.2 Investment in Recognition

The trait known as *investment* represents the investment made in a mechanism for individual re-identification, it implies a level of IRec accuracy which is fixed for life. This level of accuracy determines how good the player is at re-identifying an opponent. It is encoded at a locus on the genotype which is also labelled *investment*. Investment thus represents the proportion of resources that a player has allocated to the capability of recognition. It is in the range 0 to 1.

See 6.6 for the full algorithm for individual recognition.

### 6.5.3 Kin (Similarity) Investment

The *kin (or similarity) investment* parameter controls the accuracy of kin recognition. When it is high, the degree of similarity perceived by a player is accurate. When it is low, the player receives an erroneous degree of similarity. Kin (similarity) investment is also in the 0 to 1 range.

The precise algorithm is

```
Let p1 and p2 be two players who meet
Let k1 be the kin (similarity) investment made by p1 (between 0 and 1)
Let c be a random floating point number between -1 and 1
Let e be c multiplied by 1 - k1
Let s be the similarity between p1 and p2

Let s1, the similarity as perceived by p1, be s + e.
                                    (Limited between 0 and 1)
```

It should be noted that when we take the number of kin perception errors as a statistic, we count all instances of $e > 0.06$, which is enough to shift one row of the similarity strategy look-up table.

### 6.5.4 Appearance

Appearance is a bit string. It represents those features of the player that another would use to recognise it. It is possible to measure the similarity of two appearances by taking the Hamming distance (the number of bits where the two strings differ.) A Hamming distance of 0 means that the two players appear identical. A Hamming distance of the length of the appearance string, typically 512, means that the two players are as different as it is possible to be.

As noted above, one piece of information available to a player is the similarity of an opponent's appearance to its own. In certain circumstances this can be indicative of the degree of relatedness of that opponent to itself, because appearance is encoded in the genotype. The consequence is that the similarity of genotypically derived appearance between two players depends partly on their relatedness, and partly on the mutation rate. However, another complication of the model is that it allows a difference between genotypic appearance and phenotypic appearance, as discussed in the next section.

### 6.5.5 Diversity of Appearance

In the model, appearance is made complicated by the fact that, phenotypical appearance - which is made available to the player, and used for other similarity calculations - is not always the same as genotypical appearance. Both appearance strings are of equal length, but the phenotypical appearance can have an amount of noise is inserted into it. The amount of noise is determined by the *diversity (or divergence) of appearance* parameter. This is also an integer between 0 and 255, and represents the number of bits which can be flipped. (NB : For certain comparisons, divergence is also scaled to between 0 and 1 and is usually graphed as such.)

The exact algorithm for adding noise is as follows.

```
Let d be the divergence of appearance parameter
Let a be the genotypic array of bits
Let b be the phenotypic array of bits copied from a
```

```
Iterate d times
        Let x be a random bit in b
        There is a probability of 0.5 that x be reset to NOT x
```

Note the effects. Players who have a high divergence parameter display less family resemblance to their kin. In the actual experiments, appearance is 512 bits long. The diversity of appearance can be between 0 and 255. Hence, at maximum diversity, a genotypic appearance can have 50% of its bits randomised. This should be sufficient to make full siblings unrecognisable as such. But as the table 6.3 below shows, in general, a diversity of 1 won't entirely eliminate the possibility.

The table 6.3 of typical similarities between kin was found experimentally using the code from the model.

| Divergence | 0 | 0.2 | 0.5 | 0.7 | 1 | U |
|---|---|---|---|---|---|---|
| Stranger | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| Nephew | 0.58 | 0.57 | 0.55 | 0.53 | 0.53 | 0.55 |
| Parent | 0.68 | 0.66 | 0.61 | 0.58 | 0.56 | 0.62 |
| Grandparent | 0.59 | 0.57 | 0.55 | 0.54 | 0.53 | 0.55 |
| Cousin | 0.53 | 0.52 | 0.51 | 0.51 | 0.51 | 0.51 |
| Half Sibling | 0.59 | 0.57 | 0.56 | 0.55 | 0.53 | 0.55 |
| Full Sibling | 0.73 | 0.69 | 0.64 | 0.61 | 0.58 | 0.66 |

Table 6.3: Similarities of Appearance of Relatives when mutation rate is 1/15. A similarity of 1 is identical. A similarity of 0.5 is chance. The columns with numerical headings are for a fixed divergence of appearance parameter. The column headed "U" is for an undefined divergence of appearance. That is, one for which the divergence of appearance is encoded on the genotype. The table is generated empirically, by averaging 500 examples of each relation, so the U column is informed by a representative spread of divergence.

Note several things from the table.

- Even unrelated strangers are, on average 50% similar.
- Divergence of appearance, even when 1 is not quite enough to destroy kin perception. (In any individual case a high divergence potentially could render even full siblings imperceptible.)
- Siblings are closer than parents, cousins closer than grand-parents.

## 6.6 Individual Recognition and its Failure

The full recognition / misrecognition algorithm is as follows.

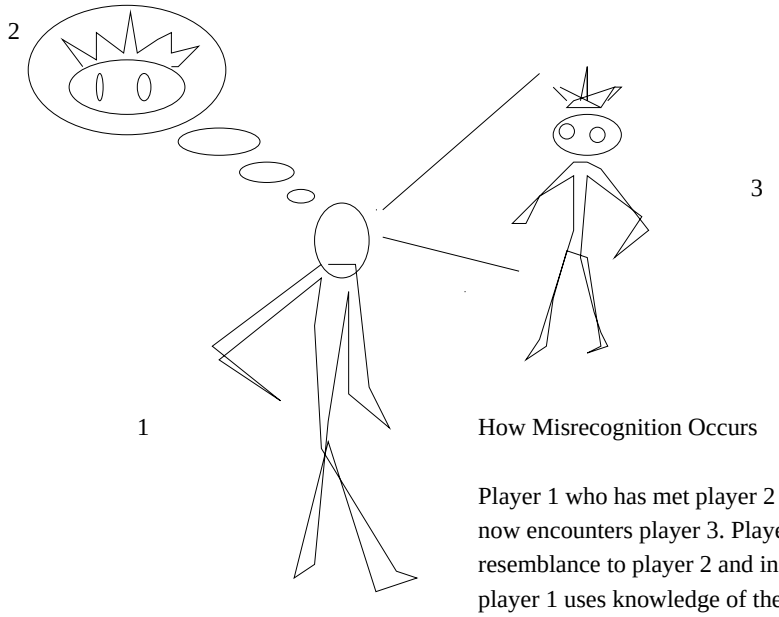


Figure 6.2: Misrecognition

```
Let p1 and p2 be players in the population
Let i1 be the investment made by p1 in individual recognition (0 < i1 <1)

If p1 has previously met p2

  There is a probability of i1 that p1 shall correctly recognise p2

  If p1 correctly recognises p2

      p1 shall receive accurate information about the pair's previous match

  otherwise

      Let e be the margin of error (1 - i1)
      Let p3 be a player, chosen at random, who
               a) has a similarity of appearance to p2 of more than e
               b) has previously met p1
      p1 shall receive information about p1's previous encounter with p3

In the case where p1 has not previously met p2
```

```
p1 shall play the move appropriate to stranger
```

## 6.7 Axelrod Diagnostics

Earlier I described a story offered by Axelrod as to the evolution of co-operation which was expressed using his three diagnostics : *niceness*, *provokability* and *forgiveness*. One of the requirements for this model is that we might try to see this process occuring.

The other reason to be interested in these diagnostics is that they can be thought of as behavioural components of strategies. When a working with a few strategies, it is possible to plot graphs showing the proportions of all the strategies in the population. Once the number of strategies goes above 5 or 6, these graphs can get confusing. Niceness, forgiveness and provokability are components or diagnostics of every strategy. One can plot these to get a sense of the broad character of a population.

### 6.7.1 Definitions

What are Axelrod's diagnostic categories of *niceness*, *provokability* and *forgiveness*? There is an issue as to whether these are to be considered as dispositional properties of individual players or whether an attempt to measure them should be based on actual behaviour over the players' lifetime.

Both can give distorted impressions. If certain situations are not arising, for example no co-operation is occuring, then that part of the genotype that would code for a forgiving behaviour is genetically drifting and measurement of forgiveness on the genotype would be misleading.

On the other-hand, provokability is best described as a willingness to punish defections. If a TFT strategy meets PAVLOV, PAVLOV will start by defecting then, on being punished by TFT, will revert to co-operating. In a long sequence, the two strategies will co-operate. The one instance of punishment unleashed by the TFT player will therefore be only a small proportion of the number of the overall number of moves. Yet this player *is* provokable. And provokability has been vital to its success.

A further problem, what of players using a mixture of individual and kin oriented strategies? Surely we would want to say that a player that played TFT a third of the time was less nice than a player that played TFT two thirds of the time. Yet Axelrod's story turns on the idea that niceness increases due to a *failure* of kin recognition. If the measurements are to capture this, then an increase in co-operation due to inadequate kin recognition, *when kin oriented strategy is being played*, must themselves count as nice.

For this reason, it seems we must adopt a behavioral measurement rather than a strategy measurement for niceness. Similarly for provokability - a strategy which is overly generous to even the most distant relative might fail to be provokable enough - and forgiveness - a strategy that is mean to relatives might unduly punish transgression.

The most serious distortions due to a lack of appropriate circumstances can be handled by correcting the measures for this. Hence forgiveness can be measured as the number of times a player has an opportunity to be forgiving, and is.

### 6.7.2 A measurement of niceness

Niceness, then, we will calculate as follows :

```
Let p be a player
Let o be the number of opportunities for niceness
    where an opportunity for niceness is a match m with an opponent q
         who did not defect in the preceding match between p and q
Let c be the number of opportunities for niceness where p co-operated
Let the niceness of p be c divided by o
```

Remember, niceness is the refusal to defect first. So it must include all the times that the player co-operates, but should not measure indifference to being defected against. So co-operations with players who previously defected do not count. Niceness is normalized (divided by the number of times that the player could have been nice) so that cross comparisons can be made between runs with different opportunities.

### 6.7.3 A measurement of provokability

The algorithm for calculating provocability of a player is :

```
Let p be a player
Let o be the number of opportunities for provocation
   where an opportunity for provocation is a match with opponent q
        who defected in the previous match between p and q
Let d be the number of opportunities for provocation where p defected
Let the provocability of p be d divided by o
```

Normalization occurs as with niceness, so provocability measures the proportion of retaliates compared to the number of opportunities for retaliation; but it is important to remember that when the overall number of opportunities is small, the genotype controlling this behaviour will be under less selective pressure.

### 6.7.4 A measurement of forgiveness

And the algorithm for calculating forgiveness :

```
Let p be a player
Let o be the number of opportunities for forgiveness
     where an opportunity for forgiveness is a match with opponent q
```

```
        who co-operated in the previous match between p and q
        but who had defected in the match preceding that
Let f be the number of opportunities for forgiveness where p co-operated
Let the forgiveness of p be f divided by o
```

Clearly forgiveness must take account of the last two interactions of the pair; otherwise any strategy that co-operated on the previous move would count as an opportunity for forgiving. This seems both counter-intuitive and would make forgiveness pretty much the same as co-operate with co-operations. What aren't assessed are the player's own previous moves. Forgiveness doesn't require that the opponent returned to co-operating because of punishment by the player. The original defection might have been the result of a rare error of kinship perception, and the following, co-operation, the return to perceptual clarity.

There are still reasons to be unhappy with this formulation. Indiscriminate co-operators will be counted as forgiving whereas intuitively, one might not have wished to include them. However, the actual control system of the players is not itself sophisticated enough to grasp the distinction. For example, TFT *is* counted as forgiving, but would not be able to distinguish the motivation of its opponent. Generously forgiving strategies such as those in Grim[18] are actually displaying a probability of indiscriminate co-operation.

## 6.8 Conclusion

This chapter has described, in detail, the model used for the empirical research in the rest of this thesis. Where variations on this model are used, such as in the spatial world in chapter 9 the differences will be described at the appropriate time. Criticisms of the model are in the final discussion chapter.

# Chapter 7

# The Evolution of Individual Recognition

## 7.1 Introduction

The first of the four topics to be investigated is the evolutionary dynamics of the growth of individual recognition. This chapter provides answers to to the following questions :

- Can TFT evolve when there's full Individual Recognition?
- How does the degree of individual recognition affect the evolution of TFT?
- What is the evolutionary dynamic of individual recognition invading a population?

The first of these questions is trivial in the sense that we already believe we know the answers to it. Can TFT evolve when there's full individual recognition? Yes. However, demonstrating that it does, corroborates the experimental framework.

The second question we don't know the answer to. No earlier work in this field, not even Crowley et al[10], has shown how co-operation or different strategies change with degree of individual recognition. We have some clues, the fact that a noisy world encourages less retaliatory or more forgiving strategies as witnessed by Lindgren and Seth may indicate something.

The third question is a more open-ended search for an interesting dynamic for the growth of IRec. What stimulates it? What suppresses it?

## 7.2 The Evolution of TFT

*Can TFT evolve when there's full Individual Recognition?*

The complex model was run for fifty generations, with a population of 80 players; of whom the best 30 were allowed to breed the following generation. As a default, two random strangers were added to the population every generation; and the two best genotypes were retained into the next generation. The population played the standard IPD with pairs being chosen and matched at random, any combination being acceptable including a player being matched with itself. Altogether

100,000 matches were shared out between the 80 players, per generation. As pairs were picked at random, not all players would have played the same number of matches but on average most players would have had around 2500 matches. Note that these parameters were found through preliminary investigations to reliably allow the evolution of TFT strategy. This experiment is number 210.

Figure 7.1 shows the count of the moves per generation. In the first generation, there are roughly the same number of co-operates as defects. However, as consistent defectors score more highly than co-operators, defection quickly increases while co-operation plummets. Around generations 3 to 6 co-operation is at a minimum, but then a change begins. We can guess that some players have discovered the benefit of mutual co-operation. From then, co-operation climbs quickly until at around generation 11 about three quarters of all moves are co-operates. Co-operation still increases but more slowly.

