The Evolution of Animal Communication Systems: Questions of Function Examined through Simulation

Jason Noble

Submitted for the degree of D. Phil.

University of Sussex

November, 1998

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

Seth Bullock and Ezequiel Di Paolo have probably had the greatest immediate influence on the thesis, and so deserve to be acknowledged first: thanks, guys.

I remain grateful to Dave Cliff for taking me on as a student in the first place, back when I was full of wild-eyed ALife enthusiasm. Dave's devotion as a supervisor is rivalled only by his skill as a rock-climber. He is also very good company—cheers, Dave.

Thanks also to Mike Scaife, who kindly agreed to take up the slack when Dave moved on. Mike has been a level-headed and pragmatic presence during the frantic final year of my D.Phil.

The research community in the School of Cognitive and Computing Sciences at Sussex has been extremely supportive, and has gone some way towards restoring my faith in academia. There are too many people to mention individually, but some names stand out. A big thank you to Jo Brook, Julian Budd, Ron Chrisley, Stephen Eglen, Inman Harvey, Phil Jones, Adrian Thompson and Mike Wheeler, plus all the members of the ALERGIC, E-Intentionality, CogPhi and EvoComm discussion groups. I would also like to thank Linda Thompson and Magdalena Portmann, without whom nothing in COGS would work.

But it's not over yet. There's a large group of people, some of them at Sussex and some not, without whom I couldn't have done it. Their contributions to the thesis range from philosophy to beer to just being patient with me (sometimes all three at once), and I want to thank them all: Argiri Arfani, Guillaume Barreau, Tom "the Cincinnatti Kid" Beament, Hilan Bensusan, Vicki Carruthers, Sirem Erkek (my *en tatlı bebek*), Berkan Eskikaya, Sevan Ficici, Ronald Lemmen, Mar Muriana-López, Sean Noble, Manuel de Pinedo García ("Market Diner, beforé!"), Jorge Ramírez-Uresti, Oli Sharpe, Tim Taylor, and Richard Watson.

Many people at SAB 96, EvoLang II, ALife VI, and SAB 98 were also helpful. Thanks.

Thank you to John Maynard Smith and Peter Caryl for the advice and encouragement, and thanks to Bruce MacLennan for being such a good sport about chapter 5.

For their financial support I would like to thank the Association of Commonwealth Universities and the British Council, who were foolish enough to take me on as a Commonwealth Scholar. At this point I should probably also thank Mum and Dad, who gave a lot more than just cash. Finally, my undying gratitude goes to various Australian record companies who, through inaction, encouraged me to do a D.Phil. in the first place.

Preface

Parts of the thesis are based on work that has been previously presented or published. Chapter 4 borrows heavily from a paper presented at the Fourth European Conference on Artificial Life (Noble, 1997). Chapter 5 is based on a paper published in the Proceedings of the Fourth Conference on the Simulation of Adaptive Behavior (Noble & Cliff, 1996)—Dave Cliff was a co-author of this paper but all of the work actually presented in the chapter is my own. Chapter 7 is based on a paper published in the Proceedings of the Sixth Conference on Artificial Life (Noble, 1998b) and presented at the Second International Conference on the Evolution of Language (Noble, 1998a); an extended version of the work is currently in press in the journal *Adaptive Behavior*. Chapter 8 is based on a paper published in the Proceedings of the Fifth Conference on the Simulation of Adaptive Behavior (Noble, 1998c).

The Evolution of Animal Communication Systems: Questions of Function Examined through Simulation

Jason Noble

Summary

Simulated evolution is used as a tool for investigating the selective pressures that have influenced the design of animal signalling systems. The biological literature on communication is first reviewed: central concepts such as the handicap principle and the view of signalling as manipulation are discussed. The equation of "biological function" with "adaptive value" is then defended, along with a workable definition of communication. Evolutionary simulation models are advocated as a way of testing the coherence of a given theory. *Contra* some ALife enthusiasts, simulations are not alternate worlds worthy of independent study; in fact they fit naturally into a Quinean picture of scientific knowledge as a web of modifiable propositions. Existing simulation work on the evolution of communication is reviewed: much of it consists of simple proofs of concept that fail to make connections with existing theory. A particular model (MacLennan & Burghardt, 1994) of the evolution of referential communication in a co-operative context is replicated and critiqued in detail.

Evolutionary simulations are then presented that cover a range of ecological scenarios; the first is a general model of food- and alarm-calling. In such situations signallers and receivers can have common or conflicting interests; the model allows us to test the idea that a conflict of interests will lead to an arms race of ever more costly signals, whereas common interests will result in signals that are as cheap as possible. The second model is concerned with communication during aggressive interactions. Many animals use signals to settle contests, thus avoiding the costs associated with fighting. Conventional game-theoretic results suggest that the signalling of aggression or of strength will not be evolutionarily stable unless it is physically unfakeable, but some recent models imply that cost-free, arbitrary signals can be reliable indicators of both intent and ability. The simulation, which features continuous-time perception of the opponent's strategy, is an attempt to settle the question. The third model deals with sexual signalling, i.e., elaborate displays that are designed to persuade members of the opposite sex to mate. The results clarify the question of whether such displays are the pointless result of runaway sexual selection, or whether they function as honest and costly indicators of genetic quality.

The models predict the evolution of reliable communication in a surprisingly narrow range of circumstances; a serious gap remains between these predictions and the ethological data. Future directions for simulation work are discussed.

Submitted for the degree of D. Phil.

University of Sussex

November, 1998

Contents

List of Figures

List of Tables

Chapter 1

Introduction

While going about their business of survival and reproduction, most animals influence, and are influenced by, the behaviour of other animals. In some cases we describe this as communication. For example, vervet monkeys *Cercopithecus aethiops* use vocal alarm signals to warn their fellow troop members of the approach of three or four distinct kinds of predators. Mantis shrimps *Gonodactylus bredini* wave their claws to threaten other mantis shrimps in disputes over territory. Male Túngara frogs *Physalaemus pustulosus* produce distinctive sounds in order to advertise their size and attract females. Why do animals exhibit these sorts of behaviours? Such communication systems, i.e., dispositions to produce and respond to certain signals, did not spring up overnight: they evolved by a process of natural selection. If we want to know why a particular communication system exists, we have to examine the pressures that are likely to have shaped its evolution, and to determine the adaptive purpose that it serves. For instance, why do mantis shrimps produce threat signals, instead of attacking immediately or doing nothing—what is the selective advantage of the behaviour? Why do threatened shrimps pay any attention to the threat? What would happen to a mutant shrimp that behaved differently to the established norm? The aim of this thesis is to use a particular style of computer simulation as a tool with which to cast light on such questions.

1.1 Ways of explaining animal behaviour

The thesis will be concerned, then, with questions of adaptive purpose or function. To describe the function of something is one way of explaining it: one might explain a can-opener to a curious child by saying that its function was to open cans. However, functional explanations are by no means the only explanations that can be offered of animal behaviour. Tinbergen (1963) suggested that there were four independent questions that could be asked about any particular feature of an animal's morphology or behaviour. Tinbergen's four questions concerned mechanism, ontogeny, phylogeny and function. The reader may ask, with some justification, why only one of those questions will be addressed here.

The question of mechanism means asking how the behaviour or trait is physically realized; in other words, how does it work? A mechanistic account of, for example, vervet monkey warning signals would involve describing the precise way in which the nervous system of the signalling vervet processes the incoming visual information, recognizes a certain patch of yellow as an approaching leopard, stimulates muscles in the vocal tract so as to make the appropriate sound, etc. The flight behaviour of the vervets that hear the warning would have to be explained in a similar way. Providing a *complete* mechanistic account of vervet communication would require a marriage of ethology and neuroscience. Extreme technical difficulties in recording the activity of neurons, nerve fibres and muscles in living, unrestrained animals would have to be overcome. There could be only limited help from first principles (i.e., analysis of the computational requirements of the task) because the various elements of the system, such as leopard recognition, are likely to be multiply realizable. Mechanistic explanations will not be considered here because of the nature of the chosen research tool: the thesis is essentially a theoretical one and does not involve the sort of detailed empirical work necessary to provide a mechanistic explanation for any given communication system. Whilst it is conceivable that a sufficiently detailed, physically realistic computer model could be used to test well-defined mechanistic theories, the computer simulations employed in the current project are, on the contrary, quite simple and abstract.

The question of ontogeny is about the way a trait or behaviour develops over the lifetime of the animal. An ontogenetic account of the vervet monkey's communication system requires a description of such morphogenetic details as the way the vocal tract develops in the embryo and the growth of the relevant neural modules. An ontogenetic explanation would also describe the way in which young vervets learn to adjust their use of the warning signals in the light of feedback concerning their early efforts.[1] As with mechanistic explanations, the practical problems of investigating ontogeny will be left to field biologists and neuroscientists.

The question of phylogeny concerns evolutionary history. A phylogenetic explanation of vervet monkey alarm signals would involve specifying the evolutionary predecessors of the current behaviour patterns: for instance, perhaps ancestral vervet monkeys gave an undifferentiated alarm call when confronted with any dangerous situation, and then later came to differentiate by giving increasingly specific calls in response to each of the types of predators commonly encountered. Or perhaps the original alarm call was given only in response to leopards, and the other calls evolved separately and independently. However, evolution is famously difficult to observe in action, and investigating phylogeny necessarily has the flavour of detective work. Phylogenetic questions have traditionally been the domain of palaeontologists, but in the case of behavioural traits such as communication (as opposed to physical traits) the problems of determining phylogeny are particularly acute: after all, patterns of behaviour do not fossilize. In these cases the comparative method—in which the behaviour and, sometimes, the actual genomes of related species are compared—is the only way to test theories of phylogeny. Returning to the vervet monkey example, if we were to examine several closely related species and find that in each case only a generalized alarm call existed, this would lend a degree of support to the theory that the current differentiated calls evolved from a single ancestral call.

The complexities of phylogenetic investigation will not be dealt with here. However, whereas

[1] This is not meant to appear as pre-judgement of an empirical question: it is, of course, possible in principle that there is no learning component to the vervet monkey's communication system and that it is entirely innate, as most animal communication systems appear to be. However, observational data indicate that there is indeed a learning phase in which young monkeys receive feedback from adults: adults will not join in the alarm calling, nor will they perform predator avoidance behaviour, if a call is given in response to a non-predatory species (Cheney & Seyfarth, 1990; Caro & Hauser, 1992).

the thesis must be completely silent on mechanism and ontogeny, there is in fact some crossover between the question of function, with which we will be concerned, and the question of phylogeny. This is because a phylogenetic theory cannot propose evolutionary predecessors for a behaviour without regard for the functional value of those predecessors. For instance, whatever the comparative or palaeontological evidence, we would be suspicious of a theory that suggested that vervet monkey alarm calls were derived from earlier calls that were designed to *attract* predators. Intermediate forms in the evolution of any modern trait must themselves have had adaptive value, and thus analyses in terms of function may help to restrict the space of plausible phylogenetic hypotheses.

Tinbergen's question about function (originally presented as the third in the list) refers to the selective advantage that a trait has for the animal; the question has also been described as one of "survival value", "adaptive value" or of "ultimate function". The latter term hints at the appeal of this style of explanation: to give the function of an evolved behaviour is to answer the "why?" question. It is the goal of the current project to say something about why animals communicate in certain ecological contexts and not in others. Of course, in the broad scientific effort to understand animal behaviour, concentrating on function to the exclusion of the other three questions would be folly. Knowing why a behaviour persists tells us almost nothing about how it is physically achieved, or what its antecedents were. However, as Grafen (1990b) notes, "we can understand something without understanding everything", and an emphasis on the question of function is not new: the field of behavioural ecology (Krebs & Davies, 1981, 1997) can be defined by its reliance on this approach.

1.2 Theoretical models and functional explanations of animal communication

In biology, as we shall see in chapter 2, verbal arguments and mathematical models of evolution have been prevalent in discussions of the function of communication behaviour. For example, Dawkins and Krebs (1978; Krebs and Dawkins, 1984) have argued that communication is best seen as the manipulation of one animal by another; the function of signalling behaviour is thus to cause the receiver(s) of the signal to do something of benefit to the signaller. In contrast, Zahavi (1975, 1977, 1987, 1991) has repeatedly insisted that communication systems are kept honest because of the cost of the signals produced. Zahavi refers to this as "the handicap principle", and it implies that the function of a signal is to provide—due to its being costly to produce—an accurate index of some quantity of interest to the receiver. Models of both processes have been put forward, e.g., Noble (1998b) and Grafen (1990a) respectively. The two positions are, at least on the face of it, incompatible, but there is as yet no consensus in the literature as to which account best describes reality.

Readers with a taste for empirical evidence may immediately wonder what good could come from the continuation of this sort of abstract theoretical debate. Surely some experiment or observation could settle the question? The answer must be that the observation of communication behaviour in modern animals is often insufficient to decide between conflicting theories concerning the evolved function of that behaviour. Whilst it is true that careful empirical explication of the mechanisms underlying a particular communication system can sometimes make the function of that system clearer to us—von Frisch's (1967) work on the waggle dance in honeybees (*Apis mel-*

lifera) is an excellent example—determining the selective advantage of signalling and response behaviours is often not so simple.

Returning again to the alarm signals of vervet monkeys, it might seem obvious that the function of the communication system is to help the monkeys avoid predation. At one time, when the theory of group selection was held in higher regard than it is today, this explanation would have been accepted. However, the orthodox position in evolutionary biology (Williams, 1966; Dawkins, 1976; Maynard Smith, 1993) now says that animals are best understood as products of their selfish genes: animals do not do things for the good of the species, but in order to propagate copies of their genetic material. From this viewpoint, avoiding predation is likely only to be the function of the *response* behaviour. The function of the signalling behaviour is not so obvious: why should a monkey that has spotted an approaching leopard warn its conspecifics? Giving the alarm signal may well increase the risk to the signaller, by drawing the leopard's attention to itself.

The point here is not that the function of the alarm signal now becomes completely mysterious: several theories, such as kin selection and reciprocal altruism, provide candidate explanations. The point is that the switch from group-selectionist to selfish-gene thought was not inspired by empirical observations, but by theoretical arguments about how evolution really works. While each theory or viewpoint implies very different things about the evolutionary process, and about the evolutionary history of particular species, the two theories would not necessarily make different predictions about the likely alarm-signal behaviour of modern vervet monkeys. This rather extreme example of the under-determination of theory by data provides a paradigm case for the sort of argument that will be made in this thesis: when the data we have access to (i.e., ethological observations of modern animals) are incomplete, then choices between competing theoretical models must sometimes be made on the basis of non-empirical criteria, such as the internal consistency of the model, and the fit between the model and other, well-established theories. In our current theoretical understanding of animal communication, there are problems like the one above, where more than one model can account for what is observed, and there are also problems where a model can account for one phenomenon only at the cost of making something else inexplicable. The issue is surely complicated by the fact that communication is a co-evolutionary process, in which the advantage of using any one communication strategy will depend on the current distribution of strategies in the population. It is hoped that this thesis will contribute to the process of selection between, and improvement of, the existing theoretical models.

1.3 The use of computer simulations

The simulations presented will involve the explicit modelling of individual organisms interacting in a shared environment (although both the organisms and the environment will be very simple). Artificial evolution will be incorporated: organisms that are more successful, where success is defined by a criterion analogous to energy accumulation, will have a greater likelihood of passing on their genetic material to the next generation. Variation will be introduced through mutation, i.e., the occasional random alteration of the transmitted genetic information.

Conditions that influence the evolution of communication between organisms can thus be explored. If a mutation leads to proto-communicative behaviour in a newly-generated individual, and the behaviour is beneficial, then typically that individual will be selected to reproduce and the behaviour will be perpetuated. Possibly the communication system will become more complex through the accumulation of further mutations. The simulation method permits exploration of the conditions for such developments. If we envisage evolved signalling systems as points in a space of possible strategies, a simulation might help us to conclude—for example—that there will be no evolutionary path to communication if the organisms' sensory systems are thus rather than so, or if the benefits of dishonesty are above a certain threshold, etc.

Through varying the initial conditions (i.e., the genetic makeup of the organisms in the first generation), we can look at the conditions for both the stability and the emergence or origin of communication. Conditions for stability are examined by starting with a population that *already* communicates. Conditions for the emergence of communication, on the other hand, are examined by starting with a population in which there is no signalling behaviour—this logic is based on the evolutionary axiom that any modern communicating organism must have had a non-communicating ancestor.

Theories about the function of communication in a particular ecological context are easily tested: the simulation is constructed such that the hypothesized selective advantages for signalling and for responding are in fact available to the simulated organisms. If communication evolves under these conditions, then the theory is supported; if not, our confidence in the theory is reduced. For example, we could test the idea that vervet monkeys give alarm calls for the benefit of other troop members, despite there being a fitness cost to the signaller in doing so.

The simulation method is inspired by recent work in the area known variously as artificial life or the simulation of adaptive behaviour—for introductory reviews, see Langton (1989) and Meyer (1994) respectively. Simulation is not new as a research tool in biology, but conventional biological simulations tend to model whole populations, abstracting away from the individual organism, and they tend to be extensions of simple game-theoretic models, thus incorporating radical simplifying assumptions such as random mating and the absence of a spatial distribution. Communication is self-evidently about interactions between individuals mediated by an environment, and it is hoped that artificial-life-inspired methods will better reflect these key aspects of the phenomenon.

It is worth stressing that the simulation results will not be presented as a substitute for empirical evidence. If a simulation establishes the plausibility of a particular hypothesis about function, that is not the same thing as establishing its truth. The claim here is only that simulation methods can demonstrate the logical coherence (or indeed incoherence) of a particular model, and that they may suggest new hypotheses for empirical investigation: the shift from group-selectionism to selfish-gene thinking mentioned above illustrates that it is only when we have a particular theoretical picture in mind that we know whether to view an empirical observation as surprising or as unproblematic. These issues are explored at much greater length in chapter 4.

1.4 Problems in defining communication

Animals influence each other's behaviour in many different ways. Pinning down exactly which kinds of influence that we wish to call communication can be a troublesome business. It is uncontroversial to say that vervet monkeys are communicating when one gives a leopard alarm and the others scramble for the safety of the trees. But is a camouflaged insect signalling to its predators? By running away, is an antelope signalling to a cheetah? In both cases the answer is yes under certain definitions of communication that have been adopted in the biological literature. Intuitions differ about how such borderline cases should be treated; mimicry and deception are two other notable problem areas.

This definitional problem is despite the fact that in ordinary language we have a clear idea of what we mean by communication, or at least an archetypal image: a sender imparts information to a receiver via some sort of signalling channel, e.g., one person says truthfully to another, “It’s raining outside.” This has been dubbed the conduit metaphor (Reddy, 1979; Lakoff & Johnson, 1980).

The logic of the thesis requires measuring the existence and degree of communication in the evolutionary simulations that will be presented. In order to do so, it will be necessary to take sides in the debate about just what constitutes communication; a definition will be defended in chapter 3. The discussion will not be pre-empted here, but the definition argued for is based on the work of Millikan (1984, 1993), and turns out to be not far removed from the imagery of the conduit metaphor. Alternative definitions of communication—notably those phrased in terms of behavioural influence, information transmission, or the intent to communicate—will be considered and rejected.

1.5 Human language

One of the reasons we study the evolution of animal signalling systems is the suspicion that they may have something to tell us about the origins of human language. The way in which vervet monkeys use apparently arbitrary sounds to denote different kinds of predators, for instance, is certainly reminiscent of the arbitrary connection between words and their referents. Could language have had its beginnings in something like an alarm call system amongst our savannah-dwelling primate ancestors? Possibly, but caution is called for: speculating on the origin of language has a long and disreputable history.[2] Whilst it is more than reasonable to suggest that *Homo sapiens* is descended from creatures that did not have language but possessed only relatively primitive signalling systems, it would be a mistake to thereby suppose that human language is *merely* another—albeit complex—animal signalling system. Human language is very different from all other forms of animal communication, and our very familiarity with our own linguistic abilities, combined with a laudable desire not to be species-chauvinist, can sometimes lead us to underestimate the distance between ourselves and the rest of the animal kingdom.

One major difference is that human language has recursive syntax: complex meanings can be built up by combining smaller units. Animal communication, on the other hand, either consists *only* of atomic meaning-units, as in an alarm call system, or has some “syntactic” structure but does not appear to utilize that structure for the purpose of constructing complex meanings, e.g., bird or whale song. Syntax makes human language an extremely expressive system, and therefore allows us to do such complicated things as talking about past and future events, talking about conditional relationships, combining arbitrary actions, properties and objects, and producing and understanding sentences that we have never heard before. There is no evidence for anything like

[2] Bickerton (1994) informs us that as long ago as 1866, the members of the Linguistic Society of Paris were so tired of wildly speculative papers on the origin of human language that they imposed a ban which apparently stands to this day. It is not clear what they would have thought of Bickerton’s own speculations on the topic.

this kind of sophistication in the communication behaviour of any other animal species (Hauser, 1996). Unsurprisingly, discussions of the evolution of language are often, at heart, discussions of the evolution of syntax.

A second difference is that human language is overwhelmingly a learned system, whereas most animal communication systems are innate. Of course, the human ability to process and produce syntactically structured utterances, i.e., to use grammar, does seem to be innate, as Chomsky (1957, 1968, 1975) has argued. However, we flesh out this capability by rapidly learning many thousands of words in one or more natural languages, a feat that no other animal has achieved. Oliphant (1997) even suggests that it is this ability to learn a large vocabulary by imitation, rather than the ability to use syntax, that stands as the "cognitive bottleneck" that explains why no other animals have comparable communication systems.

Finally, human language is problematic because some of its basic features may not have had communication as their original function. Chomsky argued that our linguistic talents derive from the presence in our brains of a "language organ" that at one time allowed us to carry out combinatorial calculations in the service of some other activity, and which has since been co-opted for the purpose of communication. Other authors (Dennett, 1991a; Bickerton, 1994) have also suggested that language originated in the re-application of some older ability. If any of them are correct in their suspicions then it would be a mistake to conclude that language has been shaped by the same selective pressures as simpler animal communication systems.

It is therefore safe to say, at the very least, that human language is an exceptional case of animal communication. The thesis will not take up a position in the debate on whether human language is "continuous" with animal communication—this is left as a matter for others to argue over. Instead, a case will be made that coming up with functional explanations for the simpler forms of animal communication is not as unproblematic as some syntax-oriented theorists might suppose. The thesis will not deal with the complexities of syntax nor with learned communication; it will be seen in chapter 2 that this nevertheless leaves significant problems to be explored, mostly around the issue of how reliability is maintained in a signalling system. The question of language is touched upon briefly, however, in relation to the concepts of communication, information and intentionality in chapter 3. The temptation to indulge in idle speculation about language origins will be resisted as strongly as possible.

1.6 Overview

1.6.1 Outline of the thesis

The simulation models presented in later chapters do not represent *ad hoc* ideas on the function of communication, but rather test and extend existing theories. Therefore in chapter 2 the biological literature on communication is reviewed, with specific reference to theories of function. Central concepts such as Zahavi's handicap principle, and Krebs and Dawkins's view of signalling as manipulation, are discussed. The modelling techniques used in theoretical biology, e.g., game theory and population genetics, and the implied theoretical stance, in which evolution is seen as an optimization process, are also considered.

Chapter 3 deals with some conceptual problems: the thesis concerns the evolved function of communication behaviours, but in order to discuss this coherently we need to take a position on the notion of function in biology, and to settle on a workable definition of communication. In both instances the work of Millikan (1984, 1993) proves useful. Chapter 3 also covers the concepts of information and intentionality, both relevant to the problem of defining communication.

Chapter 4 presents arguments justifying the use of computer simulation as a tool for investigating evolutionary phenomena. Evolutionary simulation models are advocated—within the framework of a Quinean view of science—as a way of testing the coherence and consistency of a given evolutionary theory. Simulation models are shown to be, approximately speaking, analytic tools and not alternate worlds worthy of study in and of themselves, as some of the recent work in artificial life would have them. The relationship between computer simulation models and the older tradition of mathematical modelling in biology is explicated, and connections are drawn between simulated evolution and the notion of biological function in Millikan's work.

Chapter 5 reviews existing work in the artificial life and simulation of adaptive behaviour literature in which computer simulations are used to model the evolution of communication. Given the perspective taken in chapter 4, it is argued that much of this work consists of isolated proofs of concept, and could be improved upon by closer attention to links with existing theory. In order to illustrate the point, a simulation model of the evolution of communication by MacLennan and Burghardt (1994) is replicated and critiqued in detail.

Chapter 6 is a pause in the argument. Many of the theories reviewed in chapter 2 are worthy of investigation through the construction of evolutionary simulation models, but space and time preclude doing this exhaustively. This chapter argues for the choice of problems and hypotheses in the subsequent modelling chapters; the choices made are of necessity somewhat arbitrary, but an attempt is made to cover a range of ecological contexts. There is an emphasis on situations where the function of a signalling system is controversial. Simulations allow the investigation of many factors that are closed to, or difficult to capture with, traditional mathematical treatments, e.g., the effects of distributing the population in space: chapter 6 also justifies the selection of certain of these factors for closer attention.

Chapters 7, 8 and 9 consist of original simulation work. The evolved priorities of animals have famously been summarized as "the four Fs": feeding, fighting, fleeing, and reproduction. The evolutionary simulations presented will model animal communication systems relevant to all four of these categories.

Chapter 7 deals with communication about feeding and fleeing, i.e., food and alarm calls. With respect to these calls, it is clear that signallers and receivers can have common or conflicting interests. Being informed of the presence of food or predators is usually beneficial, but it may or may not pay to so inform one's conspecifics. This provides convenient grounds for testing an aspect of Krebs and Dawkins's (1984) theory: the idea that a conflict of interests will lead to an "arms race" of ever more costly signals and ever more sceptical reception strategies, whereas common interests will result in signals that are as cheap as possible while still being detectable.

Chapter 8 is concerned with communication during aggressive interactions, e.g., aggressive posturing, threats, bluffs, and signals by which animals assess each other's strength. Fighting is energetically expensive and carries a risk of injury or death; it is intuitively plausible that animals might evolve to use signals to settle fights and thus avoid these costs. However, conventional game-theoretic results suggest that the signalling of aggressive intent or of fighting ability will not be evolutionarily stable, unless the signal is unfakeable in some way. In contrast, a few recent models imply that cost-free, arbitrary signals can be reliable indicators of fighting ability. A simulation model, featuring continuous-time decision making and perception of the opponent's "intention movements", is constructed in an attempt to settle the question.

Chapter 9 deals with sexual signalling, the elaborate displays or signals produced by one sex (typically the male) in order to persuade members of the opposite sex to mate. The basic controversy in the literature is over whether these displays are the "pointless" result of runaway sexual selection (Fisher, 1930), or whether they function as costly-and-therefore-honest indicators of genetic quality (Zahavi, 1975). While mathematical models have indicated that both processes are plausible, there is disagreement as to their relative importance. Models of the costly-indicator theory have almost all failed to incorporate heritable variation in male quality. A simulation is presented in which that defect is corrected, and in which both mechanisms can be evaluated in a common framework.

Chapter 10 presents the conclusions and limitations of the thesis. Overall, the simulation and game-theoretic models presented predict the evolution of stable communication in a surprisingly narrow range of circumstances. It would appear that there is a serious gap between these predictions and the ethological data—possible explanations for the discrepancy are offered. Promising directions for future simulation work are also discussed.

1.6.2 Original contributions

The main contributions of the thesis can be placed on a continuum between philosophy and theoretical biology. In the former category there is a an application of Millikan's (1984, 1993) ideas in order to construct a defensible definition of communication. The thesis also presents an argument for the use of evolutionary simulation models as a research tool—in biology and possibly in other disciplines—that supplements traditional methods such as game theory and population genetics. This builds on a program of research first suggested by Miller (1995).

Moving towards the scientific end of the spectrum, there is a critical review of recent work on the evolution of communication within the new field of artificial life. This includes the replication of a seminal model by MacLennan and Burghardt (1994).

The core contribution of the thesis consists of three original evolutionary simulation models. However, the purpose of these models cannot be briefly summarized or distilled into a memorable slogan: each of the chapters detailing a model (chapters 7, 8 and 9) explores specific technical points from the theoretical-biology literature. Krebs and Dawkins's (1984) conspiratorial whispers theory is modelled in the context of food and alarm calls. Krebs and Dawkins's ideas have been of some influence in biological thinking on communication, but few explicit models of their theories have been constructed. Enquist's (1985) and Hurd's (1997b) controversial views on signalling during contests are opposed to more traditional game-theoretic views (Maynard Smith, 1982)—evolutionary simulation methods have not previously been applied to this issue. Finally, a simulation of sexual signalling is developed. Iwasa, Pomiankowski, and Nee's (1991) model of handicap signalling of heritable quality is tested with some of its restrictive assumptions relaxed; this has been highlighted by Andersson (1994), in a major review of the sexual-selection literature, as work that needed to be done.

Chapter 2

The biological literature

The purpose of this chapter is to review theories from biology concerning the function of animal communication, i.e., ideas on why animals have evolved to communicate.[1] The review proceeds in a roughly historical fashion, starting with some of Darwin's thoughts on animal signals. The mathematical modelling techniques that have often been used to justify theories of function, and certain relevant aspects of the history of ideas in evolutionary biology, are also discussed.

The biological literature includes a wealth of painstaking empirical studies of the communication systems of different species (see Hauser, 1996, for a comprehensive review of work on auditory and visual communication systems). The work of von Frisch (1967) on the dance "language" of honeybees has already been mentioned; other notable examples include Tinbergen's (1953) descriptions of sexual and aggressive signalling behaviour in herring gulls *Larus argentatus*, Cheney, Seyfarth and Marler's (1980; Cheney & Seyfarth, 1982, 1990) work on the alarm calls of vervet monkeys and other primates, and Møller's (1988, 1989, 1991) ingenious experiments on sexual advertisement signals in passerine birds.[2] However, the goal of the thesis is to construct simple models that capture general principles in the evolution of communication and thus help to extend, refine or refute existing theory. The goal is *not* to build detailed computer simulations of communication systems in particular species (although this would also be a worthwhile project). Therefore empirical studies will only be covered here inasmuch as the authors contribute to theories concerning the selective advantage of communication.

2.1 Darwin on communication

Questions about the selective advantage of communication would have been almost unintelligible prior to the publication of Darwin's *The Origin of Species* (1859), in which he presented his theory of natural selection. Darwin saw that his theory applied not just to physical traits like wings or weaponry but also to behavioural traits such as social activity and communication. Animals could be expected to vary: for example, in wing length, or in their tendency to vocalize. Any variation

[1] The structure of this chapter owes much to Hauser (1996), in particular to his discussion of the history of biological thought regarding the evolution of communication. The influence of review articles by Harper (1991) and Johnstone (1997) must also be acknowledged.

[2] Passerine birds are those of the order *Passeriformes*, which includes perching song-birds such as the sparrow.

that gave a fitness advantage—that is, resulted in the animal leaving a relatively larger number of offspring—would be preserved. The animal's progeny would tend to inherit the variant trait, thus inheriting the fitness advantage, and over generational time the proportion of the population exhibiting the variation would increase.

2.1.1 Signalling of emotional state

Darwin was a keen naturalist, and he observed on many occasions that animals appeared to communicate with each other. According to the logic of his theory, he believed that this communication behaviour must have some selective advantage, i.e., it must have a function. Darwin believed that one important function of communication was the accurate transmission of information about some aspect of an animal's internal state. He developed a theory of "expression" (Darwin, 1872) in which he argued that animals, including man, had undergone selection for the unambiguous communication of their emotional state. A central element of the theory was the principle of antithesis, which maintained that pairs of signals indicating opposing emotional states were likely to be of opposite physical form. For example, dominant or aggressive animals produce low-pitched vocalizations and attempt to make themselves appear larger (through hair-bristling, etc.), whereas submissive or fearful animals produce high-pitched sounds and attempt to make themselves look smaller. Darwin expected this kind of relationship to be universal, because there is a necessary connection between—in the current example—large size, dominance, and low-pitched vocalizations.

However, the principle of antithesis is really a theory of signal form, and not one of signal function. The implication in Darwin's work is that there is some inherent selective advantage in the honest transmission of internal states. So presumably we can expect an animal that expresses its fear and submission to be more successful than one that does not, perhaps because the former will escape attacks from dominant animals while the latter will be perceived as antagonistic and be punished accordingly. Similarly, a male that expresses his confident, dominant state should be fitter than one that does not, perhaps because he will receive uncontested access to a group of females whereas the latter male will have to fight for such access. Effectively, Darwin asks us to believe that honesty is the best policy; in section 2.3 we will meet a challenge to this view.