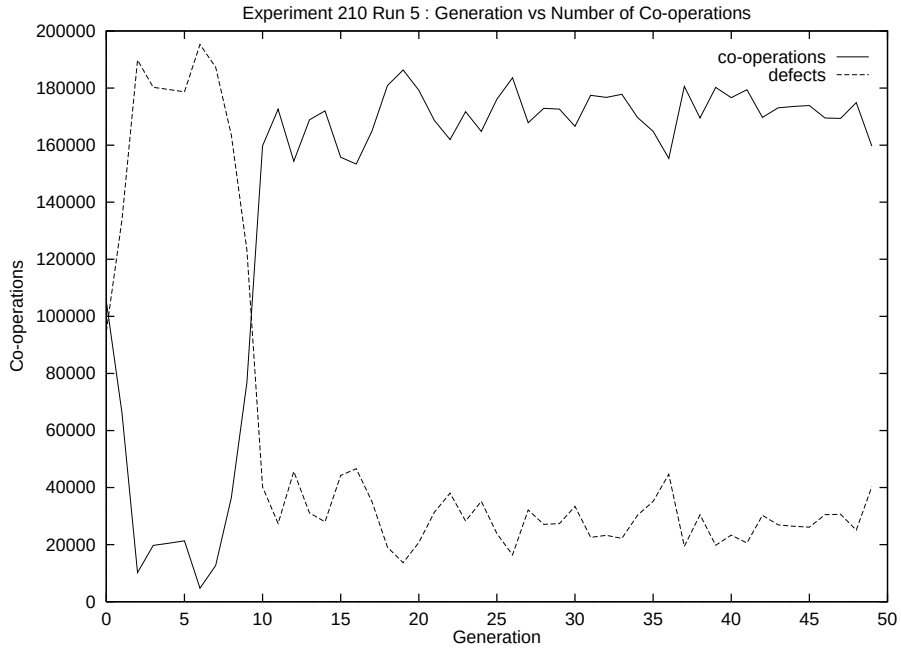


Figure 7.1: Experiment 210 : Run 5 : No. of Co-operates and Defects. Total of each move made within an evolving population of IPD players.

As you would expect with the prisoner's dilemma, mutual co-operation scores more highly than mutual defection. And the next graph 7.2 shows how the average score; and the score of the elite player behave with co-operation.

### 7.2.1 Niceness, provokability, and forgiveness in the evolution of co-operation

We presume that this co-operation is supported by reciprocal altruism, and that players are using a TFT strategy. In figure 7.3 we see the average niceness, provokability and forgiveness during the same run (210.5). The graph of the number of co-operations and defections from that run are

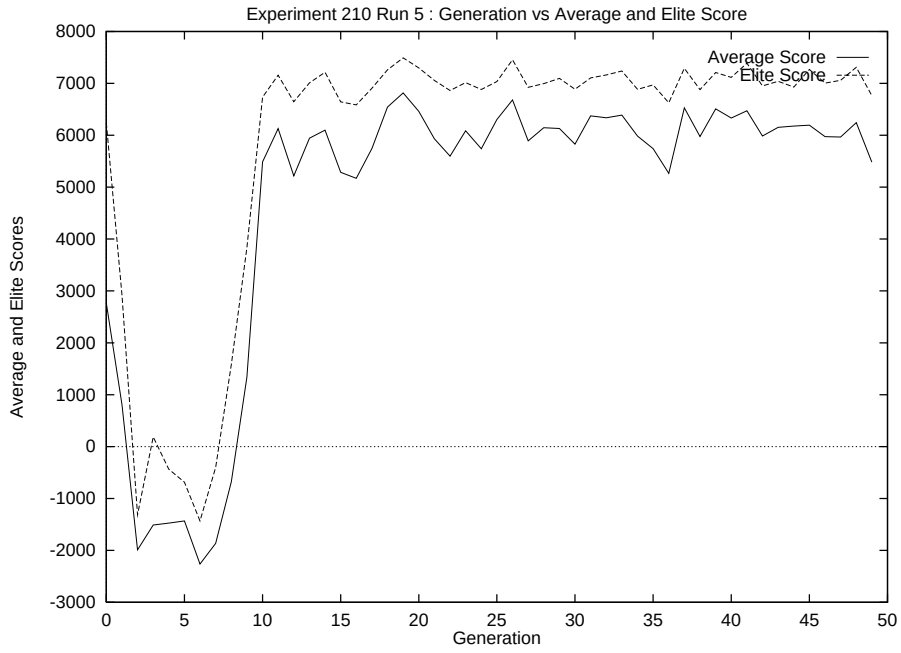


Figure 7.2: Experiment 210 : Run 5 : Average and Elite Scores.

included for comparison.

These graphs corroborate our presumption that the overall strategies are TFT-like in that all three of the diagnostics are high.

The dynamics of the the diagnostics are also worth noting. Provokability, being any tendency to return defect with defect climbs up with the initial invasion by ALLD. Co-operation is already starting to recover by generation 7 and both it and niceness increase together. Forgiveness follows them up, lagging perhaps 1 generation behind.

## 7.3 The Effect of Individual Recognition on Co-operation

*How does the evolution of a TFT strategy change with IRec?*

Throughout the experimental work I use a number of standard sequences of runs of the experiment. Sequence 1 is a sequence of eleven experiments (200 to 210) where players have a fixed degree of individual recognition, ranging from 0 to 1 in steps of 0.1.

Figure 7.4 shows the amount of co-operation over this sequence. The value of each data-point is found by averaging the values of the last generation (the 50th) from 10 runs with the same parameters.

Unsurprisingly, co-operation is high when the players have a high degree of individual recognition, and low otherwise. What might be less expected is that the jump occurs instantaneously between a recognition ability of 0.6 and 0.7 rather than as a more gradual function of recognition ability.

Nevertheless, this is quite explicable. In the low co-operation cases, the population, without

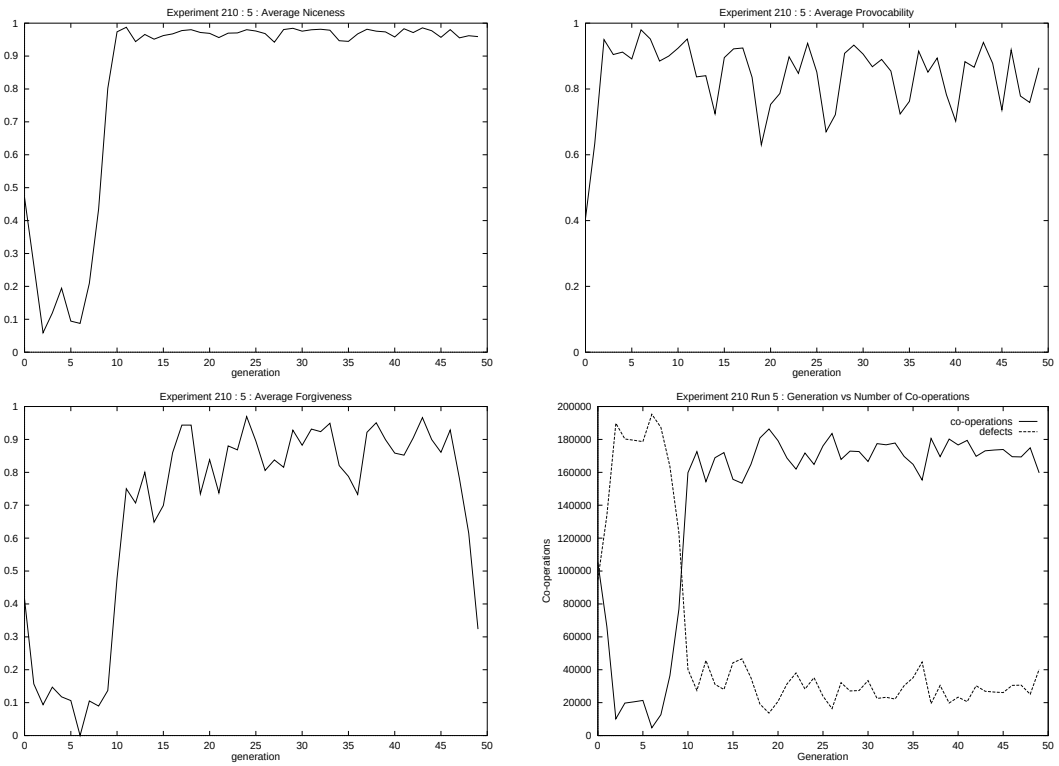


Figure 7.3: Experiment 210 Run 5 : Niceness, provokability and forgiveness

reliable recognition, is evolving a general strategy of indiscriminate defecting (ALLD). In the high case, it has switched to a tit-for-tat like strategy.

### 7.3.1 The effect of IRec on Axelrod's diagnostics

We can learn more about the underlying strategies by looking at the Axelrod diagnostics as IRec is changed in figure 7.5. Provokability remains high across all values of IRec. Under our definition, a pack of amnesiac ALLDs have as high provokability as TFT. Niceness jumps at the threshold. More, interesting is forgiveness which rises between IRec of 0.6 and IRec of 0.9.

Were we expecting the model to behave this way? We might have thought that forgiveness, depending on there having been a prior defection, would actually be higher when IRec was an imperfect 0.7. There would be more mistaken defection and therefore a greater need for tolerance of the odd slip. But here it looks as if this is not the case. (Though that may explain why forgiveness seems to decrease between 0.9 and 1. There is less need for it at perfect IRec.)

The conclusion. When IRec is unreliable, strategy is closer to GRIM than TFT.

Figure 7.6 shows a slightly curious view of experiments 206, 207 and 210 were each run 10 times. And then each generation was averaged across each run. That is, a single point was generated for experiment 206, generation 1 by averaging the values of generation 1 across all 10 runs. And the same for each succeeding generation. This is usually a confusing summary because when we look at progress over generations, we are looking at the fine detail of historical

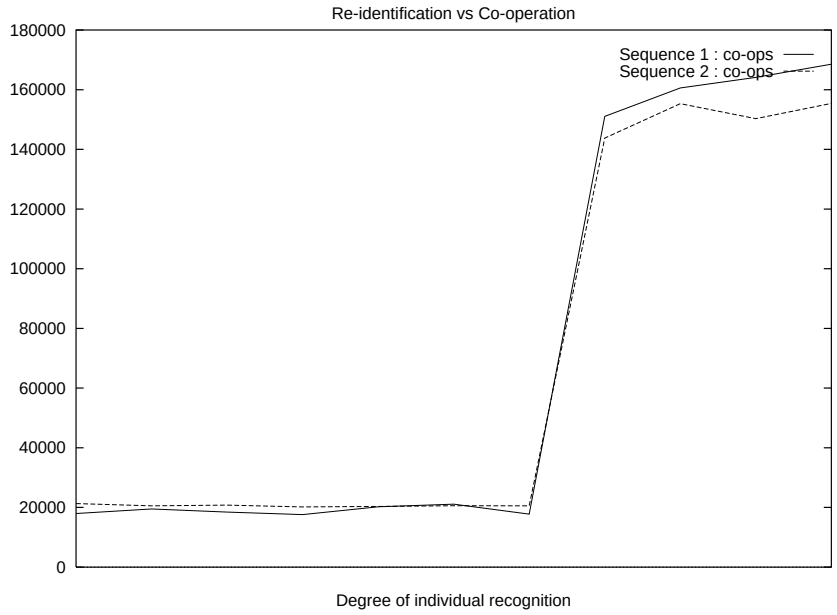


Figure 7.4: Co-operation as a function of individual recognition scaled between 0 and 1. The transition occurs between IRec = 0.6 and IRec = 0.7 . Last generation and average of generations. Sequence 1 shows the average of the total co-operation at generation 50, over 10 runs. Sequence 2 shows the average over all 50 generations, then averaged over 10 runs.

events. Even when events of the same type occur in other runs, the chances are that they occur at slightly different historical moments, so such averaging loses the fine detail one would want from a generational view. But in these simple experiments, the main events - the invasion of the population by defection, and the subsequent invasion by reciprocal altruism - occur reliably at the beginning of all runs.

Where individual recognition is 0.6, just too low to allow a reciprocally altruistic population to take off, niceness and forgiveness fall to residual rates but, as predicted, provokability is high. Where individual recognition is 0.7, just high enough to allow reciprocal altruism, niceness and forgiveness rise, and so does provokability. There is no sense in which these populations have become complacent. Finally, where individual recognition is perfect, everything rises.

## 7.4 The evolutionary dynamics of individual recognition

*What are the dynamics of an evolution by individual recognition?*

The result of the previous experiment shows that co-operation within a society is not a continuous function of individual recognition but that there is a threshold degree of individual recognition above which, co-operation can flourish. Partial IRec does not support a mix of co-operation and defection.

When individual recognition is, itself, allowed to evolve, we might then wonder how it arrives. From the previous result it seems that there isn't a reason for an intermediate degree of individual

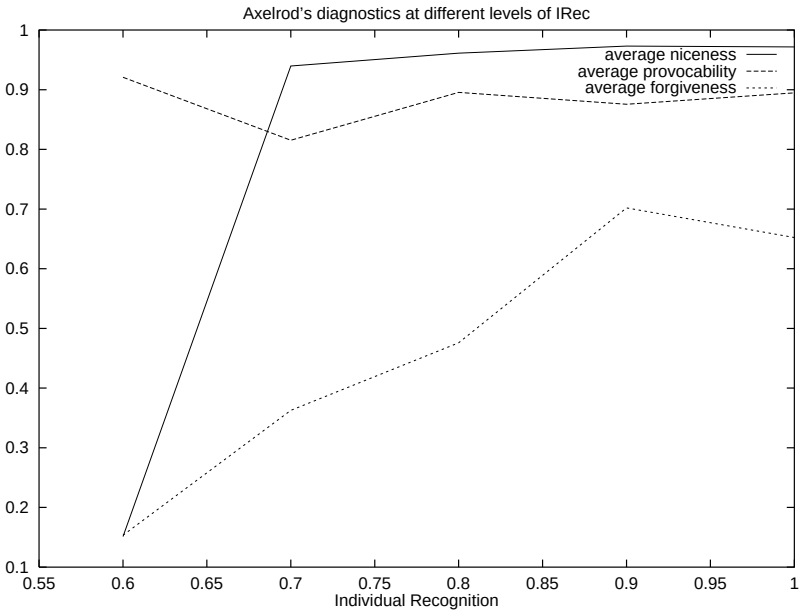


Figure 7.5: Niceness, provokability and forgiveness against individual recognition

recognition to be selected for. If this is the case we will be forced to hypothesise that the cognitive faculty of individual recognition did not evolve because of the need to play reciprocally altruistic games.