2.1.2 Sexual advertisement signalling

Darwin also developed the theory of sexual selection (1871). Sexual selection is a distinct subset of natural selection; the idea is that evolution is an exam with two papers: in order to have offspring an animal must not only survive to adulthood, but, in a sexual species, it must gain mating opportunities with members of the opposite sex. This latter kind of selection, whereby fitness advantages accrue to those animals most attractive to the opposite sex, is termed sexual selection. Darwin's insight was that sexual and natural selection could potentially exert opposing evolutionary pressures. If, for some reason, females came to prefer males with elaborate and costly ornaments, such as the peacock's tail, then sexual selection would push towards yet more costly ornaments, because males with longer tails experience greater mating success. At the same time, natural selection would push for less costly ones, because males with longer tails are more vulnerable to predation and less likely to survive to adulthood.

Darwin believed that sexual selection gave rise to another kind of communication, in which one sex produced signals (ornaments, songs, dances, etc.) that had the function of attracting the other—he observed that these signals were typically produced by males to attract females. Darwin (1871, p. 56) went as far as suggesting that human language had originated in this kind of sexual advertisement signalling:

> When we treat of sexual selection we shall see that primeval man, or rather some early progenitor of man, probably used his voice largely, as does one of the gibbon-apes of the present day, in producing true musical cadences, that is in singing; we may conclude... that this power would have been especially exerted during the courtship of the sexes... The imitation by articulate sounds of musical cries might have given rise to words expressive of various complex emotions.

As has been pointed out many times since, the main weakness in Darwin's theory of sexual selection was that he provided no explanation for the female preferences that male sexual signals were designed to exploit. In primeval humans, why should the females have found male singing attractive? Darwin apparently thought of these preferences as something akin to an innate aesthetic sense, and offered no theory as to their function. In sections 2.4 and 2.7 we will review more recent ideas on sexual selection.

2.2 The ethological view

Ethology came to prominence as a discipline in the 1940s and 50s, and defined itself in opposition to comparative psychology. Whereas the comparative psychologists studied animals in controlled laboratory settings and were working towards a general theory of learning, the ethologists believed that animal behaviour could only be studied by observing animals in their natural environment. Tinbergen, Lorenz and von Frisch are generally considered the founders of ethology.

The ethologists followed Darwin in assuming that the function of most animal communication systems was to accurately convey information about internal states. However, they went into more detail concerning the evolutionary origin of signalling systems. A clarification is necessary here: the origin of any particular signalling system is properly a question of phylogeny, as discussed in section 1.1. But general theories of how communication systems get started *are* relevant to the question of function. Recall that function has been equated with selective advantage (a move justified in chapter 3); in order to understand the selective advantages inherent in a communication scheme, we must consider what the earliest signallers and receivers stand to gain against a background of no communication.

The two key concepts in the ethological picture of the evolution of communication (Tinbergen, 1952, 1964) are "derived activities"—non-signals which provide the raw materials for signal evolution—and the subsequent "ritualization" of the nascent signal. Tinbergen credits Selous (1901, 1933) and Huxley (1923) with these notions.

2.2.1 Derived activities

Derived activities are actions or cues that are associated with a specific internal state and are thus predictive of an animal's future behaviour. For example, a male monkey might place one hand on the female's head to ensure his balance before copulation—head-touching would thus count as a derived activity that was predictive of an attempt to copulate. The ethologists suggested that derived activities are the precursors of signals.

Tinbergen (1964) also uses the phrase “derived movements”, but this is a historical accident. Tinbergen presented a typology of signals based on form: movements and postures, brightly coloured structures, scent signals, sounds, and tactile signals. He focused on movements because, at the time that he was writing, they were relatively better understood from an evolutionary viewpoint (i.e., through comparative analysis). The discussion here will also emphasize movements, but the idea is supposed to apply to actions in any modality.

Why might there be incidental correlations between observable (derived) activities and internal state or future behaviour? One straightforward possibility is that the derived activity is the first element of some complex response, e.g., if a snake rears up as the initial step in the act of striking. In this case the derived activity would be an “intention movement”. Tinbergen (1964) noted that the form of many animal signals was suggestive of their having originated as intention movements:

> . . . many signaling [sic] movements resemble incomplete versions of movements which themselves have another function. For instance, many threat postures involve the first stages of fighting in which the weapons are brought into a position of readiness; birds point the bill at an opponent or lift the carpal joints; fish may open their mouths; many mammals bare their teeth.

Intention movements have not been selected for *per se*; they are simply a physically necessary step in performing an action. A mammal that intends to bite an opponent *must* bare its teeth before doing so. Intention movements thus provide information about future behaviour, and it is not difficult to see how such movements, coupled with the complementary ability to recognize them, might provide the seeds for the evolution of a communication system.

Indeed, Tinbergen seemed to think that it was obvious that intention movements could provide the basis for signal evolution; he was more interested (Tinbergen, 1964) in the idea that movements expressing motivational conflict could also provide the raw materials. The ethologists subscribed to a “pneumatic” theory of motivation: animals were seen as possessing several innate drives; pressure on each drive built up at varying rates depending on the circumstances of the animal; the pressure was released and the drive satiated with the performance of an appropriate behaviour. For instance, Lorenz (1967) discussed at length the drive for aggression. He believed that aggression inevitably builds up over time, but at a faster rate if an animal is crowded or stressed. When the aggressive drive reaches a certain level, it finds expression in behaviour such as an unprovoked attack on a conspecific.

Tinbergen suggested that animals often find themselves with two or more drives simultaneously activated: in a mating situation, for example, both sexual and aggressive or territorial instincts might come into play. He argued that such conflict between drives could be expressed as an observable derived activity, representing some sort of behavioural compromise. The possible outcomes listed by Tinbergen are summarized below.

1. The successive combination of heterogeneous components: typified by the zigzag dance of the male stickleback which moves alternately towards and away from the female.

2. The simultaneous combination of heterogeneous components: such as the final posture of the “meeting ceremony” of black-headed gulls, in which two birds display aspects of ag-

gressive posture (raised carpal joints, used prior to wing-beating) and escape tendencies (lateral orientation and facing away).

3. Compromise movements: for example, facing side-on to the the opposing animal. This is found in mammals, birds and fishes, and Tinbergen assumed that it represents a compromise between an aggressive approach and a retreat.

4. Redirected movements: such as attacking another, typically less dominant, animal.

5. "Displacement" or "extraneous" movements: for example, unexpected feeding behaviour when there is apparently a conflict between attacking and escaping.

Like intention movements, these movements have not been directly selected for, i.e., they have no function. They are, however, on the way to acquiring one. Whether a derived activity is a simple intention movement or the expression of motivational conflict, the result—in the eyes of the ethologists—is the same. The activity is more or less reliably associated with a particular internal state, and possibly the performance of a particular behaviour in the near future. Therefore the activity could convey information to an appropriately equipped observer, and the stage is set for it to be transformed into a true signal through the process of ritualization.

2.2.2 Ritualization

Ritualization is what happens when an initially irrelevant movement such as teeth-baring or a sideways stance, or indeed an action in some other modality, such as the release of a pheromone, starts to be of informational value to other animals. The ethologists, assuming along with Darwin that the transmission of information carried an inherent selective advantage, thought that the original cue—the derived activity—would be exaggerated or stylized in the interests of reducing ambiguity. Thus the term "ritualization".

In short, the original proto-signals are expected to evolve to become more efficient at transmitting information. A signal implies a receiver, and while the ethologists were interested in the problem of responsiveness to signals, their account of ritualization nevertheless focuses on the way the signal itself is shaped over evolutionary time. Tinbergen, noting that such speculation was dangerous in the absence of experimental work, offered the following hypothetical selection pressures on the ritualization process.

1. General improvement: signals will become bigger, louder, brighter, etc., so as to be more easily perceived. This is balanced by selection pressure to avoid predation by other animals that can also detect the signal (Huxley, 1923).

2. Intra-specific distinctness: signals within one species will evolve to become easily distinguished from one another, i.e., ambiguity will be reduced. This relates to the suggestion by Morris (1957) that animal signals are characterized by a "typical intensity" (i.e., they have a stereotypical form) in order to make their accurate recognition easier.

3. Inter-specific distinctness: initially similar signals present in different species may diverge so that they are distinctly species-specific—this is especially important for signals associated with mating, as a way of avoiding fruitless cross-species copulations.

4. Inter-specific similarity: conversely, convergent evolution will lead to similar signal forms across species in some cases (Marler, 1957). For example, all songbirds that give alarm

calls in response to hawks are under the same pressure to give a distinct and noticeable call that is nevertheless hard for the predator to locate. There may also be selection pressure for alarm-calling species living in the same area to give similar calls, and thus benefit from each other's vigilance. This principle contradicts Lorenz's (1935) earlier idea that signal forms were essentially arbitrary.

5. Indirect selection effects: a wide variety of unrelated selective pressures may have some effect on the form taken by a ritualized signal. Tinbergen gives the example of the kittiwake gull, in which the facing-away displays of the chicks are influenced by their precarious cliff-side habitat.

To reiterate, the implicit but central hypothesis in ethology concerning the function of animal communication systems is that they are for transmitting accurate information about an animal's motivational state and likely future behaviour (this may strike some readers as the *only possible* function for animal communication systems; in section 2.3 we will meet an argument that it is not). However, the above list of selection pressures acting on the ritualization process suggests several secondary functions: for instance, the function of a particular mating call might be to communicate a readiness to mate while at the same time, through some acoustic peculiarity of the signal, to *fail* to attract the interest of local predators.

2.2.3 Ethology and group-selection thinking

The reason that the ethologists never questioned the idea that honest communication is necessarily a good thing was because they assumed that the process of natural selection occurred primarily at the level of the group. That is, they believed that animals behaved in the interests of the group or even of the species, rather than acting in accordance with their own interests. For instance, Huxley (1966) argued that the function of the ritualization of signals used in animal confrontations was to promote more efficient information exchange and thus to "reduce intra-specific damage". Tinbergen also worked within a group-selectionist perspective, as evidenced by his definition of communication (1964):

> One party—the actor—emits a signal, to which the other party—the reactor—responds in such a way that the welfare of the species is promoted.

Note that for both Tinbergen and Huxley it is reasonable to suppose that an animal might signal at some cost to itself, e.g., give an alarm signal that increases its own degree of risk, purely for the benefit of others in its group. Given this perspective, there is no reason to imagine that there might sometimes be a selective *dis*-advantage to honest signalling.

2.3 The rise of behavioural ecology and sociobiology

The ethological view of signal evolution was challenged in the 1970s, by the new disciplines of behavioural ecology and sociobiology. Both of these disciplines were concerned with adaptive explanations of social behaviour. An argument was made that, far from maximizing information transmission, many animal signals should be expected to maximize *ambiguity* about the signaller's internal state and future behaviour. To understand the logic behind this argument, it is necessary to review some historical trends in evolutionary biology.

2.3.1 Group selection or selfish genes?

Although Darwin's original presentation of the theory of natural selection (1859) properly suggests that the struggle for existence is a struggle between individuals, and although the modern synthesis (Fisher, 1930; Wright, 1931; Haldane, 1932) of evolutionary theory with Mendelian genetics does likewise, the ethologists were not the only post-Darwinian scientists to fall into the trap of imagining that selection operates primarily at the level of the group or species. For reasons that are beyond the scope of this work, the group-selectionist view steadily gained credence, culminating in the publication of a notorious book by Wynne-Edwards (1962), who maintained that many animals sacrificed their own fitness for the sake of the group. Wynne-Edwards believed that this occurred mainly in the service of population control, e.g., that birds deliberately refrain from breeding when their population density is high, in order to avoid over-taxing the group's food supply.

Williams (1966) was so irritated with what he saw as the fallacious logic of Wynne-Edwards's book that he published *Adaptation and Natural Selection* as a refutation. Williams did not present original experiments or observations but rather used logical arguments and simple mathematical models. Taking Wynne-Edwards's example in which birds supposedly refrain from having offspring in order to prevent a population crash, Williams asks us to imagine the introduction of a mutant that does not possess this altruistic tendency to limit its offspring for the general good, but instead has as many offspring as it can. The selfish mutant will, almost by definition, be fitter than the altruists. Although the population as a whole may suffer food shortages and even extinction, the selfish, fast-breeding mutants will always leave relatively more progeny than the altruists, and thus will come to represent a larger and larger proportion of the population. The somewhat unpalatable message of Williams's work is that animals are just as susceptible to the tragedy of the commons[3] as are selfish, individually rational humans.

Selection at the group level is certainly possible in principle: E. O. Wilson (1975), citing models by Levins (1970) and Boorman and Levitt (1972, 1973), described the conditions under which it can occur. If a large meta-population is distributed across many small, local populations that are relatively reproductively isolated, and if the rate of extinction of these local populations is especially rapid when few altruists are present, then group selection can potentially counteract individual selection. An example would be selection for lower rates of reproduction, and thus lower levels of virulence, in parasitic micro-organisms. Populations of parasites that kill their host too quickly—through rapid increase in their own numbers—will become extinct before transmitting themselves to other hosts, whereas "altruistic" parasites that breed at a slower rate will manage to infect more host organisms before their current host dies, and will thus do better in the long run. Selection over the groups of micro-organisms living in different host bodies would counteract the effect of selection for individuals with higher reproductive rates. However, Wilson argued that this sort of case was exceptional, and that, *contra* Wynne-Edwards, altruistic restraint from breeding was unlikely to occur in large, stable, inter-breeding populations such as seabird colonies and rabbit warrens. In the years following the publication of Williams (1966), group selection came to be seen as a minor force in evolution.[4] Biologists realized that they could no longer explain animal

[3] A situation in which selfishness on the part of individuals leads to disaster for the group or society, so named because of over-grazing of common lands by individual farmers in mediaeval times.

[4] More sophisticated versions of a theory of group selection have since been defended by, for example, Wade (1978) and D. S. Wilson (1975, 1980). These models tend to parcel out the effects of a trait into within- and between-group

behaviour by saying that it was done "for the good of the species".

At around the same time, Hamilton (1964, 1970) was working on the problem of altruism: if individual selection is seen as far more powerful and pervasive than group selection, then explaining animal altruism presents difficulties. Why should animals perform actions that incur a fitness cost in order to help others? Hamilton developed the theory of kin selection, in which he suggested that animals could be expected to behave altruistically towards their close relatives. Hamilton's argument was that, strictly speaking, the units upon which natural selection operates are genes and not organisms. The direct way for genes to maximize their representation in the next generation is to play some role in ensuring that the animal in which they reside is fit, i.e., that it survives to have many offspring. Thus, a gene that causes an animal to metabolize food more efficiently will be favoured. Another way for genes to maximize their representation in the next generation is to cause the animal in which they reside to behave in such a way as to increase the fitness of *other* animals that also carry copies of that particular gene. Animals generally have no way of knowing what another animal's exact genotype is, but close kin are statistically likely to carry the same genes. Thus, a gene that causes an animal to share surplus food with (and only with) its siblings and cousins will be favoured.[5] Hamilton's explanation "takes the altruism out of altruism" by showing that even though an organism may behave in a way that benefits others at a cost to itself, the genes are, one way or another, always selfishly trying to maximize their own rate of reproduction. After Hamilton's work, evolutionary biologists thought of animals as maximizing their *inclusive fitness*, i.e., maximizing their own number of offspring, and maximizing the number of offspring of each of their relatives, with the latter factor weighted by the degree of relatedness between the two organisms.

The ideas of Williams and Hamilton were popularized by Dawkins (1976) in the aptly-titled book *The Selfish Gene*. Dawkins described genes as the original replicators: blind, self-reproducing machines that, over millions of years, have developed increasingly sophisticated ways of reproducing themselves, such as the bodies and behavioural repertoires of complex, multi-cellular animals.

2.3.2 Animals as maximizers and the application of game theory

When the muddy waters of naïve group selectionism had drained away, theoretically minded biologists found themselves with a simple prediction: that animals can be expected to maximize their inclusive fitness. This greatly facilitated the mathematical modelling of animal behaviour, and particularly social behaviour such as communication. If a plausible function specifying inclusive fitness in relation to a range of behavioural variables could be given, then the mathematics of optimization could be used to predict the evolutionary outcome. Consider the time allocation problems facing a bird caring for several nestlings. The bird's inclusive fitness might reasonably be measured in terms of its chances of survival until the next breeding season, and the chances of its chicks surviving. These probabilities of survival would in turn be complicated functions of the

effects, and the (perfectly valid) argument is that if the latter is stronger than the former then group selection can occur. However, Dugatkin and Reeve (1994) have pointed out that these more advanced models can all be re-phrased in terms of individual selection, in much the same way that Hamilton's work takes the apparently group-level concept of altruism and re-casts it in terms of individual advantage.

[5] Note that a gene for altruism directed towards any and all group members and conspecifics would not be favoured—kin selection is not group selection. Indiscriminate altruism could only succeed if, through some peculiarity of its lifestyle, an animal happens only to interact with close kin.

amount of time spent by the parent away from the nest, the rate at which the parent could collect food, the risk of predation when the nest was unguarded, etc. Some of these values are under the bird's control, such as the duration of foraging trips, and others are not, such as the hourly risk of predation of an unguarded nest. If we could specify the bird's inclusive fitness as a function of those variables under the bird's control, we could then find the maximum value of that function. The logic of Williams and Hamilton predicts that the various aspects of the foraging behaviour of this species (e.g., average length of foraging trips, average inter-trip interval) should evolve so that it achieves, or at least approaches, the identified maximum fitness. Birds that behave sub-optimally will have fewer offspring and thus will tend to be weeded out by natural selection.

In fact, many of the models developed in theoretical biology have considered only the simple case of interactions between non-relatives, and thus have only had to deal with straightforward individual fitness. However, there is another complicating factor: often the fitness consequences of a particular behaviour are not static, but depend on the behaviour of other animals. This applies especially to social behaviour. For example, the overall fitness consequences for a male peacock of having a large, ornamental tail are dependent on the mate choice strategies of the local females. The fitness consequences of having an unusual diet, perhaps eating leaves instead of fruit, depend at least partly on how many of one's conspecifics share the unusual preference—there may be an advantage at first, because leaves are plentiful, but then if the majority of the population switches to the new behaviour, the advantage disappears. These sorts of problems inspired Maynard Smith (1974b, 1979, 1982) to apply a branch of mathematics known as game theory to animal behaviour.

Game theory (von Neumann & Morgenstern, 1953; Binmore, 1992) looks at the question of what strategy a rational player should adopt in a given game. The player is assumed to want to do as well as possible, i.e., to extract the maximum profit or payoff on some utility metric, but the player knows that they face an opponent or opponents who *also* want to do as well as possible. For example, if we take the game of poker: one "strategy" for doing well is to play against opponents who perversely give away all of their money. Game theory is silent on such cases. However, if one's opponents are rational agents who want to win, then game theory can be used to find the balance between, say, bluffing and conservative play that is most likely to maximize one's own winnings.

Game theory is applied to animal behaviour by assuming that animals are playing games against each other: for instance, that a male seeking to attract mates must choose an advertisement strategy, and that a female must decide on a choice strategy, and that they then play out the sexual selection game. The quantity that the animals want to maximize is, of course, their inclusive fitness—in order to model communication, for example, the game theorist must assume that such things as the cost of making a particular signal or the benefits of inducing a particular response can be quantified in this common currency. The further necessary assumption that players are rational at first appears to present problems: few would want to claim that a cockroach or an amoeba was capable of rational thought. However, the cumulative effect of natural selection on genetically specified strategies takes the place of the rationality assumption (Maynard Smith, 1982; Binmore, 1992). Over evolutionary time, animals will come to behave *as if* they were rational agents, because anything less than rational play will be open to exploitation by other players and thus lead to lower fitness.

An important concept in game theory is the idea of a Nash equilibrium, which denotes a situation where all players are using strategies such that no player can improve their expected payoff by changing his or her strategy. Each player is giving the "best reply" to the strategies used by their opponent(s). Game theory predicts that players capable of adjusting their strategies over time—through pure rationality, through trial and error, or through evolution—will sooner or later arrive at a Nash equilibrium.[6]

Maynard Smith and Price (1973; Maynard Smith, 1982) took the Nash equilibrium idea and devised the more specific concept of an evolutionarily stable strategy (ESS). If a population of players are repeatedly playing a particular game, an ESS is a strategy that, if used by almost all players, cannot be invaded by any other strategy. Uninvadability is achieved mainly because the ESS is a best reply to itself. However, it must also be the case that if an equally good alternative reply exists, then two players adopting the ESS against each other fare better than two players playing the alternative. Technically, if Π_{ij} is the expected payoff for a player using strategy i against a player using strategy j, and m represents a mutant strategy, then the strategy s is an ESS if, for all m:

$$\begin{gathered}\Pi_{ss} > \Pi_{ms} \\ \text{or} \\ \Pi_{ss} = \Pi_{ms} \text{ and } \Pi_{ss} > \Pi_{mm}.\end{gathered}$$

In other words, for a strategy to be an ESS requires that other strategies will not do as well if pitted against it. The ESS idea proved extremely fruitful in the modelling of animal behaviour, particularly as a way of discounting hypotheses: if, given some plausible payoff values, it could be established that a hypothesized strategy was *not* an ESS, then game theory predicted that such a strategy would not persist over evolutionary time but would be invaded by another. Biologists interested in offering functional explanations for animal behaviour now had to consider whether their accounts qualified as ESSs or not.

2.3.3 A new perspective on signalling

The idea that animals are inclusive-fitness maximizers, and the use of game theory as a modelling tool, led to a view of animal communication that was very different from that of the ethologists. The argument was as follows. Communication systems are often associated with competitive contexts, e.g., males signal their quality when competing for access to females; aggressive signals are used in disputes over food or territory. (Indeed, selfish-gene logic suggests that, by default, all animals should be seen as being in competition with their conspecifics to leave the most offspring.) In these competitive interactions, game-theoretic analysis shows that it will not be evolutionarily stable for animals to honestly signal their motivations or intentions, because a strategy of bluff and exaggeration will invade any honest strategy. Therefore the appearance of "signalling" and "communication" in these interactions is in fact an illusion; the animals in question are not attempting to transmit information but are trying to avoid doing so.

[6] Those readers familiar with game theory will realize that this statement is something of an oversimplification. For instance, some games have *no* Nash equilibria, and sometimes a Nash equilibrium exists but has no basin of attraction in the strategy space of the game and is thus highly unlikely to be reached.

For example, Maynard Smith (1974b, 1982) described a model of conflict over resources in a species that lacks any dangerous weaponry. The model is referred to as the "war of attrition", for reasons that will become clear. Maynard Smith asks us to suppose that two dung-flies are competing for possession of a dunghill. They cannot injure each other, and so the battle for possession is won or lost based on which fly is prepared to out-wait the other. However, waiting incurs costs, such as the risk of predation, the loss of feeding opportunities, and so on. Might it be possible for the dung-flies to avoid these costs by using a communication system, in which, at the beginning of the contest, each fly signalled the length of time that they would be prepared to wait, and possession of the dunghill went to the fly that indicated a greater degree of patience? The answer is no. Any such communication system would not be an ESS. A dishonest mutant that wildly exaggerated its planned waiting time would always win against honest signallers. Then selection would favour flies that "called the bluff" of the mutant, and settled down to wait regardless of the exaggerated signal. The communication system would soon disappear. As game theorists are wont to say "talk is cheap".

The ESS in the war of attrition in fact calls for the two opponents to randomly choose waiting times from a negative exponential distribution: this ensures that their likelihood of giving up the duel during the next arbitrary time unit is always constant. (Randomly choosing a waiting time of between one and ten seconds, for example, could not be an ESS, as it would open the way for a mutant that was always prepared to wait up to ten and a half seconds.)

The war of attrition model suggested that animals with conflicting interests will not communicate, but instead maintain a "poker face" concerning their intentions. A dung-fly that gave some sign, such as moving its wings or rubbing its antennae together, indicating that it was about to give up would be open to exploitation by its opponent: the second dung-fly would be motivated to wait just a little longer. Thus stereotyped "signals" will be favoured. If all animals display the same signal in a particular context, regardless of their intentions, the signal carries very little information about what the animal will do next. The interpretation of the phenomenon of typical intensity (Morris, 1957) is completely reversed: it is explained as a way of maximizing ambiguity about the animal's underlying motivations.

Even in situations where animals do not appear to be in conflict, ESS models often indicate that communication should not be expected to be stable. As noted in section 1.2, for example, it is not an ESS for animals to give alarm calls for the benefit of unrelated conspecifics if calling increases the risk to the signaller. This is because a cheating strategy, whereby an animal never gives alarm calls but gains the benefit of others' calls, will be able to invade. The introduction of ESS modelling thus highlighted the problem of signal honesty: what keeps a given signalling system from degenerating into bluffing or cheating? To date, this question remains central to the theoretical biological literature on communication.

The exception to the general finding that so-called signals were actually about minimizing information transfer was the case of unfakeable signals of strength or size. Maynard Smith (1982), discussing animal conflicts, notes that there are two kinds of information transfer that are relevant:

> (i) Information about 'motivation' or 'intentions'. Because any message about motivation can be sent, with little cost, there is no reason why such messages should be accurate, and therefore no advantage in paying attention to them.

> (ii) Information about 'Resource-Holding Power'... RHP is a measure of the size, strength, weapons, etc. which would enable an animal to win an escalated contest. It can be evolutionarily stable to transmit information about RHP, and to accept such information to settle a contest, provided two things are true. It must be impossible to transmit false information about RHP, and it must be expensive to acquire high RHP in the first place.[7]

For instance, due to the physical connection between body size and the pitch of a vocalization, deep roaring sounds may count as unfakeable signals of size and strength. This depends on the reasonable assumptions that there is no way to fake the low pitch of the roar, and that size and strength are costly to acquire. The notion that signals might be made unfakeable through the costs associated with their production is explored further in section 2.4.

2.3.4 Reinterpretation of the ethological data

Attempting to verify the behavioural-ecological position that held animal signals to be uninformative, Caryl (1979) re-analyzed ethological data collected by Stokes (1962a, 1962b), Dunham (1966) and Andersson (1976) on threat displays in birds. Caryl found that aggressive displays were poor predictors of subsequent attack, i.e., they did *not* carry reliable information about aggressive intent. This result is exactly what we would expect if animals were under selection pressure not to give away information in contests. On the other hand, Caryl found that certain displays were good predictors of the intent to flee; however, this is not as damaging to the theory as it might seem. Knowing when an opponent will flee is not the converse of knowing when the opponent will attack. Furthermore, as Maynard Smith (1982) points out, once an animal has made the decision to abandon a resource, the situation changes: it makes sense to signal an imminent retreat and thus avoid being attacked in the meantime.

Hinde (1981) responded to Caryl, arguing that he had made a straw man of the traditional ethological position. Hinde claimed that animal threat displays might well not have a direct and simple correlation with subsequent attacking behaviour. Instead, they could be interpreted as having conditional content, such as "I will stand my ground and will retaliate if you attack." If such a threat was effective in dissuading opponents, then it would not actually have to be carried out, and we would not expect to find a correlation between the performance of the threat display and a subsequent attack. Hinde's argument does point to some of the difficulties in analyzing ethological data. It is often the case that although we know what the animals have done, we are more interested in counter-factual questions, e.g., in what they *would* have done had their opponent not backed down.

Nevertheless, Caryl (1982) rejected Hinde's conditional-content hypothesis, primarily because of its lack of parsimony. Caryl claims that the most economical way to account for the data of Stokes, Dunham and Andersson is to suppose that the animals involved do not possess reliable systems by which they communicate about aggressive intent. Caryl allows that there may well be statistical complexities arising from the interaction between the protagonists, but argues that unless testable hypotheses are formulated then all else is handwaving. In this context it is ironic to note that Hinde baldly asserts verbal arguments about signalling strategies, such as "... but if 'Stay

[7] Maynard Smith credits Parker (1974) with the original term "resource-holding power", but in the more recent literature, RHP has come to denote the equivalent phrase "resource-holding potential".

or probably flee' is shown, the reactor may do well to go in and supplant the signalling individual". Hinde appears not to realize that the game-theoretic approach advocated by the behavioural ecologists is an excellent tool for untangling such questions as the adaptive value of a conditional threat strategy.

2.4 Zahavi's handicap principle

Zahavi (1975, 1977) suggested that honesty could only be maintained in a communication system if the signals were costly in some way. He was working within the framework of individual-selection thinking—well aware that animal behaviour had to be explained in terms of fitness maximization—and he proposed the counter-intuitive idea that signallers sacrifice some of their fitness (i.e., impose a handicap on themselves) in order to produce signals that will be believed by receivers. The handicap principle will be of particular importance in later chapters and so it will be introduced here at some length.

2.4.1 Paradoxical logic of the theory

Zahavi intended his theory to account for communication systems of all kinds[8] but it is most easily explained with reference to sexual advertisement signalling. Assume that males vary in quality (i.e., vigour or viability), and that females are interested in mating with high-quality males. The stage is then set for a communication system in which males signal their quality to females, and are rewarded with a mating episode if they "convince" a female that they are of high quality. The process of signal ritualization might begin with some visible trait that was correlated with male quality, e.g., if high quality males tended to have slightly longer tails, then long tails could become exaggerated into a signal. Zahavi agreed with the behavioural ecologists that if growing a longer tail was cheap, i.e., if it had little deleterious effect on male fitness, then the signalling system would be vulnerable to bluffing in the manner described in section 2.3.3. *All* males would come to have long tails and the female preference for longer-tailed males—the reception component of the communication system—would no longer have any selective advantage and would therefore disappear.[9] The temptation for the males to bluff would have destroyed the stability of the system.

The critical point in Zahavi's logic was to consider what would happen if growing a long tail was costly in fitness terms, e.g., if the metabolic resources necessary to grow the tail detracted from a male's ability to resist parasites, or if having a long tail made it more difficult for a male to escape from predators. In this case, he argued, the communication system cannot be corrupted by bluffers: lower quality males cannot afford to devote the necessary resources to growing a long tail. Tail length becomes an honest indicator of male quality because "cheating" is prohibitively expensive. Zahavi reasoned that only those communication systems in which the ritualized signal happened to be costly would escape collapse due to bluffing. Therefore, the stable communication

[8] The very broad scope that Zahavi believes his theory to have has become particularly clear in recent years; in Zahavi (1987, 1991) and Zahavi and Zahavi (1997) the handicap principle is presented as an explanation for everything from suicide attempts to method acting.

[9] The prediction that the female preference would actually *disappear* assumes that there is some level of cost associated with it, e.g., the time cost involved in examining tails and deciding between their owners. If having the preference incurs minimal or zero cost, then it might remain, leading to a situation where females preferred long tails even though they were not a reliable signal of quality: this is effectively the Fisher process, discussed in section 2.7 and in chapter 9.

systems that we observe in nature are maintained by this mechanism, which Zahavi dubbed the handicap principle.

Bullock (1997a), among others, has noted the similarity between the handicap principle and the economic idea of conspicuous consumption, originally described by Veblen (1899). Conspicuous consumption is the notion that wealthy individuals display their status in an ostentatious but wasteful manner, by such means as paying for elaborate ice sculptures at a dinner party, or through filling a swimming pool with champagne. These displays are necessarily accurate signals of wealth because no-one of more modest means could afford to produce them. Their message is "Look at me, I can afford to throw money away like this, so I must be genuinely rich." Zahavi's theory appears to be a good candidate for explaining similar levels of ostentation in some natural signals, the peacock's tail being the obvious example. The message behind the peacock's tail can be seen as something like "Look at me, I can afford to produce this elaborate display and I am still alive, therefore I must be of genuinely high quality."

Recalling Darwin's discussion of sexual advertisement signalling (section 2.1.2), we can see that the handicap principle, if it works in practice, can provide an explanation for otherwise mysterious female preferences: it may be that peahens are interested in the size and splendour of peacock's tails because they are a source of reliable information about male genetic quality.