Either

- the faculty is adaptively inexplicable, ie. it appeared as a single mutation or due to some non-adaptational story;

  or

- the faculty evolved for some other purpose and was later co-opted for individual recognition.

In figure 7.7 (Experiment 100 Run 0) we see an example of the evolution of an individual recognising, co-operative society. Individual recognition manages to invade with ease, and a co-operative society takes over. But it confirms the unfortunate conclusion. Individual recognition jumps instantaneously, *prior to the take off of the altruistic society*. It is not caused or influenced by reciprocal altruism or any of the sub-behaviours.

### 7.4.1 The oscillation between GRIM and TFT

The fluctuations of forgiveness catch the eye, why the sudden drop between generation 35 and 40? As the drop in forgiveness occurs, the co-operates and score dip slightly, but provokability increases. Individual recognition holds steady, and the situation soon recovers. Compare runs 1 and 2 of this same experiment in figure 7.8 Even though individual recognition remains high, there are periods where forgiveness is low. Table 7.1 throws some light on the subject. It shows the frequency of each strategy within the final generation of three runs of experiment 100. As expected

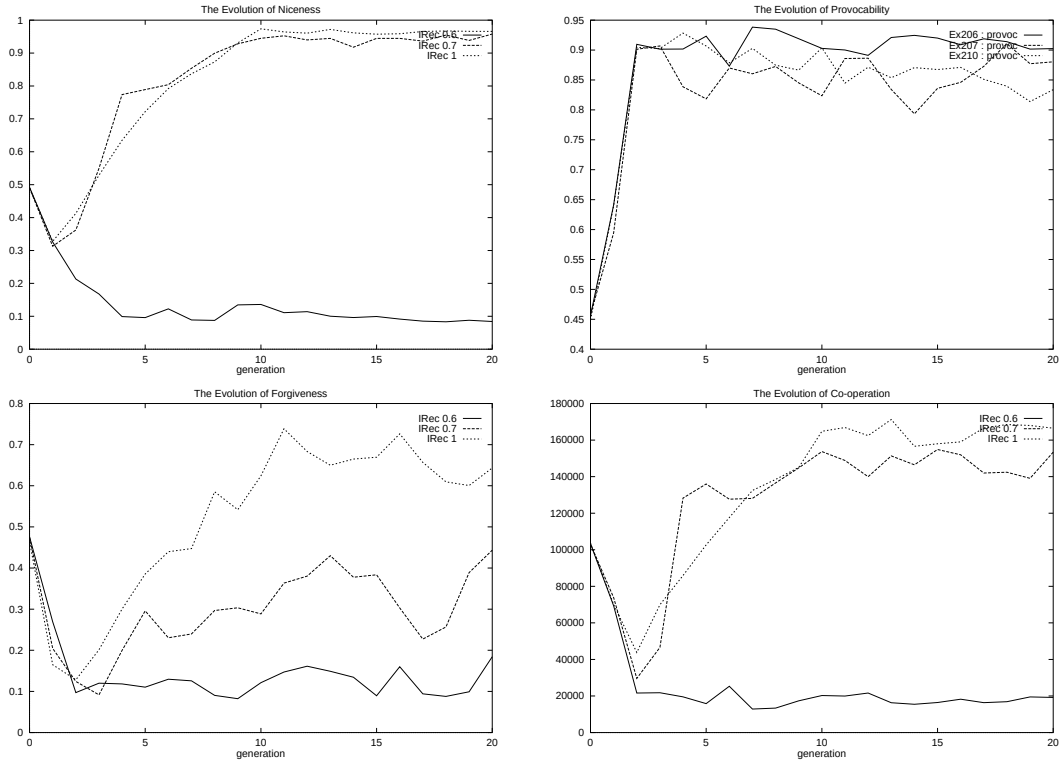


Figure 7.6: Averaged evolutionary processes. This shows the "typical" evolution of the three categories over time for three fixed values of individual recognition. Each of the lines is the average of 10 evolutionary runs. Because these populations were so reliable in their behaviour the events that led to either reciprocally altruistic or defecting populations happened at the same time in each run. If this were not the case, the diagrams would be confusing.

in run 0, where forgiveness is high, TFT (CCDDC) is the dominant strategy, although only half of the population play it. In run 2 where forgiveness is low, it is overshadowed by CDDDC (GRIM).

It would be neat to combine this with the discovery in the previous section, that forgiveness was lower when IRec was lower, and to be able to point to fluctuations in forgiveness corresponding with drops in IRec. But evidence for that has *not* been found here. Once the reciprocal society is established, IRec remains high whether in support of TFT or GRIM.

## 7.5 Conclusion

We have been able to answer the questions stated at the beginning of the chapter, using the framework 3 model. The answers are new, and interesting. But the third is disappointing. The reciprocally altruistic society can only evolve when IRec gets higher than a certain threshold. There is no interesting dynamic where IRec increases coupled with a component of reciprocal behaviour. In this model, IRec is shown to be something which has to arise through mutation before any move towards reciprocity takes place.

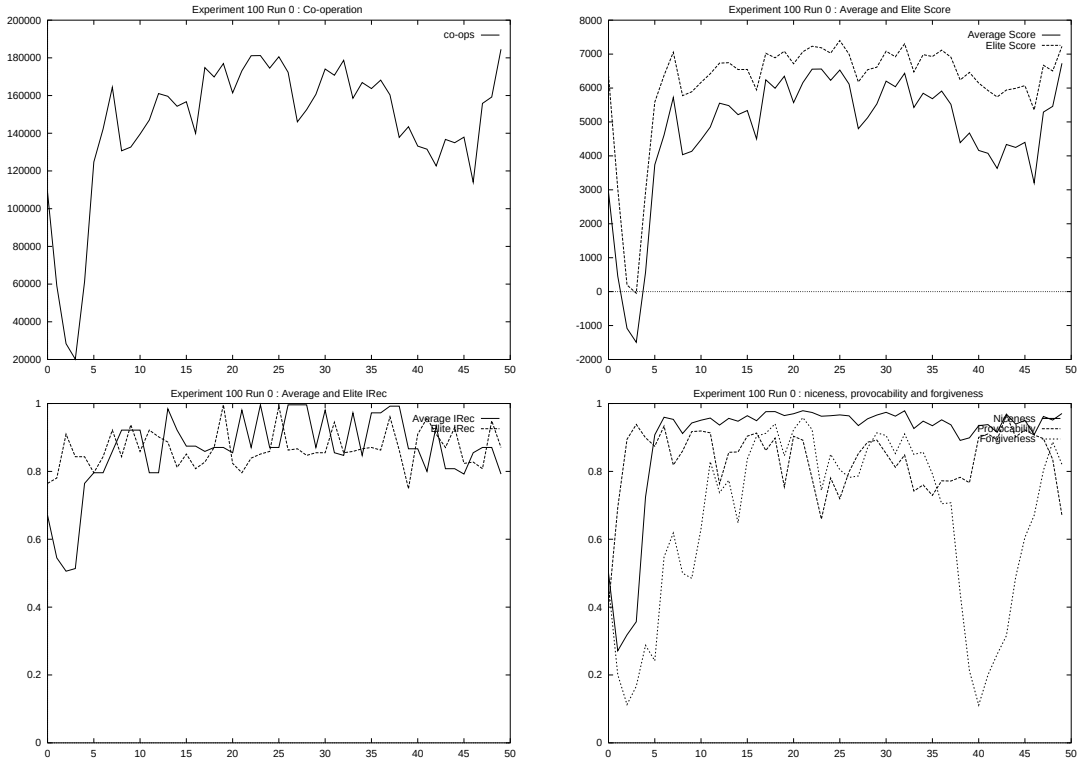


Figure 7.7: Experiment 100 Run 0 : Individual recognition invades a population.

One other interesting observation has been made. When IRec is unreliable, GRIM tends to beat TFT. Forgiveness is a luxury for those with excellent IRec capabilities. Once IRec is established societies can still slide back into unforgiving, GRIM behaviours.

| Strategy | Run 0 | Run 1 | Run 2 |
|---|---|---|---|
| CCCCC | 2 | 0 | 0 |
| CCCDC | 4 | 5 | 2 |
| CCCDD | 2 | 0 | 0 |
| CCDCC | 8 | 8 | 5 |
| CCDCD | 0 | 2 | 0 |
| CCDDC | 42 | 40 | 20 |
| CCDDD | 6 | 3 | 4 |
| CDCCC | 0 | 1 | 0 |
| CDCDC | 2 | 1 | 2 |
| CDCDD | 0 | 1 | 0 |
| CDDCC | 5 | 7 | 2 |
| CDDCD | 1 | 1 | 0 |
| CDDDC | 4 | 7 | 36 |
| CDDDD | 0 | 1 | 1 |
| DCDCC | 0 | 0 | 1 |
| DCDDC | 3 | 1 | 2 |
| DCDDD | 1 | 0 | 0 |
| DDCDC | 0 | 1 | 0 |
| DDDCC | 0 | 1 | 0 |
| DDDCD | 0 | 0 | 1 |
| DDDDC | 0 | 0 | 4 |

Table 7.1: Strategy distribution of 50th generation, Experiment 50, Runs 0,1 and 3. The strategies should be interpreted as follows. The first item refers to the move a player makes when both it and its opponent co-operated on the previous move. The second item for when the opponent co-operated on the previous move, but this player defected. The third is when this player co-operated but the opponent defected. The fourth when both defected and the fifth is how this player treats strangers.

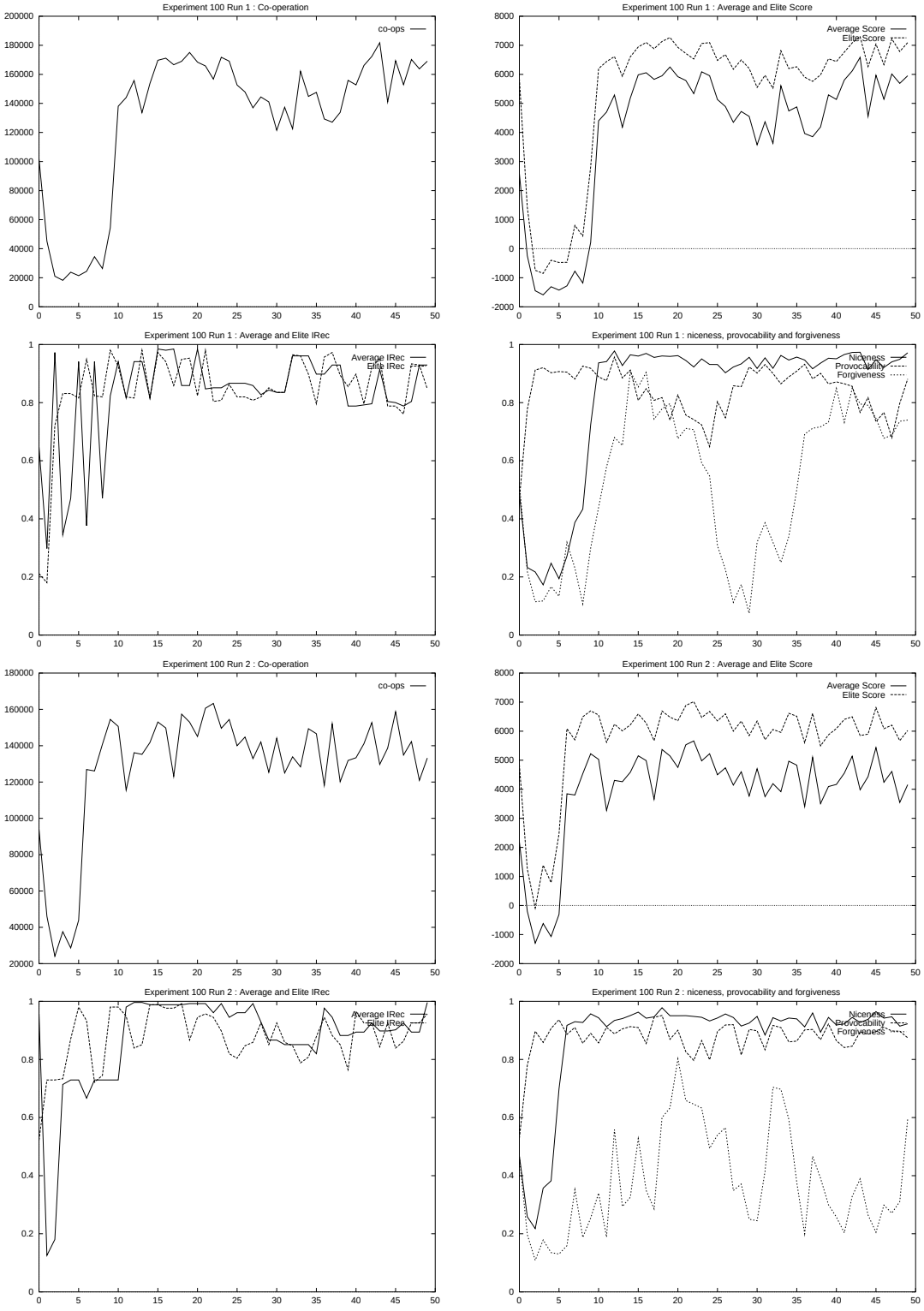


Figure 7.8: Experiment 100 Runs 1 and 2 : Individual recognition invades a population. Compare with run 0 of this experiment. Forgiveness shows the most variability.

# Chapter 8

# The role of Kin Oriented Altruism in the evolution of TFT

## 8.1 Introduction

In these next two chapters I present experiments that might shed light on the two possible scenarios described by Axelrod about the early stages in the evolution of reciprocal altruism.

These are

- initial support from kin oriented altruism; and
- support for co-operation due to the effects of locality.

This chapter focuses on the support from KOA story. The hypothesis is that the growth of reciprocal altruism (TFT) is helped by the existence of kin oriented altruism (KOA). Clearly the notion "helped" here is vague. The term can be understood in one of the following ways.

- 1) KOA helps TFT if, given two populations $p1$ and $p2$ starting with equal proportions of TFT. TFT can invade a mix of ALLD and KOA when it couldn't invade an equivalent proportion of pure ALLD strategy.
- 2) KOA helps TFT if, given two populations, $p1$ and $p2$, starting with equal proportions of TFT, TFT can invade a mix of KOA and ALLD more rapidly than it invades the same proportion of ALLD.