2.4.2 Controversial status and variant interpretations

When the handicap principle was first introduced, it was generally not accepted by theoretical biologists. Population-genetic models[10] (Maynard Smith, 1976; Kirkpatrick, 1986) seemed to show that it could not be evolutionarily stable. Dawkins (1976) suggested that although the offspring of a successful male will inherit their father's high quality, they will also inherit the genes for the costly handicap that their father used as a signal: thus they may be no better off than unhandicapped males of low quality. However, the potential effectiveness of the handicap principle has been validated by several mathematical models in recent years; foremost among these is Grafen (1990a). Grafen's model, framed in terms of sexual selection, establishes that the handicap principle can work, but specifies an important proviso: the unit cost of producing the signal must be greater for a low quality signaller than for a high quality signaller. In other words, the fitness cost of extending one's tail by an extra centimetre must be higher for unhealthy or weak males than for healthy strong ones.

The handicap principle was maligned and misunderstood because Grafen's proviso about differential unit costs was not clear from Zahavi's original formulation, and because several distinct interpretations of the principle are possible. Zahavi's tendency towards a rhetorical style of argument probably bears some of the blame for this. Iwasa et al. (1991), following Maynard Smith (1985), attempted to cut through the confusion. They detailed three variant interpretations of the handicap principle, and suggested that different findings concerning the evolutionary stability of handicapped signals could be explained by the fact that some authors were modelling one variant and others another.

[10] Population-genetic models are mathematical models of changes in actual gene frequencies in an evolving population; for an introduction see Maynard Smith (1989). Game-theoretic models are, in comparison, more abstract: they deal with the evolution of behavioural strategies without regard for the details of the underlying genetics. The problem of which kind of model to use to capture a particular biological phenomenon is touched upon in section 2.7.1.

The variants are detailed below; they are described in terms of sexual selection. (Indeed, Iwasa et al.'s typology does not sit easily as a description of handicapped signalling in non-sexual cases.) The male advertisement trait is assumed to be tail length. Males also differ on a general viability trait which may be genetically or environmentally determined, and is a measure of their quality as mating partners. Females cannot perceive the viability trait directly, but it is in their interests to mate with males of high viability. Females can, however, perceive the length of a male's tail, i.e., the phenotypic expression of his advertisement trait.

Pure epistasis handicap

In this version, a particular gene wholly determines a male's tail length, and the longer his tail, the less likely he is to survive to reproductive age. However, his survival is also determined by his viability: more viable males are more likely to survive, and for any given level of viability, a male is more likely to survive if he has a *shorter* tail. Therefore the males that are most likely to die before reaching reproductive age are those of low viability with long tails. Observing the adult population, one would find a correlation between the genes for viability and tail length. In technical terms, epistatic selection has resulted in linkage disequilibrium; in plainer language, long tails are linked to high viability, because all the long-tailed low-viability males died young. In consequence, a female's preference for mating with long-tailed adult males will mean that she is more likely to achieve her goal of mating with a high-viability male.

The function or selective advantage of having a long tail, in this version of the theory, is that it will serve as a genuine handicap, increasing a male's risk of premature death, but also increasing his prospects of being selected as a mate if he survives. The earliest models of the handicap principle were of this simple form (Davis & O'Donald, 1976; Maynard Smith, 1976; Bell, 1978), and generally concluded that such a system would not be evolutionarily stable—Iwasa et al. (1991) concur.

Conditional handicap

In this case a long tail still reduces a male's chances of survival to reproductive age, and again survival is primarily determined by viability. However, the expression of the gene for tail length is modified by the viability trait: males of lower viability will not realize their full, genetically specified tail length but will grow a proportionately shorter tail. It is assumed that only the most viable males have the resources to fully realize the tail length encoded in their genes. Because the expression of the tail-length gene is viability-dependent, observable tail length is correlated with viability even before mortality has taken its toll. A female preference for long tails will therefore translate into a preference for more viable males.

The function of a long tail is thus to stand as a surrogate or signal for high viability. This signal cannot be faked, because long tails are developmentally impossible, or at any rate too costly, for low-viability males to produce. The model by Grafen (1990a) that famously vindicated the handicap principle was approximately of this form.[11] Other, similar models include those of Nur and Hasson (1983) and Andersson (1986). Iwasa et al. conclude that the honest signalling of viability via the advertisement trait can be evolutionarily stable, if the conditions of this version

[11] In fact Grafen's model was a little more complex, looking at the evolution not of a simple gene for tail length, but at the evolution of mapping relations between underlying viability and expressed tail length. The importance of this difference in accounting for Grafen's findings will be explored in chapter 9.

of the handicap principle are satisfied. The conditional handicap model can therefore potentially explain some real-world signalling systems.

Revealing handicap

In this version the expressed tail length of males is determined directly by a gene, as with the pure epistasis handicap. Survival to reproductive age depends on viability modified by tail length, as before. However, when the males reach reproductive age and are competing to be selected by females, only highly viable males succeed in maintaining their tails at their original, genetically specified length. Males of lower viability are less well able to withstand the rigours of their environment, and their tails are shortened due to, for example, attacks by predators or parasites. Low viability males *reveal* their status by tending to have shorter tails as adults. Females preferring to mate with long-tailed males will thus mate with more viable males on average.

The function of a long tail is therefore to index viability by being potentially vulnerable to environmental degradation, but to a lesser extent for higher-viability males. The revealing handicap is quite similar to the conditional handicap, except that the interaction between the advertisement trait and viability takes place *after* the effects of premature mortality. Hamilton and Zuk (1982) first proposed this version of the handicap principle. They phrased their model not in terms of tail length, but in terms of bright, colourful plumage that could only be maintained in good condition by highly viable males. Iwasa et al. also modelled the revealing handicap and suggest that it too could be the basis for evolutionarily stable signalling of male quality.

2.4.3 Variable costs and benefits for signallers

There has been great theoretical interest in the handicap principle in recent years, and several authors have looked at the effects of altering the relationship between signaller quality and signal cost, and also the possible relationship between signaller quality and the benefit from a positive response. Addressing the latter issue, Godfray (1991) argued that the logic of the handicap principle is not limited to situations in which signallers advertise their quality to interested receivers, as in sexual signalling, but can also be applied to situations in which signallers advertised their degree of *need* to receivers, as in the begging signals of nestling birds. In Godfray's model, signals of hunger are made reliable by the energetic costliness of begging cries. However, the condition stipulated by Grafen (1990a), that the cost of giving a particular signal should be higher for lower-quality signallers, does not apply. Instead Godfray has reversed the situation, such that the *benefit* of a particular positive response (i.e., the value of an item of food given by the parent) is greater for higher-need signallers. The cost of making a begging cry of a given intensity is the same for all signallers. Under these circumstances, Godfray shows that honest signalling of need can be an ESS. Maynard Smith (1991) reaches effectively the same conclusion in a general model of signalling between relatives.

Zahavi originally argued (1975, 1977) that the handicap principle was universal, and that only signalling systems in which the correct relationship existed between signaller quality and signal cost would admit of stable signalling equilibria. However, recent models have suggested that while Zahavi's principle can work under certain circumstances, it is not the universal mechanism that he imagines it to be. Spurred by results like Godfray's, authors such as Hurd (1995) and most especially Bullock (1997a, 1997b) have examined more closely the effects of different relationships between signaller quality, signal cost, and the benefit that signallers gain from a positive response. In the most general model to date, Bullock (1997a) shows that honest signalling of quality will be an ESS whenever the rate at which signals become cheaper for higher-quality signallers is greater than the rate at which positive responses become more valuable for lower-quality signallers. This is not an easy idea to get across in a few sentences, and the interested reader is referred to Bullock's work, but the essence of it is as follows. The condition referred to above as Grafen's proviso—that the unit cost of signals must be greater for lower- than higher-quality signallers—is neither necessary nor sufficient for honest signalling. The handicap principle describes some of the situations in which honest signalling is an ESS, but not all of them. Conversely, there will also be situations in which Grafen's proviso is satisfied, but honesty is not an ESS. To give a single example of the latter case, if low-quality males pay higher unit costs of signalling, but also gain much greater benefits from a positive response than high quality males, then the honest signalling of quality will not be evolutionarily stable. To flesh out the example, suppose that we are dealing with sexual signalling in a species like the elephant seal (*Mirounga angustirostris*), in which dominant, high-quality males monopolize harems of females, and thus a single copulation could be of enormous benefit to a low-quality male because he is unlikely to get many such opportunities.

Close attention to costs and benefits in successful models of the handicap principle reveals that it is perhaps badly named. In a model where the handicap principle can be shown to work, it is never the case that signallers are imposing a real handicap on themselves—that is, they never suffer a *net* cost because of their signalling strategy. It is simply the case that given the way that signal costs and benefits relate to signaller quality, it turns out that a high quality signaller maximizes its fitness by signalling in one way, and a low quality signaller does so by signalling in another.

2.5 Dawkins and Krebs on communication

In contrast to Zahavi's emphasis on honesty, Dawkins and Krebs argued that animal communication is fundamentally about signallers selfishly manipulating the behaviour of receivers (Dawkins & Krebs, 1978; Krebs & Dawkins, 1984). Their account is in sympathy with the basic behavioural-ecological position on animal signalling described in section 2.3.3, in which animal "signals" are not communicative but instead serve to minimize information transfer in competitive interactions. However, Dawkins and Krebs go further, and suggest that rather than simply remaining poker-faced about their own intentions, animals will actively mislead and manipulate others if it is to their advantage to do so. Dawkins and Krebs thus plug an obvious hole in the behavioural-ecological account, because they offer an explanation for why animals should signal at all (rather than remaining perfectly still or silent, for instance); they also make general predictions about the form that signals are expected to take.[12]

2.5.1 Information or manipulation?

Dawkins and Krebs (1978) note that it is often in an animal's interest to manipulate objects in its world. Sometimes the object in question is inanimate or at least immobile, as when a cow ingests grass, a rabbit displaces earth in the process of digging a burrow, or a fish pushes against

[12] Although the "signals as manipulation" view is widely credited to Dawkins and Krebs, Andersson (1994) notes that it was to some extent pre-empted by Emlen (1973). Wilson (1975) also advances a similar position.

the surrounding water in order to propel itself. On the other hand, sometimes the object is another animal: for instance, it is in the interests of predators to ingest prey, males to inseminate females, and territory holders to repel intruders.

If an animal seeks to manipulate the inanimate environment, it generally has no choice but to use its own muscle power to achieve its goals. However, when one animal seeks to influence another, a second strategy exists besides pushing the other around with brute force. Animals have sensory systems and respond to certain stimuli in predictable ways. It is often possible for one animal to stimulate the sensory system of another, and thereby exploit the muscle power of the second animal, causing it to behave in a way that benefits the first. For example, a male frog does not actively move about seeking females, but instead sits in one place and makes sounds that cause females to approach him. His croaking can be seen as a way of exploiting the females' locomotive muscle power. Similarly, the angler fish attracts smaller fish by the use of a lure that hangs near its mouth; the lure mimics the motion and appearance of a worm. Smaller fish approach seeking food and are themselves eaten. The angler fish is clearly exploiting the sensory system and response patterns of its prey.

Dawkins and Krebs (1978) suggested that animal communication should be defined in this way; that communication or signalling is what happens when one animal, the actor, has been selected to produce a response in a second animal, the reactor, such that the reactor's behaviour changes to the advantage of the actor. (Of course, the signalling behaviour need not result in a benefit for the actor every time it is performed; it need only be beneficial on average.) This is an explicitly functional definition of communication: the function of any signalling behaviour is to cause responses that increase the inclusive fitness of the signaller.

Dawkins and Krebs acknowledged that such a definition moves a long way from our everyday understanding of the word "communication", and that there is no implication, in their view, that animals should be transmitting information to each other in order to qualify as communicating. Indeed, the way their argument is presented suggests that deception should be the norm. They admit (p. 283) that they are "tempted to abandon the word communication altogether." Dawkins and Krebs argued that our perception of animal communication, and in particular the way the ethologists saw communication, has been coloured by the conduit metaphor for human language use (two speakers using language to exchange truthful information for mutual benefit). They made a strong case that animal communication serves not to inform but to persuade, and that advertising and propaganda are more apt metaphors than language for what goes on in the animal kingdom.

In Dawkins and Krebs's signals-as-manipulation definition, it is irrelevant whether the reactor profits or loses from its response to the actor's signal. Sometimes it will be to the reactor's advantage to respond as the actor would like it to, as when a bird responds to the begging cries of its chick by feeding it and thereby increases its own inclusive fitness. This presents no special difficulty for the theory, however, and Dawkins and Krebs argued that it is still reasonable to view such a case as the manipulation of the parent by the young.

If, on average, it is to the reactor's disadvantage to be manipulated, then selection will of course act to reduce its tendency to respond. Dawkins and Krebs claim that this will lead to an evolutionary arms race between actors and reactors (much like the arms races between predators and prey) in which each side will develop successive adaptations and counter-adaptations to be more persuasive on the one hand, and more sceptical and resistant on the other. For example, if there is no net benefit for female frogs in approaching a calling male, perhaps because most males are under-sized and of poor genetic quality, then females will be selected to be more and more discriminating, approaching only the deeper, louder calls of larger males, while males will be selected to call as deeply and loudly as possible. The signalling system that evolves is not expected to be stable, and may eventually collapse: for instance, a point may be reached where females cannot discriminate any more finely, and they might then do just as well to approach males randomly again. Note that these co-evolutionary arms races represent the same phenomenon that the ethologists saw as ritualization, i.e., selection for exaggerated signals in the service of reducing ambiguity—with a change of theoretical perspective the interpretation is completely different.

It may also be the case that, despite the appearances of a disadvantage to the reactor, in the long run it is worthwhile for the reactor to maintain its responsiveness. If, for example, the tendency to respond to a certain stimulus is beneficial outside the signalling context, then it may be maintained despite exploitation. As Maynard Smith and Harper (1995) put it, “There are a lot more worms than angler fish lures”: the small fish that approach what they believe to be worms are better off continuing to do so, even though sometimes the worm turns out to be an angler fish. (This is approximately the sensory bias paradigm, discussed in section 2.6.1).

2.5.2 Two kinds of signal co-evolution

Dawkins and Krebs (1978) were criticized on two points: firstly, that they had neglected the active role of receivers or reactors, and tended to caricature them as passive agents that could be manipulated towards any end; and secondly, that they had failed to address the existence of co-operative communication systems among, for example, highly social species like bees and primates (Hauser, 1996). Dawkins and Krebs responded with a revised version of their theory (Krebs & Dawkins, 1984), in which they paid more attention to the role of the receiver. Specifically, they suggested that receivers were under selection pressure to be good “mind readers”, i.e., to critically assess the behaviour of others and to exploit any tell-tale predictors about their future behaviour. In the case of signalling behaviour, mind reading means being able to extract useful information from what is inevitably an exaggerated sales pitch. For example, adult birds will be selected for their ability to discriminate between chicks that are genuinely hungry and chicks that are giving grossly exaggerated begging calls.

As before, however, successful mind readers set the stage for their own exploitation by manipulators. Suppose that a certain population of dogs have developed the mind-reading ability to connect bared teeth in an opponent with an imminent attack. When they observe an opponent with bared teeth, these dogs flee in order to avoid injury. Manipulation occurs when those being mind-read fight back, influencing the behaviour of the mind readers to their own advantage. For example, a dog could bare its teeth despite not having the strength or intention to attack, and thus scare off its mind-reading opponent. Krebs and Dawkins again predict evolutionary arms races between manipulative signallers and sceptical receivers: “selection will act simultaneously to increase the power of manipulators *and* to increase resistance to it” (p. 390).

Krebs and Dawkins admit, however, that not all interactions are competitive in nature. There are some situations in which it is to the receiver’s advantage to be manipulated by the signaller. For example, a pack-hunting predator may attempt to recruit a conspecific in order to bring down prey too large for either to tackle alone. Foraging bees, on returning to the hive, may indicate to their closely related hive-mates the direction and distance to a source of nectar. In these cases the receiver's compliance is to the benefit of both parties, i.e., there exists the possibility of co-operation. Krebs and Dawkins argue that when the two parties share a common interest in this way, then a different kind of signal co-evolution will result. Specifically, there will be selection for signals that are as energetically cheap as possible while still being detectable; Krebs and Dawkins suggest the phrase "conspiratorial whispers" to describe these signals. Rather than signallers needing to be more and more extravagant in their attempts to persuade receivers, the opposite process occurs: receivers are eager to be persuaded, and selection will favour subtle signalling and low response thresholds. An implication is that the louder and costlier signalling displays of the animal world—such as roaring contests in red deer or male plumage in birds of paradise—may have been over-represented in studies of animal communication simply because they are obvious to human observers. There may be a great deal of conspiratorial, co-operative signalling going on that is too subtle for us to notice; this intriguing idea has been too little explored.

2.6 Receiver psychology

The discussion of Dawkins and Krebs's ideas completes our review of the major theoretical positions on the function of animal signalling systems. However, several recent authors have put forward modifications to the standard behavioural-ecological position, or to its differing extensions due to Zahavi and to Dawkins and Krebs, that are of interest. The case has been made by Guilford and M. S. Dawkins (1991; see also M. S. Dawkins, 1993) that behavioural-ecological accounts of communication have, in general, placed too much emphasis on the signalling side of the equation. Guilford and Dawkins discuss "receiver psychology", and stress the idea that the receivers of signals are not just there to be manipulated, but have their own agenda and exert an effect on signal design considerations. In particular, if signals are to have their intended effect on receivers, then they must be detectable, discriminable and possibly memorable, not in some absolute sense, but in terms of the receiver's "psychological landscape".

Guilford and Dawkins propose a distinction (later taken up by Johnstone, 1997) between "strategic design" and "efficacy" in accounting for the function(s) of natural signalling systems. The strategic design of a signal refers to the way that it has been shaped by natural selection such that it is in the interests of receivers to respond to it; strategic design is about "whether or why... the receiver responds appropriately" (Guilford & Dawkins, 1991, p. 2). For example, the fact that a signal provides honest information to the receiver because of cost constraints on its production—in line with the handicap principle—is part of its strategic design. Note that Guilford and Dawkins are not really saying anything new here: their point is simply that to understand why signallers and receivers are participating in the communication system, we need to look at the underlying game that the animals are playing.

However, the strategic situation is not the whole story. Guilford and Dawkins point out that strategic theories of signalling are not likely to have any success in coming to grips with the enormous range of signal *forms* that occur in nature. Two different bird species could both evolve honest signals of male quality in a sexual signalling context, and the functional or strategic story behind each signalling system might be the same, but that could never explain why in one species the females look for bright red throat patches and in the other for long black tails. Hoping to account for these differences, Guilford and Dawkins introduced the concept of efficacy. Whether a signal is informative or manipulative, the signaller's most basic concern is with getting the message across: a good signal must be effective.

Most obviously, the signal must be noticeable, and it must be noticeable given the sensory equipment of the intended recipient. For example, if a burrowing mammal has poor eyesight but an excellent sense of smell, olfactory signals will be favoured. It should also be easy for the receivers to discriminate between the signal in question and other signals. Guilford and Dawkins give the example of roaring in red deer: this is a graded signal, where deeper, louder and longer roaring is indicative of strength and stamina. The suggestion is that selection has favoured roaring and not some other display because it is easy for receivers—given the kinds of ears and the kinds of brains that they have—to discriminate between different levels of the signal. Finally, in signalling systems that involve learning, such as the warning coloration signals of unpalatable insects, which some of their predators learn by experience, memorable signals will be favoured.

Guilford and Dawkins, along with Johnstone (1997), further argue that it is not only the sensory and psychological mechanisms of the intended receivers that will influence the form of signals, but also such factors as the physical environment in which the signal is transmitted, and the possibility of interference due to the activities of other signallers. To illustrate the influence of the environment, Guilford and Dawkins give the example of bird song. Birds living on open grassland tend to have songs of a higher pitch than birds living in forests (Morton, 1975), presumably because high frequency sounds carry well in the open, but are more likely to be degraded by reflections off dense vegetation. The principle is clear: if selection at the strategic level dictates that a signalling system should exist at all, then selection at the "tactical" level of efficacy will tend to favour variant systems that work well in their physical, psychological and social environments.

We must be careful to distinguish considerations of efficacy in signal design from Tinbergen's question of mechanism (see section 1.1). The two are related but distinct. Taking the example of vervet monkey alarm calls, to inquire about efficacy is to ask *why* selection has favoured signals that are given vocally, rather than visual or olfactory signals for instance, and why the alarm calls have come to possess the particular spectral properties that they do. The answers to these questions, according to Guilford and Dawkins, lie in the historical details of the vervet monkeys' perceptual and neurological systems, and perhaps in the acoustic qualities of the savannah. To ask about mechanism, on the other hand, is to ask *how* the signals are produced and received in the modern animal: knowing the answer to this second question will be useful in answering the first, but the questions are not synonymous.

Both efficacy-based and strategic approaches must be incorporated in any complete theory of signal function. Dawkins and Krebs (1978) allude to both factors: in signalling arms races, signallers will be selected for efficacy in the form of more elaborate and energetic signals that trigger responses from increasingly sceptical receivers, while strategic concerns come into play in determining whether receivers will tolerate long-term manipulation (because the benefits of a particular response pattern outweigh the costs) or whether the signalling system will eventually break down. Both perspectives are needed to understand natural signalling systems, as Johnstone

(1997) makes clear. Johnstone cites work by Marchetti (1993) on plumage brightness in eight species of warblers of the genus *Phylloscopus*; across the eight species the plumage becomes brighter as the species-typical habitat becomes darker. This can only be explained by appeal to signal efficacy. In contrast, Johnstone describes work by Briskie, Naugler, and Leech (1994) on the link between the intensity of chicks' begging calls and the level of extra-pair paternity in passerine birds. Chicks that are less likely to be related to their ostensible father beg more loudly; this cannot be a result of selection for signal detectability and must be explained with reference to strategic concerns.

2.6.1 Sensory biases

An emphasis on the effects of the receiver's sensory system on constraining signal design characterizes the "sensory bias" or "sensory exploitation" paradigm of Ryan and Rand (1993; see also Ryan, 1990). The idea is supposed to have broad applicability, but, like the handicap principle, is most easily illustrated in terms of sexual signalling: some male sexual signals may have tapped into pre-existing biases in female sensory systems and response patterns. The female response behaviour existed before the male display, and was either selected for in another context, or was not the result of selection at all.

The behaviour of the water mite *Neumania papillator* will serve as an example (see Proctor, 1991). These animals are aquatic predators that are sensitive to the vibrations on the surface of water caused by the movement of their prey. In their courtship display, males use their legs to mimic the vibration patterns characteristic of prey; females will orient towards males and clutch at them as if seeking a meal. Thus, males appear to be exploiting a female behaviour pattern that has been selected for in the context of feeding. Comparative analysis (Proctor, 1993) indicates that the female responsiveness existed before the male display.

Ryan and Rand's own work (Ryan, 1985, 1988; Ryan & Rand, 1993) on the Túngara frog *Physalaemus pustulosus* shows that males can also exploit female biases that are not themselves the result of selection for any function. Male frogs produce a call with a characteristic descending whine, followed by a low-frequency "chuck" sound; females are more attracted to males that can produce these sounds. However, comparative analysis across related species in the same genus indicates that *Physalaemus coloradorum* females *also* have a preference for calls with chucks, even though *P. coloradorum* males do not produce them. Ryan and Rand surmise that the preference for the chuck evolved before the chuck itself.

How might such female preferences have come about, if not through selection? Arak and Enquist (1993; see also Krakauer and Johnstone, 1995) suggest that female preferences or response thresholds for signals of a particular type or intensity are often implemented in such a way that hidden preferences exist for stimuli outside the normal range. For example, if females typically encounter males with tails between 10 and 20 centimetres long, and have evolved a preference for males with tails of at least 15cm in length, the preference might actually be implemented in the neural hardware as a cutoff threshold at 15cm and a linear increase in strength of preference above that point. This would mean that if a male with a 30cm tail was ever to appear, females would exhibit an extreme preference for this individual, whether or not he actually represented a good mate choice. The same female preference for tails of 15cm and longer might be constructed in any number of ways, with differing implications for unusual cases—the point is that as long as extreme males are rare, then selection is blind to the merely hypothetical fitness consequences of accepting or rejecting them.[13] Once a particular female preference has been established, later generations of males may develop novel qualities that allow them to exploit the preference.

The sensory bias aspect of receiver psychology complements Dawkins and Krebs's view of signalling as manipulation, because it offers some more detail about just what sorts of receiver preferences might be there initially for signallers to manipulate. It also addresses the issue of why these preferences might prove durable even when exploited: either because they are valuable in other contexts (as Dawkins and Krebs suggested) or because the receiver's cognitive circuitry is wired up in a particular way, and evolutionary change away from the old design is complicated and time-consuming in terms of the number and scope of the mutations required.

2.6.2 Information requirements of receivers

Another aspect of receiver psychology that has received little attention is the question of just what the receiver is seeking to get out of a signalling interaction. Bullock (1998) criticizes the idea that the receiver will be interested in accurate information for its own sake. While few theorists have claimed that receivers are motivated to collect truthful information *per se*, Bullock points out that when theories like the handicap principle are modelled in mathematical or simulation form, the fitness of receivers (typically females in a sexual signalling paradigm) is usually modelled as their accuracy in estimating the underlying quality of signallers (i.e., males). However, real females in a sexual signalling context are not interested in accurately determining the genetic quality of every male they meet, but in successfully mating with a male of high quality without wasting too much time on the process of search (see also Todd & Miller, 1995; Miller & Todd, 1998). Similarly, receivers in an alarm call system are not selected for accuracy in determining whether or not a predator is really approaching; on the contrary, they are likely to have a substantial tolerance for false alarms as the benefits of evading a predator will probably outweigh the costs of a number of unnecessary flight responses. It is important to recognize that for receivers in evolved signalling systems, survival and reproduction are more important than truth and accuracy.

2.7 Communication and sexual selection

Sexual selection, the process in which members of one sex experience differential reproductive success due to the mating preferences of the opposite sex, has already been covered to some extent. We have discussed Darwin's original development of the theory (section 2.1.2), and several ideas on signalling have been presented in terms of sexual selection. However, some minor points remain to be made.

Firstly, the fact that sexual selection and communication exist as distinct topics in the biological literature sometimes leads to the same idea being presented in parallel forms.[14] This is especially so in the case of the handicap principle (although at the same time models of the handicap principle have provided an important route for crossover between the two areas). It is instructive

[13] This idea is the likely explanation for the phenomenon of "supernormal stimuli" (Tinbergen, 1953) in which, for example, birds will preferentially brood an artificial egg that is of the appropriate colour and pattern for their species but is much larger than any normal egg.

[14] See Andersson (1994) for an excellent and comprehensive review of the literature on sexual selection.

to consider the role of the handicap principle in the sexual selection literature: specifically, that it is only one among several possible selective pressures leading to the evolution of female choice.

Alternative explanations for female mate choice include the possibility that there are direct phenotypic benefits for females in choosing particular males; for instance, in bush crickets the male presents the female with an edible spermatophore (Wedell, 1994), and thus there is a direct selection pressure on females to choose the male who bears the largest and most nutritious gift. However, the central debate in the sexual selection literature has been about whether female choice exists because of runaway sexual selection or because females are choosing males with good genes (i.e., because the male advertisement trait functions as a signal of underlying genetic quality).

Runaway sexual selection refers to a process first described by Fisher (1930), in which the genes determining a female preference for some male trait become linked with the genes for the trait itself. That is, females initially prefer males with tails that are slightly longer than average (for whatever reason). These females will mate with their preferred males, and their offspring will thus inherit both the genes for the father's long tail, and the genes for the mother's preference for long tails. The process will continue in a spiral of increasingly exaggerated traits and increasingly strong preferences—thus "runaway" selection—until the male trait becomes so exaggerated that it is deleterious for survival. The effects of natural selection then cancel out those of sexual selection: the trait and the preference stabilize, but remain extreme.

When the Fisher process brings about strong female preferences, the male trait cannot reasonably be described as a signal because it does not represent or indicate anything. Fisher's theory is sometimes described as the "sexy son" hypothesis: the function of a female's preference is simply to ensure that her male offspring gain from their father some of the genes necessary for an attractive advertisement trait. Both sexes are thus caught in a vicious circle: because of the prevailing fashion for long tails, short-tailed males will have low reproductive fitness; females with a preference for short tails will also lose out because their male children will not be fit.

The good-genes hypothesis is usually presented as the major alternative to Fisher's theory; this is the idea that male advertisement traits carry information about genetic quality, and that female preference evolves in order to exploit this information. The good-genes hypothesis looks a lot more like communication, and indeed such processes are referred to in the literature as indicator mechanisms (Andersson, 1994). From this perspective we can see that the handicap principle is a way of explaining why the signalling system implicit in male advertisement and female preference remains an honest one: namely, because it is stabilized through the cost of the advertisement trait.

However, an important problem for indicator-mechanism theories of female choice has to do with variation in the genetic quality of males. Such theories require some variance in male quality, because otherwise one male is as good as another and there is no incentive for females to use advertisement traits as indicators. At the same time, the theories suggest a situation in which the variance in male quality is constantly decreasing: if we imagine a lekking species[15] in which females with strong preferences choose the cream of the male population to mate with, then the next generation of males will all have similar (high) levels of quality. After several generations we

[15] A lekking species is one in which mating takes place at *leks*. A lek is a designated area where males perform displays and females assess them before deciding which male to mate with; in many lekking species (e.g., sage grouse *Centrocercus urophasianus*) breeding pairs have little or no contact outside the lek context, and it therefore becomes obvious that males are contributing only their genes to the project of raising offspring.

would expect the males to cluster very tightly around the optimum quality level, and again there would be no reason for female choice to be maintained as all males would now be similar. This is known as the paradox of the lek; recent attempts to solve it have invoked a negative mutation pressure on male quality that would maintain some variance despite the homogenizing effects of female choice (Iwasa et al., 1991; Pomiankowski & Møller, 1995).

The conflict in the sexual selection literature between runaway selection and good genes theories becomes relevant to the concerns of this thesis in chapter 9, where a simulation model of sexual signalling will be presented. In such a model, we might observe exaggerated male advertisement traits and a female preference for such males, and be tempted to say that communication was occurring. However, the Fisher process shows us that costly male advertisements and corresponding female preferences can co-evolve without necessarily having a communicative function, and thus runaway selection would stand as an alternative hypothesis that must be ruled out before we could conclude that our model exhibited handicap signalling.

2.7.1 The phenotypic gambit

Finally, there is a methodological caveat that the sexual selection literature makes clear. When building models of biological phenomena, we cannot always use the kind of simple game-theoretic model described in section 2.3.2, in which the fitness payoffs for various behavioural strategies, matched one against another, are considered. This is because sometimes the gene frequencies underlying the strategies are important—sexual selection being a case in point.

Grafen (1991) describes what he calls the “phenotypic gambit” of game-theoretic modelling. To accept the gambit is to assume that the complications of genetics can be safely ignored when modelling a particular behaviour. For instance, even though the behavioural strategies of foraging birds may be influenced by many genes in ways that we do not yet understand, we are nevertheless confident that a game-theoretic model provides a good framework for predicting and understanding the behaviour. This is because we assume that there are no genetic “dead ends” on the path to the optimal strategy; we assume that there always exists a series of possible mutations that would take the population from here to there (see Hammerstein, 1998). We therefore save ourselves the trouble of analyzing complete genetic models and rely on simpler constructs such as ESSs. Grafen argues that the gambit is popular because it is usually successful.

However, in the case of sexual selection, the genetic details are important and the gambit cannot succeed. For example, consider the paradox of the lek as described above. If we imagine constructing a game-theoretic model of the problem, it is easy to see that an important element of the male strategy involves specifying the individual’s advertisement level, perhaps as a function of his underlying quality. However, where does quality come from? One certainly cannot make the male’s quality level part of his strategy, because a banal strategy of “be optimal in quality” will prevail. On the other hand, a male’s genetic quality is not random, but depends on who his parents were. It turns out that the only way to capture this fact is to construct a population-genetic model instead, which tracks the changing frequencies of the genes encoding such traits as male quality, male advertisement and female preference.

In fact some models of handicap signalling in the context of sexual selection do not incorporate genetics, but remain pitched at the simpler game-theoretic level. Grafen’s (1990a) model is a notable example. However, there is no such thing as a free lunch, and such models pay for their simplicity by having to treat male quality as a random variable. This means that a male's suitability as a mate can be thought of as being environmentally determined rather than a genetically inherited character. If we are interested in the hypothesis that signalling systems can evolve in which males communicate their genetic quality to females (and in chapter 9 we will be) then we must employ population-genetic models.