## 8.2 KOA and TFT in framework 1

The belief that KOA can support TFT is based on the idea that KOA raises the absolute amount of co-operation in a population. And it is this which allows the few TFT mutants to begin to score more highly than ALLDs. Is this right? Does KOA increase absolute co-operation within society?

An early indication that it might is given in figure 8.1. In this graph, drawn from the analytic (framework 1) model, the frequency of TFT starts at 0.3 (TFT is 30%) of the population and, in all but one cases, goes on to invade the population. The different lines, represent the different proportions of KOA in the initial mix.

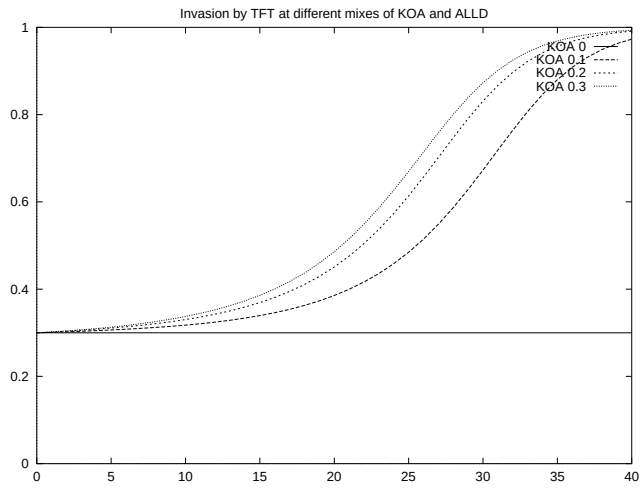


Figure 8.1: The trajectories as TFT invades a population containing different mixes of ALLD and KOA.

The only run where there is no KOA is the run where TFT fails to invade. When there is some KOA, TFT takes off. The more KOA in the initial population, the faster the invasion. In this ideal case then, both ways of taking "help" are confirmed. The hypothesis that KOA helps TFT are corroborated.

Unfortunately, this graph was produced from the original version of the analytical, framework 1 program, where KOA was doomed never to invade ALLD. Once one of the versions of the program which credits KOA with the capability of invading ALLD is used, we see a situation like that in figure 8.2. This uses the personal approach, which presumes that family members always play the same strategy. And here KOA and TFT are in competition. KOA, implemented personally has the unfair advantage that KOA players, get to treat ALLD family members as though they are KOA players.

Looking back at figure 5.6 we see that in some parts of the population, TFT does win out. But in much of the space, the trajectories are towards KOA dominated societies.

The inclusive fitness version of framework 1 behaves better, in that, eventually, TFT wins out. This is, presumably, what should happen with a plausible implementation of KOA. But there is a point where KOA demonstrably helps TFT. In figure 8.3 we zoom in to part of the ALLD-TFT-KOA space which starts with a small proportion playing TFT (about 2% ) and similarly small proportion playing KOAs.

In this model, the helping of TFT by KOA only occurs at these frequencies. Where TFT starts with a larger frequency, eg. 10%, it is already on the way to invading, and KOA strategy is merely an impediment.

But while corroborating the basic "KOA can improve co-operation and hence allow TFT to invade" story, this example has also challenged the more detailed misrecognition story. In this model, there is no kin misrecognition. The extra co-operation is getting going through pairs who

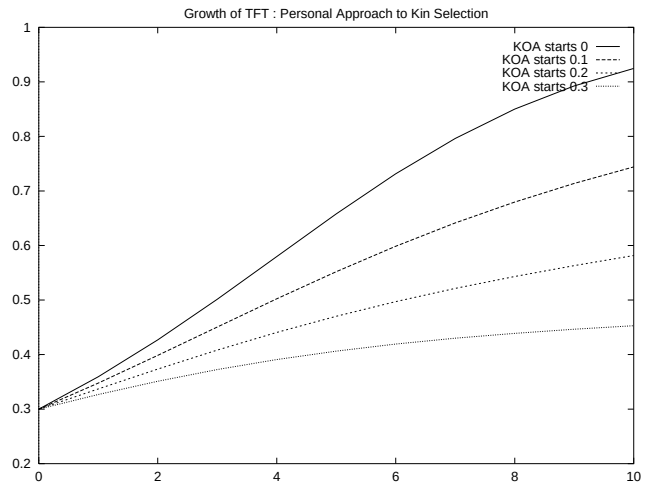


Figure 8.2: The trajectories as TFT invades a population containing different mixes of ALLD and KOA.

are related and where one member is playing KOA and the other TFT. Such pairs will co-operate.

## 8.3 KOA and TFT in Framework 3

To look at Axelrod's kin misrecognition story, we must use a model with kin recognition and kin misrecognition. The complex model also does not suffer the same problem of representing kinship as as the simple models. Here players of each generation are explicitly the children of two parents of the previous generation. Strategy is encoded in the genotype, which is derived from the genotypes of the two parents by a cross-over operation. The same genotype is also used to calculate the apparent relatedness of two individuals, which guides their kin oriented strategy.

### 8.3.1 Kin recognition and co-operation

*Does kin oriented strategy increase co-operation?*

On first observation the effect of kin recognition on overall co-operation is unexpected. Look at the left hand graph in figure 8.4 to see two views of co-operation against kin recognition in a purely kin oriented society. (Graphs are averages of 10 runs.)

Kin recognition seems to have a negligible effect on absolute co-operation, which is much the same at maximum kin recognition as at none. But, between the two we see an increase in co-operation, followed by a sharp drop. The lowest co-operation is at the point where recognition is almost but not quite perfect. This is a strange phenomena, our intuition is that kin recognition should increase co-operation. Otherwise how might it have evolved if it is not rewarded?

First, let us not be deceived by the scale of this left-hand graph. The absolute number of co-operations is small compared to the number of defections. (See the right-hand graph in the figure for the comparison.) So, could it be that kin recognition has no effect on behaviour?

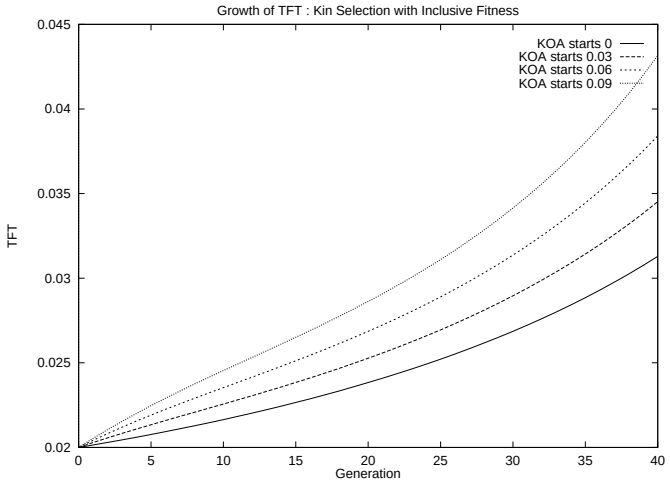


Figure 8.3: The trajectories as TFT invades a population containing different mixes of ALLD and KOA.

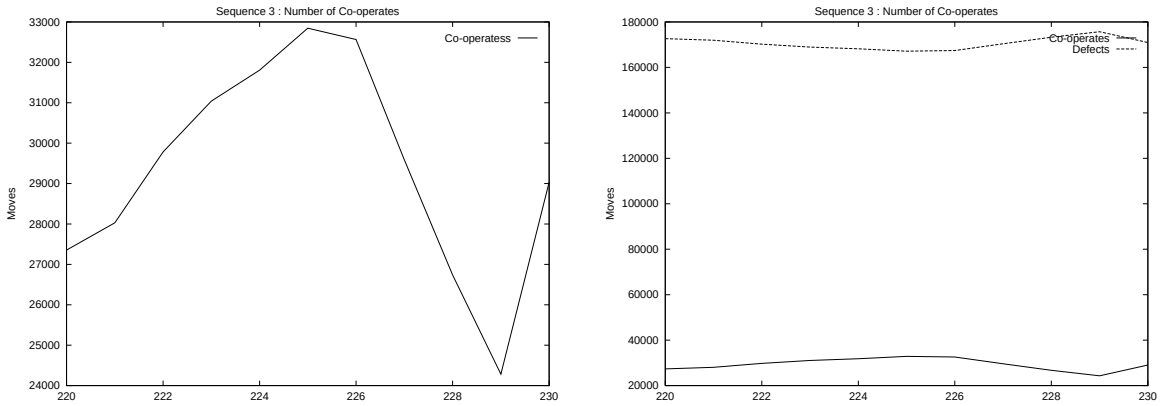


Figure 8.4: Co-operations as Kin Recognition varies from 0 to 1

Perhaps the graph is uninformative because the overall number of interactions with kin are low. To get a second opinion, we can look at the players' internal strategy tables and see if we can find some tendency towards kin oriented altruism there.

### 8.3.2 Evolving kin oriented strategy

*Is there KOA in the model?*

The players in framework 3 have a separate strategy table for dealing with kinship. This table has 11 entries indexed by degree of similarity between 0.5 and 1. Each position in the table contains a probability of co-operating with an opponent of that similarity. Thus the table represents the player's intentions towards those with a family resemblance. I shall call them a *characteristic profile* of kin oriented strategy for a player. A typical (consensus individual) profile after 100 generations, is shown in figure 8.5.

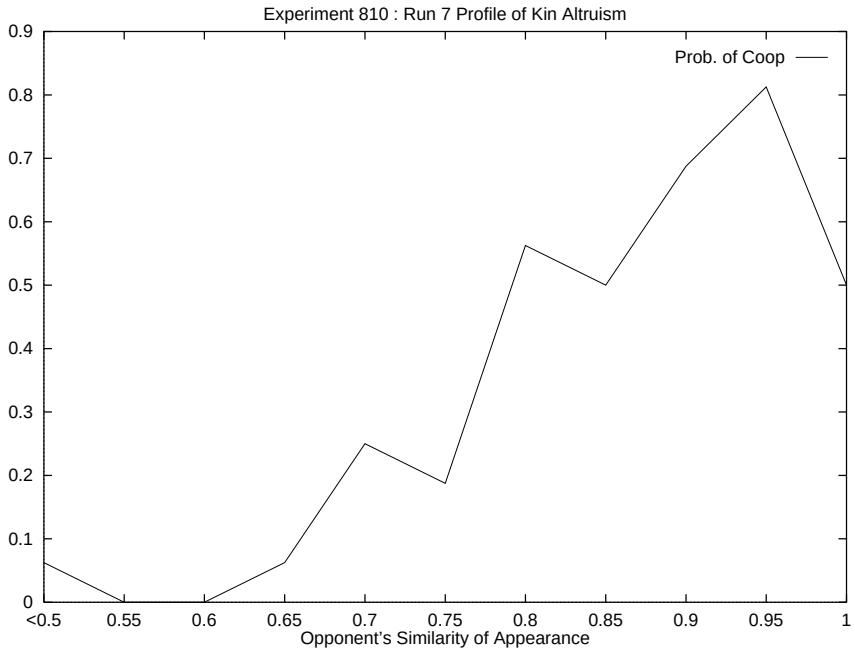


Figure 8.5: Experiment 810 : Run 7 : Profile of Kin Oriented Strategy after 100 Generations

There clearly *is* a definite bias towards co-operating with similar opponents. When the profile is extruded as a third variable is changed, we can get quite a strong visual impression of the relation between the degree of kin oriented altruism and that variable. Figure **??** shows the average characteristic profile of kin altruism that has evolved as KRec is varied between 0 to 1.

Where KRec is low, the probability of co-operating has been minimised. It is not to be trusted. As it increases we see a willingness to co-operate, starting with the only the most similar, and coming down to lesser degrees of relatedness.

## 8.4 KOA strategy in the evolution of TFT in framework 3

Unfortunately, this thesis does not contain an illustration of Axelrod's story of TFT arising, triggered by a failure of kin recognition. Either the story is not true or, as is my belief, there are limitations to the framework 3 model which mean that it couldn't capture such a history.

I will discuss this failure in the conclusion, but the problem seems to be one of competition between KOS and IOS within players.

There *are* some few hints we can get from the framework 3 model however.

*In framework 3, can we see kin recognition supporting the growth of co-operation?*

The reader may be thinking "Hold on, haven't we previously been shown that KRec doesn't increase co-operation?" This is so. But look at figure 8.7. Here the players use an equal mix of kin and individual oriented strategy. The graph shows the amount of co-operation as individual recognition is increased. Most noticeable is the fact that the gap between no kin recognition and some kin recognition is significant, but the difference between half and full kin recognition is

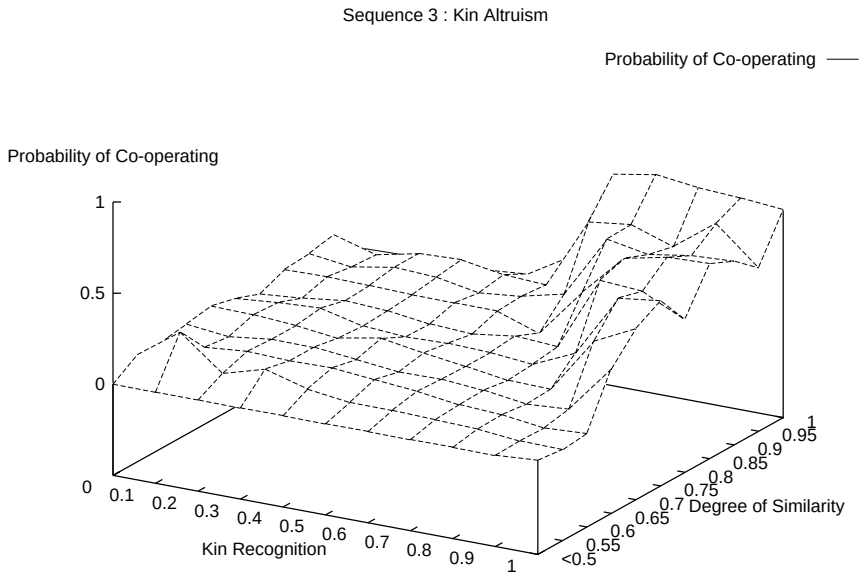


Figure 8.6: {charprof2
Evolved Kin Oriented Altruism as Kin Recognition varies between 0 and 1

trivial.