Chapter 3

Conceptual issues in the study of communication

The thesis is concerned with the function of animal communication systems. However, throughout the previous chapter, the concepts of "function" and "communication" have both been taken for granted. Relying on the everyday meanings of these words has served us well in introducing the key ideas on the evolution of communication, but in fact both of these terms are the subject of debate in biology and in a wider philosophical context. Their use in ordinary language is imprecise. This chapter explores some of the issues around both concepts, especially communication, and defends a position on each. The work of Millikan (1984, 1993) will be integral to the argument. Millikan's ideas on evolved functions and on representation have not yet penetrated mainstream biology, but various authors (Dennett, 1987; Bekoff & Allen, 1992; Allen & Bekoff, 1997; Bullock, 1997a) have pointed out their relevance and utility.

3.1 The concept of function in biology

To ask about the function of something is to ask about its purpose or point. Why is it here? What is it for? These sorts of questions are no longer asked in many sciences—few chemists would claim that nitrogen has a purpose, or astronomers that supernovae occur for a reason as opposed to being merely physically caused. However, in biology this sort of *teleological* explanation, in which a phenomenon is explained in terms of its function or goal, has persisted. As Paley (1802) argued, living things appear to have been designed for certain purposes. Animals, in particular, behave as though they were pursuing specific goals. Explanations in terms of function or goal-directedness have simply proved too useful to be abandoned.

How, though, might we isolate and identify the function of a particular biological phenomenon? For instance, we believe that the function of the mammalian heart is to pump blood, but what grounds do we have for this belief? One way of pursuing a functional explanation is to look at the current causal properties of the phenomenon—at the way it fits into a network of causes and effects (Cummins, 1994; Wright, 1994). In the case of the heart, we can amass evidence such as: when the heart stops, the blood stops flowing and the animal dies; signals from the brain cause the heart to beat at different rates depending on the demand for oxygen in muscle tissues; if the heart can be replaced with some other kind of pumping device, blood flow can be maintained and an animal can be kept alive. By thus exploring the heartbeat's role as cause and as effect, we can arrive at the conclusion that the function of the heart is to pump blood (and that the function of blood is in turn to supply oxygen to the muscles, etc.).

This position on function is the one espoused by functionalist philosophers of mind, e.g., the early Putnam (1960), and Fodor (1968). Functionalism in philosophy of mind is the view that the mental states of an animal are its functional states, i.e., that mental states are to be identified by their roles as causes and effects in relation to sensory input, other mental states, and output in the form of bodily action. This leads to what is known as the "strong claim" of artificial intelligence: that mental states do not have to be realized by the particular hardware of neurons and synapses, but could in principle be implemented by a computer that was wired up in a functionally equivalent way.

It follows that under this view, two things A and B have the same function if A can be exchanged for B without disturbing the rest of the system that A is embedded in. A carefully crafted electronic device might have the same function as a neuron, because it could be switched for the latter without upsetting the running of a brain. But similarly, any ball of rock of the right size and density could function as the Earth's moon, because it could in theory be exchanged for the original without upsetting the Earth's orbit around the sun, the cycles of the tides, the location of the Lagrange points[1], etc. Somehow this seems an unwelcome conclusion: to say that something else could have the same function as the moon implies that the moon *has* a function. A scientifically minded person does not want to make such a claim; presumably the moon is just *there*, the result of certain natural processes. It has causes and effects but it surely has no function. Millikan (1984, 1993) has argued that the causal-role view of function is really an exposition of how an object or system can *function as* something, and that there is another, deeper notion of function that is of greater relevance to biology.

The alternative view of function, and the one that will be adopted in this thesis, is close to the common-sense meaning of the word "purpose". Millikan has argued that, given a physicalist view of the universe, the only process that can give rise to something like purpose is natural selection. Millikan claims that the function of a biological phenomenon is determined not by looking at its place in a causal network in the here and now, but by examining its evolutionary history. Specifically, the purpose of a trait, or, to use Millikan's terminology, its *proper function,* is to do that which gave a fitness advantage to ancestral holders of the trait. In other words, the proper function of a trait is to do whatever it has done in the past that has led to its being here today. For example, let us suppose that a tendency to run from any sudden movement exists in a species of herbivore, and that this tendency leads, over many generations, to the differential survival of those who possess it, because they are more likely to escape attacks by predators. The proper function of this tendency is therefore to assist the animal in evading predators.

Millikan carefully distinguishes between the concept of proper function and the causal-role view of function. Continuing the example, it may be that the herbivores have an extremely low threshold for triggering their flight response. They flee from innocuous movements such as wind-blown vegetation, and thus the vast majority of their flight responses occur when there is no predator near. Nevertheless, the proper function of the response is still predator avoidance. If the

[1] The Lagrange points are basins of attraction in the gravitational field of the Earth-moon-sun system.

predatory species suddenly dies out, and thus the herbivores are no longer subject to any predatory attacks, the proper function of the flight response is *still* predator avoidance, even though the response never coincides with the approach of a predator any more. A behaviour need not necessarily fulfil its proper function every time it is performed or even on average. Millikan's point is that the function of a behaviour (or any other trait) is determined by its evolutionary history—specifically, by the historically normal circumstances in which it has proved advantageous to its bearers, and not by the causal network in which it is currently embedded.

The force of this point can be brought out in several ways. To use a communication example, imagine that we have observed that whenever one animal makes a certain "signal", another reliably performs a behaviour in response. Millikan would argue that this observation alone cannot justify the ascription of a communicative function to the signalling or the response behaviour—it would be necessary to demonstrate that each behaviour had actually been selected for in the past.[2] In a more extreme case for Millikan's theory, proponents of the causal-role view of function have conjured up a philosophical fiction known as the swamp creature: this is a being that has, by miraculous coincidence, suddenly come into existence with the same structure and capacities as, say, a person. The causal-role theorists suggest that the instinctive reactions that this creature would exhibit, such as an eye-blink response when an object moves rapidly towards the face, surely have the same function as in a normal human, because the structure of the creature's body and brain is exactly the same. Millikan takes a hard line and says that the various capacities and responses of the swamp creature in fact have *no* function, because the creature has no evolutionary history.

If we consider some probable events in the evolution of signalling systems, it will be clear that Millikan's position suggests, sensibly enough, that the earliest mutant behaviours that provide the seeds for later ritualization, or elaboration via a signalling arms race, do not in themselves have a proper function. For instance, take Krebs and Dawkins's (1984) example in which dogs have evolved a mind-reading ability to infer aggressive intent from the bared teeth of an opponent. The very first dog that manipulates this arrangement to its own advantage, by baring its teeth when it has no intention of fighting, is simply a lucky mutant in Millikan's view. The teeth-baring behaviour does not take on the function of "threat display" until there exists a history of selection for that purpose. Similarly, suppose that ancestral vervet monkeys made some noise when danger was perceived, perhaps the equivalent of a human gasp of alarm, but did not yet have the ability to respond to the audible evidence that a fellow monkey had perceived danger. The first mutant that comes along with the tendency to flee whenever a gasp is heard may well live to a ripe old age, but the flight response does not have a communicative function until it has been selected for over time. Other philosophers of biology, notably Sober (1993), have expressed similar opinions on this point.

Millikan is aware of the possible objection that the evolutionary history of a trait, particularly a behavioural trait, can be difficult or impossible to determine with any certainty. She freely admits that in many practical circumstances, the proper function of something will be ambiguous.

[2] Strictly speaking, Millikan does not claim that natural selection is the *only* possible way that a behaviour can acquire a real function. She also allows the possibility of "derived proper functions": this could include a signal or response behaviour that had been learned rather than genetically inherited, as long as the learning system itself had a proper function. However, this aspect of Millikan's theory will not be explored here as the thesis stops short of looking at learned communication systems.

For instance, is the proper function of human language to promote social cohesion, or to provide a basis for rational thought? Is it a mixture of the two, or something else entirely? Barring time travel, it is difficult to see how the evolutionary history of *Homo sapiens* could be known in enough detail for the question to be settled. Millikan would say in this case that orthodox techniques, such as theoretical modelling, archaeological evidence, comparative evidence, etc., are the appropriate tools for trying to narrow down the possible proper functions of language. Her claim is an ontological one: that proper functions *exist*, by virtue of a history of selection.[3] The epistemological problem—finding out just what they are in particular cases—is left as part of the detective work of research.

Millikan's notion of proper function will be adopted in this thesis: whenever the function of a behaviour is referred to, that should be taken as shorthand for its proper function in the sense outlined above. The proper function idea is consistent with functional explanation in Tinbergen's sense (see sections 1.1 and 1.2) and indeed sceptics might say that it was no advance on Darwin. However, Millikan has elaborated the concept with more philosophical care than either Darwin or Tinbergen. Her work licenses the use of historically determined function as an explanatory construct that does not threaten to collapse into the mechanistic notion of function-as-causal-role.

3.2 Communication and related concepts

Many different phenomena are subsumed under the term communication. In standard treatments (e.g., Lewis & Gower, 1980) animal communication is taken to include aggregational signals, alarm signals, food signals, territorial and aggressive signals, appeasement signals, courtship and mating signals, and signalling between parents and offspring—the list is not exhaustive. Borderline cases such as deception, mimicry, camouflage, and imitative behaviour may count as communication depending on the author. Communication can also be said to occur between cells in the body, e.g., across synapses, and even inside the cell: witness "messenger RNA". Finally, we use the term for such uniquely human phenomena as language use, and information transmission between artefacts such as computers.

This multitude of forms invites conceptual exploration. Do these varieties of communication have anything fundamental in common? If so, what is it? If not, what is the most practical way in which the many varieties listed above can be grouped and categorized—how can we cut nature at its joints? To paraphrase Millikan, the term "communication" does not come from scripture. If the concept turns out to be too broad or vague to be useful, then we will need to establish narrower, more specific concepts in order to continue with our inquiry.

3.2.1 Some problem cases

In the following pages we will examine several different views on how communication should be defined, and in section 3.3 a particular definition will be defended as the most appropriate one for the purposes of the thesis. Before proceeding, it will be useful to review some problematic cases that may or may not qualify as communication, depending on the definition adopted. The reader

[3] It follows that a trait can have no proper function because it has never been selected for; it may be a "spandrel" *sensu* Gould and Lewontin (1979). It is also possible that there is no single proper function associated with a trait, as it has been selected for on multiple grounds.

is invited to exercise his or her own intuition regarding each case. It is intended that these problem cases should highlight the difficulties involved in defining communication.

Incidental information transfer

Sometimes the observed behaviour of one animal can supply information to another, but the information transfer appears to be incidental; the behaviour of the first animal seems not to have a signalling function. For example, suppose that a scavenging bird notices some food on the ground and flies down to eat it. Other birds of the same species see that the first one has landed: they "assume" that it must have found food and they also fly down in order to obtain some for themselves.[4] Are we to regard the first bird as sending a signal, through the act of flying down in order to eat, about the presence of food? If this is not communication, then what if the species evolves prominent red markings so that conspecifics can spot each other easily from above? What if the bird makes some kind of call upon flying down to feed? On the other hand, what if the species evolves camouflage colours such that individuals are more *difficult* to spot from above? Which, if any, of these developments would justify calling the behaviour communicative?

Camouflage: the difference that makes no difference

Camouflage is often presented in biology textbooks as a variety of communication, because the "signaller", the cryptic animal, has been designed to affect the sensory system of the receiver in a particular way. (Similarly, mimics have also been designed by selection to influence receiver sensory systems.) Imagine that a moth is camouflaged against a tree trunk, and a predatory bird flies past. We can then assert that had the moth *not* been camouflaged, the bird would (probably) have seen and eaten it. On the other hand, had the moth not been there at all, then the bird would have acted just as it did in the presence of a camouflaged moth, i.e., it would have flown past the tree. Are we then to regard the moth as sending an "I am not here" signal? If we regard camouflage as communication, then perforce communication can include instances in which the appearance and behaviour of the sender have no influence whatsoever on the behaviour of the receiver. We are potentially drawn towards a *reductio ad absurdum* in claiming that moths hundreds of kilometres away are also sending "I am not here" signals to the bird. Nevertheless, the camouflaged moth does seem to be exploiting aspects of the bird's sensory system, and the discussion of both Dawkins and Krebs's ideas and the sensory bias paradigm in chapter 2 has shown that such exploitation is often a feature of signalling behaviour.

Deception

The human practice of telling lies represents occasional deception in a general context of truth-telling. However, many deceptive animal signals are not like this: the deceptive signaller frequently hijacks a certain responsiveness amongst a group of receivers to which it does not belong, and never sends a truthful signal. A typical example is the lure of the angler fish. This is always used to exploit the tendency of smaller fish to approach worm-like objects; obviously, it is never really a worm. Do we want to classify such deceptive animal signals as communicative? Consider a bird that flies erratically away from a predator such as a fox; the bird clearly has a broken wing

[4] Whether we mean that the animals really make an assumption here in the same way that a person would, or whether phrases like "the animal assumes. . . " or "the animal believes. . . " are only a kind of shorthand for describing behavioural dispositions, is precisely one of the difficult issues in thinking about animal communication. This aspect of the problem will be treated in section 3.2.4. For now, the scare quotes around such terms will be omitted for convenience.

and will be easy for the fox to catch. Thus the fox is gaining incidental information about the bird as discussed above; Dawkins and Krebs would say that the fox was a good mind-reader. Is this communication? Intuitively it seems odd to describe the erratic flight of the bird as a *signal*. But what if the apparent broken wing is being faked in order to lead the fox away from the nest? Because this is deceptive, it somehow looks closer to being communicative. However, if we accept this intuition, we are left with the strange result that telling the truth—flying erratically with a wing that really is broken—is not communication, but lying is.

Co-ordination without signalling

Sometimes social animals co-ordinate their behaviour so closely that communication is suggested, but this need not imply that signals are in fact being exchanged. For example, imagine that two ants co-operate in caring for larvae in the nest. One ant places each larva in turn into a brood chamber, and the other caps the chamber with a ball of mud. The behaviour of the two ants is well co-ordinated, and might seem to require communication between them. However, suppose that upon investigation it turns out that the two ants are simply following independent behavioural programs, and no signals of any kind pass between them. In an extreme case, let us suppose that they cannot even perceive each other: the second ant simply places its ball of mud whenever it finds an uncapped chamber with a larva inside. Should we regard the first ant's placement of larvae into chambers as stigmergic[5] communication, the message in each case being "please cap this chamber now"? Or should we accept that communication is not a necessary condition for co-ordinated behaviour?

Direct causation of responses

Communication is often defined in terms of one animal's actions affecting the behaviour of another (see section 3.2.2). However, there are many circumstances in which the behaviour of one animal affects the behaviour of another but which do not appear to fit the intuitive notion of communication. For example, if a cheetah successfully stalks a gazelle and then seizes it by the throat, this will certainly affect the behaviour of the gazelle: it will thrash about and eventually die. Presumably no-one wants to call the cheetah's attack a signal, however. Similarly, if A tells B to jump into the lake, it is communication; but if A physically *pushes* B into the lake, it is not (Cullen, 1972). The intuitive concept of communication seems to include the proviso that a signal is something *perceived* by the receiver and then acted upon, rather than being something that directly causes the receiver's response. There are potential problems in formalizing this intuition, however. If we believe in a physical, mechanistic universe, then ultimately *all* signals cause their responses. Is there a principled way to distinguish between responses caused by the perception of signals, and responses caused in some other more direct way?

3.2.2 Attempts to define animal communication in terms of behaviour

The problem cases discussed above should make us circumspect about our chances of arriving at a neat and simple definition of communication. Nevertheless, at first glance communication appears so straightforward as to be hardly worth defining. The naïve definition is easy: to communicate is

[5] Stigmergy is simply communication through manipulation of the environment. Of course, in a sense *all* communication must involve the manipulation of the environment, but the term is meant to refer to communication through relatively permanent changes such as (for ants) the rearrangement of soil or the depositing of pheromones.

to transmit information; to *tell someone something*. This is the conduit metaphor for communication, described by Reddy (1979) and Lakoff and Johnson (1980), in which information is passed from one agent to another via a signalling channel. The speaker or signaller might be lying, and the listener or receiver might have already known what the speaker tells them, but these look like marginal cases.