What has happened to the expected jump to reciprocal altruism at around IRec = 0.7? The answer is that this population never makes it to reciprocal altruism. This graph is a close-up look at an unco-operative society. The reason this population is doomed not to become reciprocal is that the mix of using KOS and IOS too heavily favours KOS. Looking at the co-ops (figure 8.8) at different mixtures when both individual and kin recognition are perfect, we see that at more than 20% kin oriented strategy, reciprocal altruism is unsustainable.

But the most interesting result being shown here, with regards to the Axelrod story, is that *niceness* is higher when KRec = 1 than when KRec is 0.5. This is extremely contradictory to the hypothesis that failure in kin recognition allows spare co-operative behaviour to leak into the wider population.

## 8.5 Conclusion

What has been shown in this chapter?

- First, that there are regions of ALLD-TFT-KOA space where the existence of a small amount of KOA can mean the difference between the ultimate success and failure of a minority of TFT.
- But this demonstration also shows that there is no necessity for a lapse in KRec for this KOA to help the take off of TFT.

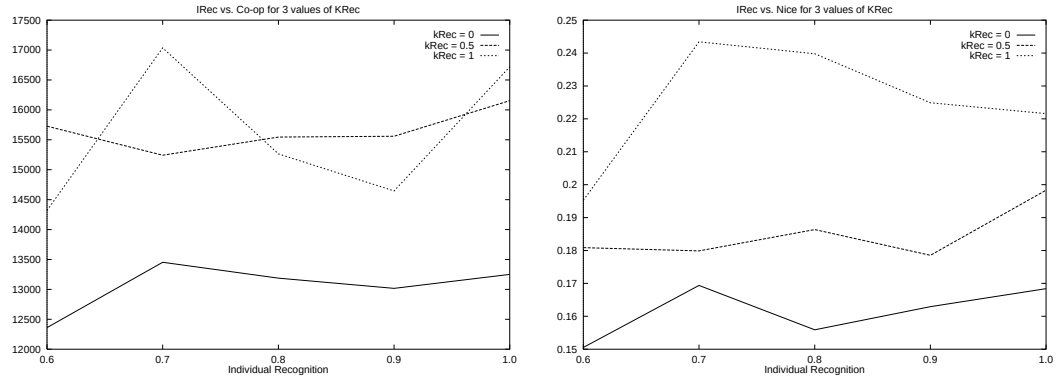


Figure 8.7: Co-operation and niceness plotted against individual recognition. Each line represents a different value of kin recognition. Note that these graphs show a low incidence of co-operation however. These populations have never achieved a state of reciprocal altruism.

- Framework 3 provides contradictory evidence. On the one hand, plotting co-operation against KRec when all players use KOS, shows that overall co-operation doesn't seem to increase between KRec = 0 and KRec = 1. But within the range there is an interesting pattern, a maximum when KRec is around 0.5 and a dramatic minimum at around 0.9[1] This suggests that some degree of kin misrecognition might increase co-operation. Though also, that a degree might lower it.
- But further counter evidence from running the model with a mixed strategy of KRec and IRec players. Here KRec is found to increase the co-operation and niceness in the community, but perfect kin recognition doing so more than imperfect.

The work presented in this chapter highlights a weakness of the framework 3 model which does not allow meaningful interaction between kin oriented and individual oriented strategy. *The two strategies can not be played simultaneously.* Any particular decision that a player makes must be due to either KOS or IOS. The way to mix the strategies is to probabilistically play one or the other. But reciprocally altruistic society, while robust in the face of invasion from rival strategies, is brittle with respect to both uncertainty due to unreliable recognition, and uncertainty when faced by an opponent using a mix of strategies. Co-operation is unsustainable when IRec is less than 0.7 or individual oriented strategy is used less than 80% of the time[2].

But if IOS needs to be used 80% of the time before reciprocal altruism becomes viable, there is a danger that the selective pressure due to KOS is effectively swamped.

[1] This is a pattern that has been found several times. It may be an artifact of the implementation of kin misrecognition in framework 3. Using one alternative, the pattern disappeared.

[2] Though, of course these numbers apply only to this model.

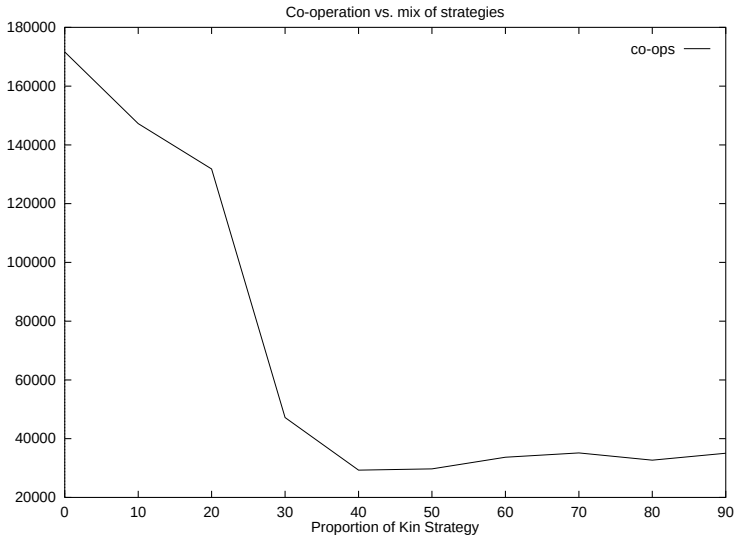


Figure 8.8: Co-operation in a range of populations where players use a mix of kin oriented and individual oriented strategies. Reciprocal altruism can only survive when individual oriented strategy is played 80% of the time or above.

# Chapter 9

# Games on a Spatial Grid

---

We believe that spatial and similarly restricted versions of the prisoner's dilemma in viscous worlds increase co-operation. But does this imply a stronger likelihood of the evolution of individual oriented, individual recognising strategies or does it just imply more indiscriminate co-operation?

## 9.1 The grid world variant on the framework 3 complex model

### 9.1.1 The spatial matching algorithm

- Each player only gets to play *n* matches, with each of the occupants of the eight, adjacent, neighbouring cells.
- The grid is toroidal (has wrap-around), so edges are considered adjacent to the opposite edge. Consequently all cells have eight neighbours.
- All players get *8n* matches, with *n* per pair.

To clarify this third point: the players are matched *n* times with each neighbour but the sequence is that each player plays a circuit, of every neighbour once; then repeats the circuit until each neighbour is played *n* times. This ensures that the players meet all of their neighbours before repeat matches take place; maintaining the possibility of mis-recognition.

### 9.1.2 the spatial breeding algorithm

Breeding is also handled differently in the spatial game. In the non-spatial version, the population comes in discrete generations. While it would have been possible to produce a *steady state* model where players were chosen to be replaced at a constant rate, I wanted continuity with the previous model, so that comparisons could be made, The breeding strategy is very like that of the non-spatial game; except in its geographical restrictions.

The algorithm is as follows.

```
For each cell c.
   Find the scores of all eight neighbours.
   The new occupant of c will be the child of the two highest scoring neighbours.
```

Where members of the previous generation survive, they retain their location. Strangers are added at random locations. Being set in a toroidal world, there is no notion of an edge to the population.

### 9.1.3 Population Maps

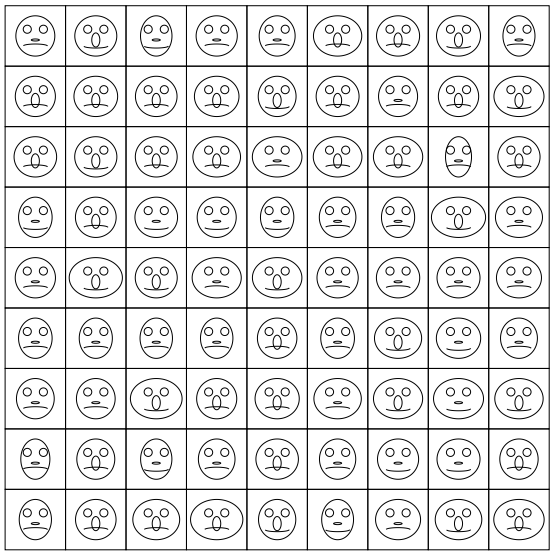

Figure 9.1: Experiment 9000 : The opening generation of a game played only with individual recognition. Individual recognition is 1 and there is no cost.

Sometimes on the grid world, it is necessary to get an overview of the types and distributions of strategies. I use a *face map* where several of a players' characteristics are represented by facial features. The "fatness" of the faces represents the average score per move of the player. The fatter the better. The mouth smiles in a nice strategy and turns downwards otherwise. The nose is long (and hawk-like) when the player is provokable. See figure 9.1 for an example of the initial population.

## 9.2 Invasion by Co-operation

*Can co-operation invade a spatial world when individual recognition is full?*

In figure 9.2 we see a small population after 6 generations. Players here have full individual recognition and within six generations a TFT strategy has pretty much taken over. We can see that the players are smiling, fat and happy. The fatness represents a high score. The smile shows

these players are nice; while the long hawk-like nose is an indication that they are still provokable. Those players who are not nice are the definite losers in this population.

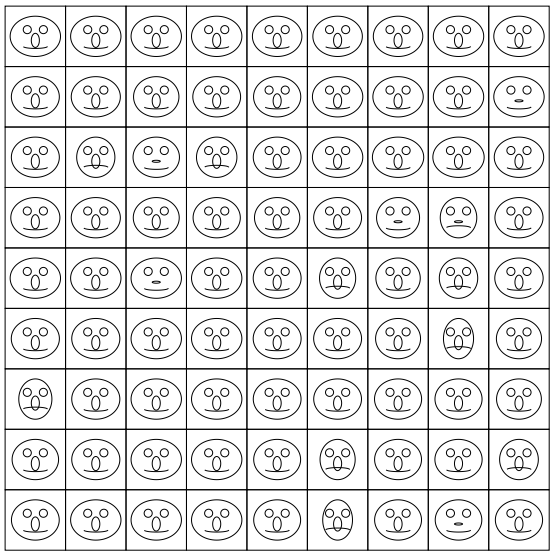

Figure 9.2: Experiment 9000 : Generation 6

*Can co-operation invade a spatial world when individual recognition is low?*

The sorry bunch in figure 9.3 have only minimal (0.1) individual recognition capability and are consequently hollow cheeked from continuous defecting against each other. Interestingly, some of the highest scoring players are nasty but not provokable. Perhaps these are Anti-TFT (defect against co-operate and co-operate with defector) who maintain small pockets of rather arbitrary co-operation.

To answer the question then. Very little co-operation takes place when individual recognition is low. But perhaps, even in this simple example, some has found a purchase. The amount is not significant (see fig 9.4).

## 9.3 IRec and Co-operation

*What is the profile of IRec on Co-operation?*

Figure 9.4 shows how co-operation increases with IRec. If we compare with the non-spatial world in 7.4 we will see that the jump to co-operation is at a lower value of IRec. (All parameters are the same.) ]

## 9.4 Discriminate or Indiscriminate Co-operation?

*Does spatialization promote greater indiscriminate co-operation?*

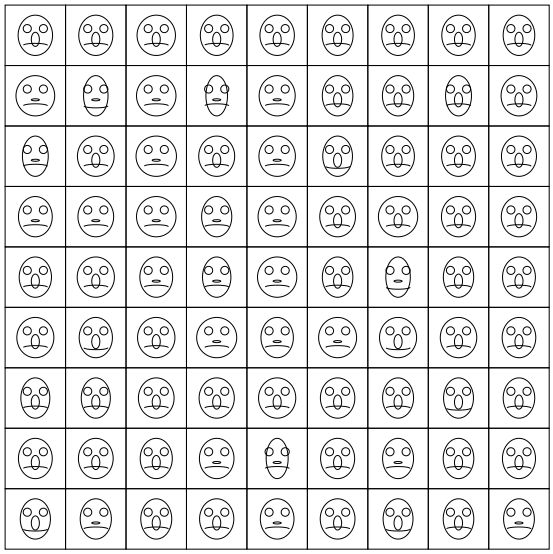

Figure 9.3: Experiment 9001.0 : Generation 6

Possibly the most indiscriminate co-operation seems to occur at the IRec = 0.9 mark. In figure 9.5 we see a predominance of TFT. However there are also sizable enclaves of ALLCs (Smiling, button noses indicating lack of provokability)

If we look at provokability as IRec increases in figure 9.6 we see a different pattern from the non-spatial world in figure 7.4. When a population in the non-spatial game make the transition to altruism provokability remains high. Here, in the spatial game, we can see that it appears to be falling off as the generations progress. At generation 6 it is still high, but by generation 30 it has dropped. But note two caveats. This graph is scaled between 0.5 and 0.95, so provokability hasn't disappeared. Also provokability seems to be higher when IRec is full.

## 9.5 Conclusion

What have we noticed in this brief visit to the grid-world?

- The shift to co-operation is still a sudden jump. It does not look as though pockets of TFT are spreading across the grid (although the grid may be too small to judge whether that would occur in other circumstances.)
- We have some evidence, through lower provokability, that co-operation is less discriminate here.
- The shift to co-operative society occurs when individual recognition is lower than in the non-spatial case.
- Non-provokable, indiscriminate co-operators do clump. Though the grid may not really be large enough to see this effect strongly.

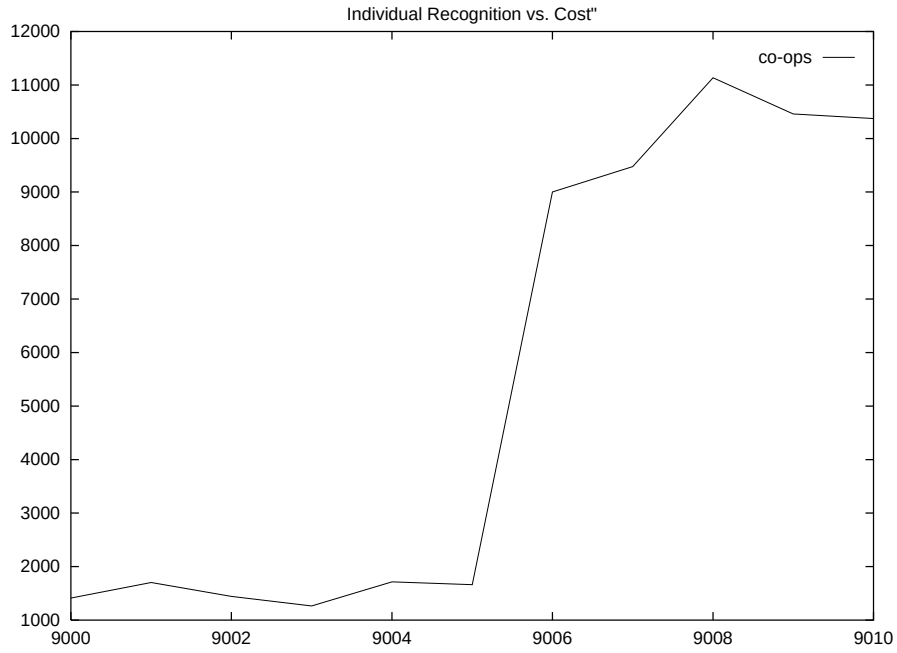


Figure 9.4: Evolution on a spatial grid.

- Non-provokable, indiscriminate co-operators are more widespread where IRec is imperfect. This could be evidence of greater generosity required in a noisier environment (See also [18]).

An unanswered question remains. When IRec evolves on grid world, would it perhaps stabilise at a lower level than in the non-grid world? This question should clearly be the next to be researched.

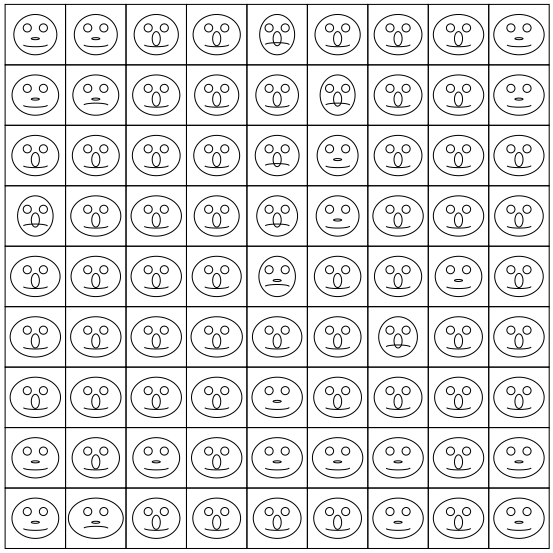

Figure 9.5: Experiment 9009.0 : Generation 6 : Individual Recognition is 0.9. Perhaps these simply survive due to the benign environment, or maybe their extra leniency is valuable in a world where recognition is occasionally mistaken.

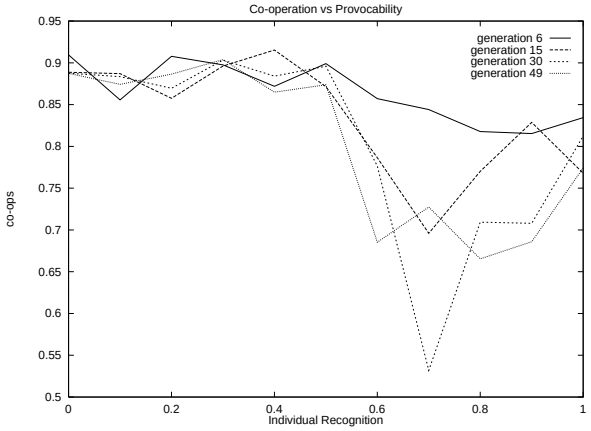


Figure 9.6: Provokability as IRec increases in a grid world.

# Chapter 10

# Discussions

## 10.1 Introduction

This chapter contains discussions of several aspects of the work described in this thesis. I hope the reader will grant me, in this place, a few brief words of meta-comment. The reader is aware that this is a re-submission of work presented earlier. That previous work had many faults, the most grievous to my mind being that several of the avenues of experiment had not been taken to their planned conclusion. Nevertheless in revising this work, I have been guided by the principle that I was not to go further and to try to produce other experimental results, but to stay within the boundary of the original work completed. Instead, the intention has been to recontextualize and better explain those experiments already undertaken. For this reason, some interesting results that were originally mentioned have been excluded as they did not really address the main questions. But there are some obvious questions that are very pertinent to completing the enquiry which are left unanswered here.

I am also conscious, that the necessary implication of these choices is a certain retrospective justification. Sometimes I have used ideas discovered in the revised background reading, or from the results of running the new, simpler analytic model, to justify some aspect of the original, framework 3 model. The reader familiar with the original presentation will know this not to be the true historical ordering. Nevertheless I hope the reader will allow this in the interest of supporting a more coherent explanation.

The rest of this chapter covers the following issues.

- On the choice of a notion of *individual recognition*
- On the results.
- On the limitations of the models.
- On the scientific status of Artificial Life.
- On future work

## 10.2 On the choice of a notion of *individual recognition*

*The distinction between what it is to be an individual or a particular instance on the one hand, and what it is to be general, repeatable, or common on the other, is as fundamental as any distinction in philosophy. It is central to the conceptual scheme with which the human mind operates.*
Milton K. Munitz[30]

This talk of conceptual schemes warns us that the above quote is from a philosophical perspective. Philosophy, particularly of the cognitive variety has had seat at the table of *cognitive science*, along with linguistics, cognitive psychology and artificial intelligence, since its first inception in the 1950s. Cognitive science arose, partly as a reaction to the earlier *behaviourism* in psychology, which asserted that only behaviours were suitable entities for scientific investigation - as only behaviours could be observed objectively. In contrast, cognitive scientists felt that it was legitimate to posit certain unobservable *mental* entities which, although hidden from observation, guided behaviour and could be inferred from it. Candidates for such hidden entities or *cognitive innards* ranged from short term memory buffers, through to beliefs and desires, through to special language learning modules. This range reflected the diversity of disciplines on which cognitive scientists drew for inspiration in their their hypotheses. Philosophers brought the idea of mind as possessing a scheme of interdependent concepts or categories, with which it interprets the world.

It is curiosity about this conceptual framework, and what seems to be the fundamental distinction between conceptualising the world in terms of individuals and conceptualising it in terms of classes, that has motivated my research from the beginning[1] At the beginning of this thesis I surveyed the approach taken by ethologists towards individual recognition. And clearly, were one to start from their position, working with the limits of what could be inferred from behaviour, one would focus more on either *a plausible mechanism* of individual discrimination and recognition; or a known *continuum of ecological significance*. In the first case, one would model perceptual apparatus and the cues of identity; and study the evolutionary dynamics of how the mechanism became refined enough to discriminate the cues. In the second case, it might be apt to consider a situation such as discrimination by parents of chicks in the nest. It should be possible to show how the capacity to distinguish one's own offspring from others' could eventually evolve into distinguishing individuals.

The rather abstract and conceptual notion of individual recognition that I have in mind has turned out to be highly problematic. I have yet to find a continuum of ecological significance, along which, another capability could be gradually drawn until it became individual recognition. The other capacity, considered here, namely kin recognition, *has* been shown to enable the evolution of a reciprocal altruism. So in a sense, one might say that this is a path of ecological significance. A need to evolve kin recognition would set up the conditions under which individual recognition tied to reciprocal altruism could thrive. Were it also possible to show that the mechanism of KRec could be refined into the mechanism of IRec, then an even richer story could be told. But the

[1] And quite possibly led me astray.

work here also signals the problem that, in as abstract a social game as the prisoner's dilemma, kin oriented strategy and individual oriented strategy can demand different responses : when a player is confronted by a defecting relative or an altruistic stranger.

The Axelrod story, which can be thought of as trying to smooth a path which can be followed between the two behaviours, could not be demonstrated here. But this may be due to weaknesses in the actual model. One clear suggestion for future work would be to look beyond the prisoner's dilemma for some social game that might allow KOS and IOS to be more aligned. Such a game would have to allow for divergence at some point (or the behaviours would be the same), but might only require this separation when individual recognition was quite mature.

## 10.3 On the results

The actual results achieved are these.

- The model demonstrates that reciprocal altruism is highly dependent on individual recognition.
- Furthermore, given a model of misrecognition which is intuitively plausible though not well grounded in biological observation[2], reciprocal altruism turned out to require IRec to be above a threshold. There seems no possibility of part IRec being traded for part reciprocal altruism.

  I conjecture that this all or nothing quality is a feature of the prisoner's dilemma situation, rather than a quirk of my implementation of misrecognition, and consequently we should expect to see it in any reciprocally altruistic circumstances. However, it also means that those alleged co-operative behaviours which are *not* really prisoner's dilemmas may not suffer the same fragility.
- The presumption that a level of non-reciprocal altruism could help a TFT-like strategy to invade was demonstrated in the simple case where TFT and IRec were shackled together. In the complex model where IRec has to evolve before TFT, it was not demonstrated.
- Finally, we confirmed that a spatial world might be friendlier to co-operation; but also that this counted against provokability which might in turn relax the pressure towards discrimination of individuals.

## 10.4 On the limitations of the models

The main model presented here is the one I have referred to as framework 3. It was designed to have a large number of "moving parts", parameters which can be set or evolved. When designing it I envisaged being able to run with most of these parameters evolving freely, for long periods and with large populations. Hence the results would be obtained by mining large datasets. However the program is inefficient. And in practice it is necessary to run with smaller populations and for short periods, keeping many of the parameters fixed. The effect of this is that points tend to be illustrated with anecdotal evidence, often found by trial and error, sifting through many uninteresting runs by

[2]Further research on this would be interesting.

eye. This means that parameters were not chosen by any systematic method but simply according to which produced a result about which I could tell an interesting story in the light of one of the hypotheses under consideration.

I don't regard this use of anecdotal evidence as wholly inappropriate. Such results can still turn out to be the counterexample that successfully falsifies a hypothesis, thus driving the experimenter to come up with a new one.

### 10.4.1 Particular problems and their solutions

The small population size might be responsible for artifacts in the results. For example in a large population, small increases in co-operation due to infrequently tested KOA might still be able to help TFT take over in circumstances it would otherwise fail in. The framework 1 model shows that there could be significant regions in the KOA-TFT-ALLD space where the proportions of each strategy were around 2%. The populations of 50 players used in my experiments would effectively filter out such regions.

The other serious problem when looking at IOS and KOS is the reliance on a single game, the prisoner's dilemma. The simplicity of this model is in the range of behaviours allowed to players. There are no intermediately good behaviours between ALLD and reliably recognising TFT.

## 10.5 On the scientific status of Artificial Life

The ALife community is very aware of the conceptual issues surrounding Artificial Life and its scientific status. It is alert to a criticism that compared to biology it is amateurish and that simulations are not scientific experiments. See, for example, papers by Geoffrey Miller[28] and Jason Noble[3][31], the latter of which frames three questions about the status of such research.

1. Is it science?
2. What does it study?
3. What constitutes good practice?

His conclusion is, broadly, that ALife should be respectfully attentive to existing work in theoretical biology, effectively becoming its modelling arm. But he makes some further analyses of what ALife is attempting to achieve. For Noble, Alife simulations are built to examine what he calls the *analytic*[4] implications of a theory.

One may hypothesise, for example, that a group of simple behavioural layers, when interacting together, will give rise to a specific overall behaviour. Observing this in action, by running the simulation, is little different from performing any other symbolic transformation on a mathematical

[3] I am grateful to Jason, Ezequiel (Di Paolo) and also to Hilan Bensusan, Darius Sokolov, Peter Elliot and Ron Chrisley who have all helped me develop my understanding of these issues.

[4] By which he refers to a philosophical sense, something like being without empirical content. Note that Ron Chrisley points out that Noble's use of analytic can be challenged from within mainstream philosophy. For example Kant's notion of the *synthetic a priori* might violate the dichotomy that Noble wants to make between analytic and synthetic. Also following Quine's[36] away from the distinction would take away much of it's force - as Noble *should* know. However my own disagreement doesn't come from this angle.

description, and adds no new information about the world. It consequently, does not count as an *empirical* observation. What it will do is show whether the emergent property really is logically derivable from the simple behaviours. By itself this is not a complete scientific activity, as the hypothesis still needs to be *tested* in the real world.

Given that it is an attempt to discover whether one's initial hypothesis is coherent, it is a very desirable thing that the program *does* correctly implement the behaviours about which one is hypothesising. Hence, an example of good practice is a clear specification of what theory is being tested, or at least what the micro-component behaviours are. That way, other researchers can re-implement the model, controlling for possible bugs in one's own implementation, and contingencies of one's computational resources. Particularly, he states : "To do bad analytic AL is to write a computer simulation that fails to capture all and only the intended assumptions ". He goes on to define further criteria for a *synthetic* mode of ALife. "it is no easy thing to generate good empirical hypotheses ... good synthetic AL is marked by the ability of the researcher to recognise likely correspondences between real world phenomena and the emergent results of a simulation." An activity which requires knowledge of the real domain that is to be investigated.

We can summarise this position by answering the three Noble questions thus :

1. ALife is *half of a science* in that it is that part of science concerned with creating theoretical models of phenomena, working out their internal consistency and their implications.
2. It studies the same phenomena as biology[5]
3. Good practice is the creation of good theoretical models, and ensuring a high rate of information exchange with biologists who will provide empirical observations to test the hypotheses.

In contrast, I think that Alife artifacts are neither models nor simulations of things that we are interested in, but actual *examples* of them. This is a position which I will call *strong*, in roughly the same sense as *strong AI*[39], and one which I plan to lay out a defence of in the next few sections. By way of an outline (and manifesto) I will answer the Noble questions like this :

1. Alife is a science in that it lives up to two genuinely useful criteria for the demarcation of science.
    - It makes hypotheses that are falsifiable by observation. Importantly, this is the observation *of the artifacts themselves*.
    - It makes claims about the relationships between natural kinds rather than historical particulars.
2. It studies artifacts which have functions and behaviours in common with those real-life entities that are studied in biology and psychology. The artifacts are, by definition, artificial. The functions and behaviours are as ontologically real, as those studied in biology.

[5]This isn't altogether true. Both Di Paolo, and I believe Noble, recognise ALife's close connection with investigations into the dynamics of complex systems and seem willing to extend the umbrella to cover areas that investigate similar underlying principles. Hence almost any mathematical research into complexity could be presented at an ALife conference.

3. Good practice, as in all science, is constituted by bold and imaginative conjecture, and openness to criticism. In ALife terms, the first of these is exemplified by the creation of inspirational concept demonstrators, while the second includes close observation and interpretation of particular models.

The key point here is the answer to 2, that the objects of investigation are members of classes which are individuated by their behavioural or functional properties. Once this argument is accepted, the broad point 1 becomes quite uncontroversial. The details of the demarcation of science and good practice, can be accepted to taste.

To defend position 2 I will do the following :

- Present the explanation of functionalism in terms of its AI / cognitive science history.
- Point out that biology is itself functionalist in many ways.
- Attempt to defend a strong position against one of its more coherent critics.

### 10.5.1 The argument for strong ALife

Strong ALife is possible because in Alife, as in AI, one can be a *functionalist*. That is one can organise the objects of enquiry into classes according to their functional properties without caring about their material or physical properties. The important feature of a functionalist approach is that it allows that the object of enquiry is *multiply instantiable* in different substances.

Functionalism is licensed in ALife because much of evolutionary biology is *already* imbued with talk of function. In the program known as *adaptationism*, the function of a trait is the focus of attempts to explain its existence. Despite prominent criticisms and suggested alternatives, this is still the mainstay of evolutionary investigation. Furthermore, cross-species comparisons are encouraged in biology, and this can be seen as an acceptance that traits are themselves *multiply instantiable*.

Consequently, unlike the case of artificial chemistry, artificial biology should have no qualms about accepting that some traits, particularly but not exclusively behavioural ones, are multiply realizable and hence allowing that the entities in computer programs can realize these traits. Extending this cross-species comparisons to species *in silico* is quite acceptable. But what it implies, is that our artificial populations are neither models nor simulations but actual ALife stuff which is being experimented on.

### 10.5.2 Functionalism

Attempts to formulate a science of the mind come up against the problem that experience is a private matter. it is thought that, no one else can feel your pain or even see what you believe. So when differences of opinion arise in an analysis of the structure of the mind; there seems little that can be done to resolve the matter. At the beginning of this century, this problem led to *Logical behaviourism* which posits that statements about the structure of mental innards, or mental entities such as beliefs and desires, are merely a short-hand summary for statements about observables

such as behaviours. "Wants to hold X" is really a rephrasing of the statement "Will reach to grasp X" in the appropriate circumstances. And other mentalistic language can be equivalently rephrased in behavioural language.

Behaviourism was challenged by rival theories of mind from a neurological perspective. An *identity theory* supposes that a mental term designates something about the brain and nervous system. There are two flavours of identity theory.

A *type identity theory* claims that a type of mental event is also a type of neural or brain event. So, we can, in principle, discover that there is a type or class of neural firing patterns; which you, I and every other person, even the apes, have whenever they have a wanting to hold the cup mental event. Both the statement in terms of the mental language and the statement in terms of the neurological language refer to the same physical event. The implication here is that in principle we could discover psycho-physical laws that relate types of brain state to types of mental state.

A *token identity theory*, in contrast, agrees that the statement "I want to hold this cup", an instance or token of a mental event, is a reference to a physical state. The difference however, is that while we could indubitably generalise from all the mental events to get types of mental event; and generalise from all the physical events to get types of physical event; these two classes do not match. The class of "wants to hold X" mental descriptions does not map onto any class of physical events that could be generalised within a physical language. In one individual it might be a spiking pattern P in region A of the brain; in another it might be spiking patterns of Q in region B. Nothing in the physical descriptions of the brain would lead us to consider that these patterns should be grouped together. Furthermore, there is no possibility of discovering psycho-physical laws.

But perhaps something might convince us that P in A should be grouped with Q in B. What that something might be is not to do with the physical properties of P in A and Q in B but with their behavioural interactions. Perhaps in their relative contexts both P in A and Q in B lead the agents possessing them to reach for the X. Both states are in the causal chain that lead to their holders acting in the same way. But further, it turns out that P in A and Q in B lead not just to the reaching for X in context C1; but also the holder responding to the question "Would you like to hold the X?" in the affirmative in context C2 and the same appropriate behaviour in a slew of other contexts.

This is a *functionalist* theory of mind. We take mental terms to refer to physical states which are classified by their functional roles within the mental economy. To say that an agent "wants to hold X" is to say that it has a physical state (such as a firing pattern P in region A) which makes the appropriate causal connections with other physical states; such as the physical state with the function of tracking the position of the X. If all the functional states are wired up in the right way; not only do they allow the agent to perform the appropriate action; but it is exactly by being able to cause the agent to perform the appropriate action; that they become the states to which mental statements refer.

One of the attractive features of the functionalist account of mental entities is that it allows for multiple realisability. If what it is to be a "desire to hold an X" is just to make the right causal

connections with other beliefs and desires (and some perceptions and action causing units) then there is no necessity that such states be realized by human or even animal brain-stuff. A computer or robot with the right architecture can also have states which can be described mentally. A computer really can "have" a mind.

This makes it possible for experiments on mind to be done in a computer. Artificial intelligence is a research program which is seen to allow *empirical* investigation into the mind. One can make a hypothesis that, for example, a particular behaviour arises due to an internal mental capacity whose structure S consists of a set of a set of interrelated functional states.

These functional states are created and connected within the computer, and by observing their interaction it is possible to see whether they really give rise to the behaviour or not. If they do, we have a *concept demonstrator*; if not we may have a falsification of the claim that the functional architecture is sufficient for the behaviour.

### 10.5.3 Function in Biology

It is my contention that much biological research is also functionalist. But can I just assert this? What does it mean?

It is clear that biology is unlike the physical sciences in that it introduces functional, and purposive, notions into scientific explanation[6]. There is no similar purposive talk in the physical sciences. Atoms and molecules are not *for* things, unless we have designed them that way. But biological traits are explained by reference to a problem they solve; or a niche of opportunity they allow the exploitation of. Wings, quite definitely seem to be for the purpose of allowing birds, or bats, or insects to fly.

Furthermore, the fact that they allow their owners to fly seems to be the best explanation of their continued maintainance. It's true that flightless birds have some sort of wing; and perhaps the possibility of any particular species having wings may depend on some morphological quirks of the ancestral line. But flight is by far the most significant factor in explaining the wing's existence and upkeep.

Theories of function fall into two main classes. *Etiological* definitions of function are those which count a trait's evolutionary history as constitutive of its function.[7] So, very roughly, *flying* is a function of *wings* of this *bird* because having wings that were good for flying was a significant fitness advantage to the ancestors of this bird. Or more generally : trait T has function F for organism O in lineage (or reproductively established family) L because F-ing using the T was a significant influence on the fitness of ancestral members of lineage L.

The alternative, *teleological* idea, is that function should be identified with having the right set of causal capabilities within a particular context. Hence, flying is a function of wings just because

[6]See a very useful book for discussions on function : "Nature's Purposes : Analyses of function and design in biology" edited by Colin Allen, Mark Bekoff and George Lauder[9]

[7]This position has been brought to the attention of the cognitive philosophy and ALife communities most strongly through the influence of Millikan[29], and her use of *proper functions* to solve problems of reference in thought and language. Note that Ron Chrisley points out that she does not use have a constitutive etiological claim - that functions *are* their history, just that history *explains* the occurrence of the function.

wings are so shaped that flapping them gives sufficient aerodynamic lift to let the bird go about its business. F is a function of T because T does F so well (in this context).

Neither of these definitions of function specifies the substance of an organ. It is, typically, not a function of an object to be made of glass, or carbon atoms. So biology has already allowed itself a substance independent functional language.

*The necessity of functionalism in biology*

There is a second reason for thinking that biology *needs* function. It uses such language to define *classes*. Since Darwin discovered that the species are not eternal, *natural kinds*[22] there has been some enquiry as to what the kinds of biology are.

Scientific laws are normally seen as ones which relate kinds. And kinds are seen as being universal, not tied to a specific location in space or moment in time. In contrast particulars are precisely located and extended in space and time. Before Darwinism, it was possible to think of, say, a horse as a natural kind. Anything which had the right number of legs, teeth, bones or whatever could be considered a horse. After a shift to an evolutionary perspective, it was more intuitive to think that anything that belonged to the horse lineage (that was related to other horses) was a horse; but anything that was even genetically identical to a horse, that had arisen from a different lineage that merely converged on the horse-style was not a horse.

Could there, then, be a science of biology or should there be a mere natural history which contents itself with collecting examples of the species and attempting to reconstruct their historical emergence? I contend that there can be a science, once traits are seen as instances of functional kinds.

The philosopher Karl Popper has argued strongly that historical reconstructions and predictions due to the extrapolation of trends are not scientific[35]. Science works on the basis of hypothesising the relations between universals, and it is this universality that allows further experiment. One can test, by observation of instances of kinds, whether the interaction fits the pattern that the laws relating the kinds predicts. On the other hand, if a prediction is based on no more than the extrapolation of a trend, then whether it is confirmed or not has no implication for a notion of a universal class. Nor can it be seen to warrant further predictions or hypotheses.

But when we take an *adaptationist* stance in biology; that is we do use function or purpose to define our kinds, a rich new possibility of true scientific research is opened up. We can not have laws that refer to hymenoptera but we could have laws that refer to eusocial animals; or to behaviours such as co-operation, breeding and feeding. And such laws have observational consequences.

Let us suppose that we hypothesise a behavioural trait T exists to exploit an opportunity in the environment. It is testable in three ways.

- Geology and paleontology can potentially falsify a claim that such a suitable environmental opportunity existed.
- Even if it did, paleontology and research into DNA history can falsify the claim that the trait appeared contemporaneously with the opportunity.

- Finally, modern experiments, and mechanical or behavioural tests can falsify the claim that the trait is sufficient to take advantage of the opportunity.

Now, this is not unique to functionally defined kinds. A trait could be defined as an example of a physical kind. Perhaps a particular trait could be defined as the outcome of a particular class of self-organising system, or basin of attraction in developmental space[8]. I accept that this leads to equally scientific predictions but deny that the understanding we gain is comparable. We have certainly not managed to eliminate all functional talk in biology in favour of such talk. Although it is not to be ruled out in the future, it is not current practice.

To summarise : biology uses functions. They feature prominently in explanations. They are also important to the definition of natural kinds, which in turn give biology its status as a science which can make falsifiable predictions. It seems only fair, therefore, to consider that since it is the multiple instantiability of functionally defined kinds that gives biology its scientific status; similar courtesy should be accorded to ALife.

### 10.5.4 Arguments against strong ALife

When the discussion of strong vs. weak Alife arises, typically the emphasis is put on the hard cases. For example, are virtual things *really alive*? And two further assumptions are made, one implicit, the other explicit.

The explicit assumption made by, among others, Elliot Sober[45] and Seth Bullock[6], is that because we do not have a full theory of life, we ought to err on the side of presuming that virtual things are not alive. The implicit assumption is that we need a grand overarching theories *before* it is possible to assume virtual things fall into *any* scientifically respectable categories.

I think many of one's intuitions can hinge on this implicit assumption, and this is what I take issue with. Bullock illustrates it in operation very well. He is always motivating criticisms of naive or superficial or inadequate "strong" assertions using the lack of a mature theory of life.

Consider this quote from Bullock : "Attempting to discover the nature of life (or some other biological phenomenon) through digital naturalism is analogous to attempting to discover the laws of aerodynamics which govern flight through drawing many different pictures of imaginary birds, imaginary flying insects, etc. ... Once many such pictures of 'flight-as-it-could-be' are rendered, one might spend years searching for principles which reveal the nature of flight. Without some theoretical framework, within which to locate this enterprise, some framework which takes existing theory of 'flight-as-we-know-it' as bedrock ... the chances of stumbling across some of the fundamental principles of aviation are remote in the extreme."

The problem is, that while he is (sort of) right his conclusion is unnecessarily pessimistic. One should note two things from the example. One is that, it is not making models which is the problem, or indeed any particular type of model. It is making models which don't grasp the right essence of the thing under investigation. Seth's researcher would be perfectly reasonable to study

[8]Two researchers who are interested in this approach are Brian Goodwin and Stuart Kauffman. See Griffiths[17] for a discussion and Kauffman[25]

flight by making *scale models* of imaginary birds. And equally, she would be perfectly reasonable to study *camouflage* by painting coloured pictures of imaginary birds. One does indeed need some existing theory of flight-as-we-know-it to recognise what the crucial property to model is.

But what I really feel like shouting at Bullock here is something like "Pah! Theoretical framework indeed. This is just a matter of common sense!"

In fact one doesn't need a great deal of theoretical framework in identifying the "right" property to be captured when studying flight. Any fool can *see* what that property is[9].

Of course, Bullock uses flight just to make a point, but wrapped up in it is a deeper assumption that the "right" theoretical framework is a holy grail, only accessible to the most scientifically pure of heart.

But in fact, many perfectly good examples don't require more theory than flight. And can be handled in a piecemeal way, rather than demanding a framework. One of the common targets of this kind of criticism is Tom Ray, who's Tierra model is often slated as an example of inappropriate "strong ALife" talk[10].

But in fact, it seems perfectly plausible to me that when Ray says he observes parasitism in Tierra, he does just that. Spotting the essential feature of parasitism-as-we-know-it is no harder than grasping flight-as-we-know-it, or co-operation-as-we-know-it. And no more in need of a more extensive theoretical framework.

Once this is accepted, the majority of ALife simulations with virtual beings can be discovered to be exemplifying some behavioural properties and are therefore members of behavioural classes upon which empirical observations may be made.

*But isn't this just wrong? Aren't such folkish, non-theoretical categories just unscientific?*

Personally, I don't believe that science is demarcated by being the study of rigorously or mathematically defined categories. Consider much of the animal behaviour literature purports to be the study of behavioural categories such as dominance, recognition etc. which, as we have seen, are still debated categories. Over time, we will undoubtedly come to sharper, better definitions of

[9] A witty reader challanges me at this point. "And what is it?," he asks. But I won't be drawn on this. The point of this statement, as I hope is explained in the next few of paragraphs, is not that flying is obvious or unproblematic to define. But that all behavioural categories are, to some extent, problematic and hard to define. And scientific respectability does not rest on starting with watertight definition of the behaviour. If it did, all the current animal behaviour literature would be equally problematic and "unscientific". Consider again that "co-operation" can be realized as blood sharing between vampire bats to predator inspection in sticklebacks. That we recognise the family resemblance is sufficient for us to start hypothesizing that results for one may be transferable to the other. On the other hand we may decide Hemelrijk's challange to the understanding of co-operation, that it has no fitness consequences, is a move too far. But the "too far" for modifying the theoretical framework will always itself be grounded in common sense, or blind conjecture, or something else external to the framework, not the framework itself.

[10] Bullock : "For example, the similarities between Ray's (1994) artificial ecological system, Tierra, and ... natural ecosystems ... may be merely superficial, whereas the differences may be telling ... without an established theory of life which specifies the grounds upon which comparisons between artificial and natural life may be made ... trivial resemblances ... might lead to bogus inferences from one class of system to the other." Sure, they might. The problem is that, there is no criteria for choosing good rather than bad properties. Bullock rightly disagrees with a position, once attributed to Chris Langton, that models of life could become so accurate that they would literally become examples of life. But he still seems to suppose an analogous fallacy : that having a large, detailed and accurate framework is the prerequisite to spotting good common properties, to grasp which is the essential property of flight to incorporate in one's models.

some of these categories.

For example the use of the prisoner's dilemma in studying co-operation helps us give a more rigorous definition of what co-operation is. And earlier I rejected some of Hemelrijk's findings on the grounds that her notion of co-operation didn't imply a fitness cost to the altruistic individual, and hence wasn't *real* co-operation. Maybe my approach, to define co-operation in terms of a fitness cost, will become standard and it will be the scientifically respectable notion. Or maybe anti-adaptationists will reject a fitness based definition of behaviour.

However that particular question turns out, the current lack of consensus on categories does *not* disqualify observations made using them from being *scientific*. If it did, most of biology would be equally unscientific. And theoretical biologists would be open to the same criticism as ALifers - that they played with theoretical models which were untestable.

One might argue that the categories in use in biology, while still debated, are nevertheless more mature, and hence closer to the ideal, than those in ALife. In some cases this may be true for some categories that are genuinely new in ALife. But more often the categories in ALife and biology are the same : *co-operation, aggression, extinction, predation, parasitism*.

### 10.5.5 Substance chauvinism

The second major source of criticisms of strong ALife is based on *Substance chauvinism*. A traditional view of life might go : "there is a substance of *life* which lurks inside living things". Those espousing the mainstream positions in biology and ALife have rejected this in favour of accepting that lifelikeness is a sort of organisational structure. This is of course very close to the functionalist, substance independent view, that I am trying to promote. But, the twist is, that the substance chauvinists hold that it must be a particular kind of organisation of a particular physical substance.

One example comes from Margaret Boden who argues that life is defined by several behavioural criteria including *metabolism*. But metabolism is to be defined as a certain kind of management of energy. And energy can not be defined functionally. Hence, while reproduction, evolution and other behaviours might all be realized by a computer program, metabolism can't because it requires the right kind of causal connections with energy.

A more general argument comes from H. H. Pattee[34]. Pattee first distinguishes precisely the categories we are interested in, namely *simulations* and *realizations*. Simulations are metaphoric representations of certain classes of entities; realizations are actual examples of them. He does this to correct a remark attributed to Langton that simulations could become so good that they become realizations. Pattee is clear that simulations and realizations are two entirely different sorts of things. While similarity is a virtue for simulations, no degree of it will turn a simulation into a realization. On the other hand, realizations of a functionally defined class such as life, must include the right kind of relations with such other concepts as evolution and strong emergence. He further believes that our theories of evolution and emergence are not yet good enough to tell us whether we have realizations of life. But, he thinks the chances are that they are not.

Once again I find myself much in agreement with this. But I suggest that artificial creatures

don't have to be really alive before they are really behaving or functioning. To make an analogy. It seems to me quite correct to say that *companies* are capable of certain behaviours. They can compete or co-operate in a cartel. Yet companies are not normally taken to be alive. Nevertheless they are one of the legitimate objects of enquiry of the science of economics.

I am sure I have not exhausted the arguments against strong ALife. But I hope to have made the case that it can not be trivially dismissed. And if so, the implications for ALife as science are profound. A strong Alife is an empirical investigation into the interactions of behaviours and functionally defined entities. In this it has exactly the same status as biology, psychology, artificial intelligence and perhaps the social sciences.

## 10.6 On future work

The prisoner's dilemma has been an invaluable schema to model a social environment. But to my mind, it has also been a great restriction on this work. The binary language of co-operation and defection is just too simple for us to read sophisticated stories about evolutionary dynamics. And the attempt to use Axelrod's niceness, provokability and forgiveness as behavioural decompositions of the population strategies has been largely unsuccessful. I would not recommend continuation with these categories or the straight prisoner's dilemma. It is certainly time to move on to a world with richer behavioural repertoire.

Having complained against the simplistic behaviour available in the prisoner's dilemma, it might seem strange that I have not experimented with a stochastic strategy such as those used by Nowak[32] which would place a probability of co-operating in each slot of the individual oriented strategy table. This is one way it might be possible to produce a probabilistic ramp towards full TFT.

The spatial version of this model is in very preliminary stages. More work needs to be done. Particularly to see what average level of IRec is supported on a spatial grid. Is it less than in the non-spatial world?

### 10.6.1 More biological modelling

More detail from the biological world could be brought in. It should be possible to incorporate some of the following :

- overlapping generations
- dominance hierarchies

*Overlapping generations*

Overlapping generations provide the possibility of enriching the model of kin oriented strategies to include parental care and alloparental helping. Alloparenting is the care of young by non-parents, sometimes non-relatives. Current explanations for helping in the care of young by older siblings includes straight kin oriented altruism, the idea that immature siblings are practising and refining their own parenting skills in harsh environments. Care by apparent non-relatives suggests

that some species rely on location to recognise in. (Help anyone who's in mum's nest, or in the colonial nursery.) This implies either that location is a very good indicator of kinship, or that the cost of false positives (over-helping unrelated young) is less than the cost of false negatives (under-helping related young) and so discrimination is of little importance.

My model includes a number-of-survivors-from-one-generation-to-the-next parameter with the hope that a little of this might be explored. (Remember the, perhaps unexpected fact that average parents / offspring similarity is not the same as full sibling similarity.) However, nothing of interest is discernible in the experiments so far. Further investigations along these lines would be made more interesting by having players that developed over their lifetime. So that perhaps strategy would change with age, and players would be given age recognition.

An experiment to test some of the assumptions about alloparental care could be implemented as follows. Allow strategy and some notion of competence of strategy to change with age. See if kin oriented altruism increases at ages with less competence. A variant would be allow amount or success breeding to change with age, or let players have a probability of dying each "year". See if kin oriented altruism increases at ages of less successful breeding, or in grandparents, who have already successfully produced a number of children.

*Dominance hierarchies*

As mentioned in chapter 1 dominance hierarchies are tricky things. Trivers's prediction was that hierarchies would diminish reciprocation, in the sense that a game freely entered into for personal benefit would be replaced by a game where the dominant individuals would have the power to take resources from subordinates. He predicted that the situation would be reversed in more sophisticated societies where agents needed to make political connections and seek support. We might study individual recognition in such a simulation of dominance signalling, negotiable access to resources and political intrigue.

# Bibliography


[1] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[2] Robert Axelrod. *The Evolution of Cooperation*, page 51. Basic Books, 1984.

[3] C. J. Barnard and Theodore Burk. Individuals as assessment units - reply to breed and bekoff. *Journal of Theoretical Biology*, 1981.

[4] Ken Binmore. Review of the complexity of co-operation. *The Journal of Artificial Societies and Social Simulation*, 1(88), 1997.

[5] Michael D. Breed and Mark Bekoff. Individual recognition and social relationships. *Journal of Theoretical Biology*, 88:589–593, 1981.

[6] Seth Bullock. *Phd Thesis*. PhD thesis, School of Cognitive and Computing Sciences, 1997.

[7] H. Burda. Individual recognition and incest avoidance in eusocial common mole-rats rather than reproductive suppression by parents. *Experientia*, 51:411–413, 1995.

[8] William H. Calvin. The ratchets of social evolution. In *The Throwing Madonna*, chapter 5. Bantam, 1983.

[9] George Lauder Colin Allen, Mark Bekoff, editor. *Nature's Purposes : Analyses of function and design in biology*. MIT Press.

[10] Philip H. Crowley, Louis Provencher, Sarah Sloane, Lee Alan Dugatkin, Bryan Spohn, Lock Rogers, and Michael Alfieri. Evolving cooperation: The role of individual recognition. *Biosystems*, (37):49–66, January 1996.

[11] Charles Darwin. *The Origin of Species*. John Murray, 1859.

[12] Carlos Drews. The concept and definition of dominance in animal behaviour. *Behaviour*, 125(3-4):283–313, 1993.

[13] Lee Alan Dugatkin and Michael Mesterton-Gibbons. Cooperation among unrelated individuals : reciprocal altruism, by-product mutualism and group selection in fishes. *BioSystems*, 1996.

[14] G. Goodall Gheusi and R. G. Dantzer. Individual distinctive odours represent individual conspecifics in rats. *Animal Behaviour*, 53:935–944, 1997.

[15] Christoph Goessman, Barbera Hemelrijk, and Robert Huber. The formation and maintainence of crayfish hierarchies. *Behav Ecol Sociobiology*, pages 418–428, 2000.

[16] S. J. Gould and R. C. Lewontin. The spandrels of san marco and the panglossian paradigm: a critique of the adaptionist programme. *Proc. of the Royal Society of London, ser. B*, (205):581–598, 1979.

[17] Paul E. Griffiths. Darwinism, process structuralism and natural kinds. In *Philosophy of Science, Proceedings*, number 63, pages S1–S9, 1996.

[18] Patrick Grim. The greater generosity of the spatialized prisoner's dilemma. *Journal of Theoretical Biology*, (173):353–359, 1995.

[19] W. B. Hamilton. The evolution of altruistic behaviour. *American Naturalist*, (97):354 – 356, 1963.

[20] Barbera Hemelrijk. Dominance interactions, spatial dynamics and emergent reciprocity in a virtual world. In P. Maes, M. J. Mataric, J-A Meyer, J Pollack, and S. W. Wilson, editors, *SAB 96 : From Animals to Animats*, volume 4, pages 545–552. The MIT Press, 1996.

[21] Barbera Hemelrijk. Co-operation without genes, games or cognition. In Phil Husbands and Inman Harvey, editors, *European Conference on Artificial Life 97*, pages 511–520. MIT Press, 1997.

[22] David L. Hull. *Historical Entities and Historical Narratives.*

[23] Pierre Jouventin, Thierry Aubin, and Thierry Lengagne. Finding a parent in a king penguin colony : the acoustic system of individual recognition. *Animal Behaviour*, 1999.

[24] Christa Karavanich and Jelle Atema. Individual recognition and memory in lobster dominance. *Animal Behaviour*, 1998.

[25] Stuart A. Kauffman. *The Origins of Order*. Oxford University Press, 1993.

[26] Marty L. Leonard, Andrew G. Horn, Cory R. Brown, and Nicole J. Fernandez. Parent-offspring recognition in tree swallows, tachycineta bicolor. *Animal Behaviour*, 1997.

[27] Kristian. Lindgren. Evolutionary phenomena in simple dynamics. In S. Rasmussen D. Farmer, C. Langton and C. Taylor, editors, *Artificial Life 2*, 1991.

[28] G. F. Miller. Artificial life as theoretical biology : How to do real science with computer simulation. Technical Report 378, School of Cognitive and Computing Sciences, Sussex University, 1995.

[29] Ruth Garrett Millikan. *Language, thought and other biological categories.* MIT, 1984.

[30] Milton K. Munitz. Introduction. In Milton K. Munitz, editor, *Identity and Individuation*, page iii. 1971.

[31] Jason Noble. The scientific status of artificial life. Technical report, School of Cognitive and Computing Sciences, Sussex University, 1997.

[32] M. Nowak. Stochastic strategies in the prisoner's dilemma. *Theoretical Population Biology*, 38:93–112, 1990.

[33] Ezekiel Di Paolo. A little more than kind and less than kin : the unwarranted use of kin selection in spatial models of communication. In D. Floreano, J-D. Nicoud, and F. Mondada, editors, *European Conference on Artificial Life*, pages 504–513, 1999.

[34] H. H. Pattee. Simulations, realizations, and theories of life. In *The Philosophy of Artificial Life*. Oxford, 1996.

[35] Karl Popper. *The Open Society and Its Enemies.* Routledge, 1954.

[36] William v O. Quine. Two dogmas of empiricism. *Philosophical Review*, 60(1):20–43, 1951.

[37] R. M. Sayfarth and D. L. Cheney. Grooming, alliances and reciprocal altruism in vervet monkeys. *Nature*, 1984'.

[38] Laela S. Sayigh, Peter L. Tyack, Randall S. Wells, Andrew R. Solow, and Michael D. Scott. Individual recognition in wild bottlenose dolphins : a field test using playback experiments. *Animal Behaviour*, 57(1):41–50, 1998.

[39] John Searle. Minds, brains and programs. *The Behavioural and Brain Sciences*, (3):417–424, 1980.

[40] Anil K. Seth. Interaction, uncertainty, and the evolution of complexity. In Phil Husbands and Inman Harvey, editors, *Fourth European Conference on Artificial Life*, pages 521–530. MIT Press, 1997.

[41] Paul W. Sherman, Eileen A. Lacey, Hudson K. Reeve, and Laurent Keller. The eusociality continuum. *Behavioural Ecology*, 6(1):102–108.

[42] Paul W. Sherman, Hudson K. Reeve, and David W. Pfennig. Recognition systems. In John Krebs and Nick B. Davies, editors, *Behavioural Ecology : An Evolutionary Approach*. Blackwell, 1997.

[43] Alexander F. Skutch. *Parent Birds and Their Young*, pages 318–320. University of Texas Press, 1976.

[44] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.

[45] Elliott Sober. Learning from functionalism. In M. Boden, editor, *The Philosophy of Artificial Life*. O.U.P., 1990.

[46] Robert L. Trivers. The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46:35–57, March 1971.

[47] G. S. Wilkinson. Reciprocal food sharing in the vampire bat. *Nature*, 308:181–184, 1984.

[48] G. C. Williams. *Adaption and Natural Selection*. Princeton University Press, 1966.

[49] David Sloan Wilson and Elliott Sober. Re-introducing group selection to the human behavioural sciences. *Behaviour and Brain Sciences*, 1994.

[50] E. O. Wilson. *Sociobiology (Abridged Edition)*. Belknap Press, Harvard University Press, 1980.