As we have seen in chapter 2, the naïve definition has not been satisfactory for most evolutionary biologists. Krebs and Dawkins (1984) defined a "signal" as an action or structure which increases the fitness of an individual by altering the behaviour of other organisms. Similarly, Burghardt (1970) defined "communication behaviour" as a behaviour on the part of a signaller that is likely to influence the receiver in a way that benefits, in a probabilistic manner, the signaller or some group of which it is a member (see also Wiley, 1983; Endler, 1993; Johnstone, 1997). Some biologists have been more restrictive: for example, Lewis and Gower (1980) allowed that a behaviour could qualify as communication only if one animal influences another in some way and that both signaller and receiver on average benefit from the exchange. Tinbergen (1964), as noted earlier, suggested that as long as the interaction tends to benefit the *species*, it is communicative.

Why have these biologists seen fit to define communication in terms of actions that alter the behaviour of others and thus benefit the signaller (and possibly the receiver as well) rather than in terms of one animal transmitting information to another? One important reason is that such definitions are parsimonious: they are behaviourist in spirit and thus avoid the troublesome issues of imputing mental states to animals and ascribing specific content to signals. The definitions allow an investigator, in principle, to observe certain behavioural regularities and label them as communication without ever having to assert such contentious claims as "after observing A's signal, B now believes that there is a predator approaching", or "A is trying to reduce B's uncertainty about A's suitability as a mate", or "this signal *means* 'feed me'." Although Krebs and Dawkins's definition, for instance, can be justified because it is consistent with their broader argument that animal communication is about manipulation rather than information transfer, the point remains that their definition focuses solely on observable behaviour and does not require the ascription of anything like beliefs and desires to communicating animals.

This reluctance among biologists to deal with the mental lives of animals should not be regarded as mere squeamishness. As Einstein reputedly said, a theory should be as simple as possible, but no simpler. If a successful theory of animal communication can be constructed without a commitment being made regarding mental states and so on, then so much the better. But are these straightforward definitions of communication adequate? Do they capture all the phenomena that we intuitively think of as communication, and exclude the phenomena that we do not? We will see below that some theorists argue that theories of animal communication must at least involve talk about information (section 3.2.3), while others have gone further and suggested that a commitment to viewing animals as rational agents is necessary (section 3.2.4).

If we consider Krebs and Dawkins's definition—that a signal must influence the behaviour of other organisms in a manner that benefits the signaller—it does seem to capture at least some typical animal signals. Male sexual advertisements, for instance, often influence the behaviour of female observers in a manner that benefits the signalling male. Alarm-calling qualifies as a signal, although only if the caller is closely related to the receiver of the call and thus gains a fitness benefit from the latter's response (i.e., flight from the predator) due to kin selection. An odd implication of Krebs and Dawkins's definition, therefore, is that if we observe alarm-calling behaviour but cannot establish any inclusive-fitness benefits to the signaller, then the behaviour is not signalling.

The definition also casts a very broad net. As noted above under "Direct causation of responses", there are many occasions when the behaviour of one animal influences the behaviour of another to the advantage of the first—thus qualifying as communication under Krebs and Dawkins's definition—and yet the result does not look like communication because the first animal has in a sense "directly caused" the behaviour of the second. Furthermore, as discussed under "Incidental information transfer", it may be the case that a certain behaviour fits the definition, and yet does not appear to be a signal. Returning to the example given earlier, it might well be to the advantage of bird A that bird B should notice that it (A) has found food, perhaps because the two are related. However, that would mean that A's behaviour in simply observing food and approaching it counts as a signal; this is certainly a counter-intuitive result.

Hasson (1994) pointed out the latter problem and tried to improve the definition offered by Dawkins and Krebs with the additional proviso that signals must reduce fitness in contexts other than interactions with other organisms. This is another way of saying that there must be some cost involved in their production. Hasson's proviso would exclude "flying down to obtain food" from being a signal, because it is a behaviour that is useful in a variety of contexts. Specifically, it is useful even when there are no conspecifics around to observe it. Maynard Smith and Harper (1995), in turn, point out a problem with Hasson's correction:

> For example, merely being large may alter the behaviour of opponents in contests, and may well be costly in other contexts, but we would not wish to classify large size as a signal…There seems no alternative, therefore, to including in the definition the notion that a signal has features specifically adapted to alter the behaviour of others.

Maynard Smith and Harper assert that signals are behaviours or structures that not only influence other animals, but whose *purpose* is to do so. An appeal to function is needed in order to back up our intuitions about what should and should not count as a signal. This conclusion echoes Millikan, who would say that a behaviour is only a signal if it can be shown that signalling is its proper function. However, if we need to make reference to function in defining communication, then it cannot be defined solely in terms of behaviour observable in the here and now. The evolutionary history of the trait, at least, must also be taken into account. Therefore we must conclude that attempts to define animal communication in purely behavioural terms (e.g., Krebs & Dawkins, 1984) cannot succeed.

3.2.3 The role of information

If behavioural definitions of communication will not work, then perhaps a return to the naïve definition and the notion of information transfer will be more successful. Hurd (1997a) provides a good example of communication defined in informational terms:

> Information is said to be received whenever an agent changes it's [sic] expectations about the consequences of an action, and communication has occurred whenever the action of one animal transmits information to another.

One problem is immediately apparent with such a definition: it is still over-inclusive, because incidental information transmission qualifies as communication. As with behavioural definitions, a bird's action in noticing food and approaching it can count as a signal, because other birds may gain information (in Hurd's sense) by observing the behaviour.

Nevertheless, the informational definition comes close to pinning down our intuitions about communication, and there is a temptation to narrow it down by simply excluding the incidental transmission of information. Such a move would be ill-advised, however, as the philosophical problems of the informational approach run deeper. Ever since Shannon and Weaver (1949) developed the mathematical theory of information, biologists have tried to apply the concept to animal communication without much success.

One of the main difficulties has been in determining what the units of information transmission are in natural signalling systems. Shannon and Weaver's theory concerned maximally efficient information transmission using a pre-arranged code, but real animal communication appears to be highly redundant, and cannot easily be dissected into semantic atoms. Shannon and Weaver's theory suggests that information can be measured in bits, where a bit represents a twofold reduction in the receiver's level of uncertainty. This idea does not transfer easily to real animal communication. For example, it would require millions of bits of stored information to digitally record the roar of a red deer stag; no-one believes that this is a measure of how much information the roar transmits to the stag's opponent. In theory, the informational value of the roar for the receiver can be determined through experiment: observing the different responses that receivers tend to make to roars of different intensity, and so on. However, in practice this has proved difficult. Even in the more accessible case of human language, measurements of information do not square up with common-sense ideas about how much real information there is in a message: the complete works of Shakespeare, for instance, require fewer bits of Shannon-information to describe than a nonsense text made up of the same words arranged in a random order. Both messages would hold little informational value for someone who spoke only Chinese. As Dennett (1987) notes, we seem to grasping for a theory of *semantic* information—of what a signal or message means to a particular observer in a particular context—that must be distinct from the formal mathematical notion.

Thus, a second problem with the informational approach is to specify the content or meaning of a signal. If information is to be used as a theoretical construct, then we should surely be able to measure not only the *amount* of information, as Shannon and Weaver's theory purports to allow us to do, but we should be able to specify just *what* information a signal or message is carrying, i.e., to specify its content. This assertion is based on Quine's principle of "no entity without identity": if we have no procedure for specifying whether or not two signals are identical, i.e., whether or not they carry the *same* information, then it is unprincipled to introduce "information" into our ontology. However, trying to ascribe content to signals brings us up against a problem that is well known in theories of mental representation: the problem of error.

It should be noted that theories of mental representation bear (or perhaps *should* bear) a close relationship to theories of communication: roughly speaking, the former are about how a single agent might store facts about the world; the latter are concerned with how one agent might make another aware of a fact about the world. In terms of mental representation, the problem of error is as follows. Let us suppose that brain state C is the mental representation "there is a cat". That is, on seeing a cat, a person typically goes into brain state C. One dark night our experimental subject sees a skunk and mistakes it for a cat, and therefore goes into brain state C. Must C then be redefined as the representation of "there is a cat-or-skunk"? If so, then C starts not to look like the representation of "there is a cat" at all, because the list of what C really represents must be broadened to include all possible cases that could possibly give rise to the mistaken belief that one was looking at a cat. Fodor (1987, 1990) and Dretske (1981) have both wrestled with the problem of error; it arises for any theory of mental representation that says the content of a representation is determined by its causal relationship with the world, i.e., that what causes or triggers a representation gives it its meaning.

The problem of error applies in exactly the same form to the domain of communication. Suppose we have a working hypothesis that a certain bark is used by vervet monkeys to represent the approach of a leopard, i.e., that the informational content of the bark is, for vervets, "leopard approaching!". One day we observe the bark being made when a hyena approaches. We are then faced with a problem of indeterminate content. It does not make much sense to say that the bark means "leopard or hyena or yellow Land Rover or Robert Seyfarth with a speaker in the bushes or something else approaching". On the other hand, we cannot evade the problem by stipulating that the bark *means* leopard and that all other uses of it are mistakes, because that would be begging the question about its content.

Thus, although definitions in terms of information transmission come close to satisfying our intuitive conception of communication, there are problems involved in specifying both the amount of information transmitted in a particular exchange, and the content of a given signal. We cannot rely on the idea of information transmission alone for a precise definition of communication.

3.2.4 Intentional communication?

The word "intentional" is used in two senses in philosophy. The first is in the everyday sense of a deliberate act, as in "he intentionally dropped the glass." The second sense of the word, introduced by Brentano, relates to the problem of meaning. Intentionality in this second sense is "aboutness", and refers to the fact that some things in the world—such as words, mental representations, road signs, and alarm calls—seem to be *about* other things in the world.

Information-based definitions of communication, and their associated problem of error, have already brought us face to face with some of the difficulties involved in naturalizing intentionality (the project of explaining intentionality or meaning within a scientific, physicalist world-view). As we have seen, taking an informational stance requires us to think of the participants in a communicative episode not as mere automatons, but as agents that can execute different behaviours based on information that they have received from other agents or from the environment. Perhaps light would be shed on communication if we were to go one step further and adopt the intentional stance (Dennett, 1987), which would mean viewing communicating animals as intentional beings; as rational agents that have beliefs and desires and act accordingly. The move is a tempting one, as the language used in the conduit metaphor for communication, and even Shannon and Weaver's dry mathematical treatment of information transmission, implies intentionality in both senses of the word: the sender *intends* to communicate some fact about the world via a message that *means* something.

Grice (1969) put forward the case (later developed by Bennett, 1976) that considering intentionality allows us to pick out a special kind of communication that is genuinely worthy of the name. Grice and Bennett rely on the intuition that there is a difference worth marking between a situation in which causal automatons exchange signals, and a communication system in which participants really mean what they say. Their argument is that real communication can be roughly equated with human speech acts, and must involve, at a minimum, something called third-order intentionality.

To have first-order intentionality is to be a basic intentional system, i.e., to have beliefs and desires concerning the world, such as "I *believe* there is a predator nearby", or "I *want* to mate with this animal", but not to have any beliefs or desires that are themselves *about* beliefs or desires. Second-order intentionality is to have beliefs (or desires) that can be about beliefs (or desires), such as "I *want* this animal to *believe* that there is a predator nearby." Finally, third-order intentionality means being able to hold beliefs about beliefs about beliefs (and desires about desires about desires, etc.). Thus, we come to Grice's formulation for a true speech act: that the speaker *intends* the hearer to *recognize* that the speaker *wants* the hearer to produce a particular response. For instance, if one person asks another to "please pass the salt", then although the speaker wants the salt, she does not intend to exploit some salt-passing reflex in the listener, but rather that the listener should come to believe that the speaker wants the salt and therefore pass it to her. Grice and Bennett argue that this sophisticated form of communication is what distinguishes true language from simple signalling systems.

Dennett (1987) suggests that we might be able to use this Gricean view of differing orders of intentionality to identify real communication. Dennett also believes that only third-order intentional systems (or better) can really communicate. He gives an example of second-order intentionality that fails to qualify: "I *want* you to *believe* I am not in my office; so I sit very quietly and don't answer your knock. That is not communicating."

Dennett's intentional stance means assuming that the agents of interest (e.g., animals) are rational, and trying to predict their behaviour from that assumption combined with educated guesses as to their beliefs and desires. This approach would allow us to test hypotheses about the order of intentionality involved in an apparently communicative system. For example, there may be some debate as to whether vervet monkey alarm calls exhibit first- or second-order intentionality. If the former, then a calling vervet *wants* its hearers to run for the safety of the trees, for example. If second-order intentionality is involved, then the caller may *want* its hearers to *believe* that there is a leopard approaching. Dennett suggests that careful experimental work could distinguish between these two hypotheses. Note that the second-order hypothesis, for instance, implies that the vervets have some conception that other agents in their environment can have beliefs. If the monkeys never exhibit this ability—perhaps their attempts at "deception" are always completely unsophisticated, indicating a failure to appreciate that other monkeys can see for themselves that things are not as the would-be deceiver would have them—then we must fall back on the first-order hypothesis to explain their behaviour.

Standing beneath even this, argues Dennett, is the "killjoy" null hypothesis of zero-order intentionality. This is the prospect that the monkeys do not even have first-order beliefs, but behave in accordance with simple tropisms. In the presence of leopards, they experience "leopard anxiety" and instinctively make a certain sound; those who hear this sound experience an equally blind reflex compelling them to make for the trees.

Dennett's program for an empirical approach to intentionality leads us to two problems with intentional definitions of communication. Firstly, even if it could be demonstrated that an apparently communicative system involved mere tropisms, we may well still want to classify the system as communicative. If, in a very simple organism such as a protozoon, the behaviour of one animal regularly but blindly causes a particular response in another, and both the signal and the response behaviour appeared to have been selected *qua* signal and response, then to deny that the system counts as communication seems churlish. While there may well be some substance to Grice's and Bennett's arguments that third-order intentionality is one of the things that makes human language different from what most other animals do, that does not mean that we want to restrict the application of the term "communication" to such high-level systems only.

The second problem has to do with the assumption of rationality in Dennett's intentional stance. Allen and Bekoff (1997) compare Dennett's and Millikan's notions of intentionality or meaning in natural systems, and remind us that Dennett's intentional stance is supposed to be effective to the degree that the organism being studied conforms to an idealized notion of rationality. The animal under investigation is supposed to have certain beliefs and desires, and is predicted to behave in a manner consistent with the logical pursuit of those desires given those beliefs. For example, if a monkey *wants* food currently in the possession of another, and *believes* that the other would abandon the food if it thought there was imminent danger, we could predict that the first animal might try a false alarm call.

On the other hand, Millikan's ideas on intentionality (described below) appeal entirely to evolutionary history and make no assumptions about rationality. Millikan's position, according to Allen and Bekoff, allows us to recognize that "an animal may have very specific cognitive abilities with respect to particular intentional states of other organisms without having the general ability to attribute intentional states to those organisms." Thus it is entirely possible that an animal might behave in Machiavellian third-order ways but only in specific contexts, e.g., one monkey wanting another to believe that it thought it was unobserved in the context of some deceptive food-hiding scheme. Millikan sees no reason why this could not occur despite a complete failure on the part of the monkey to exhibit third-order intentionality in other situations: she argues that natural selection tends to produce cognitive capacities that match specific ecologically relevant tasks, rather than an all-encompassing reasoning ability. Therefore a definition of communication in terms of higher-order intentionality would be founded on the dubious premise that animals either unambiguously did or unambiguously did not possess such intentional capabilities.

3.2.5 Millikan's alternative classification scheme

We have seen that attempts to define communication solely in terms of observable behaviour, in terms of information transmission, or in terms of the intent to communicate, are unsatisfactory. How, then, should we proceed? The suggested way forward is to define communication in terms of pairs of behaviours—signal and response—that have as their proper functions the encoding and decoding respectively of some fact about the signaller's world (see section 3.3). Such a definition makes reference to behaviour and the evolutionary history thereof; information transmission is implicitly included. It has already been argued that communication and mental representation present parallel conceptual problems, and in order to flesh out this definition, it will first be necessary to review a scheme proposed by Millikan (1993) for classifying representational phenomena.

Rather than perpetuating a debate about what representations really are, Millikan is interested in "lay[ing] down some terms that cut between interestingly different possible phenomena so that we can discuss their relations." She outlines the categories of tacit supposition, intentional icon, inner representation and mental sentence. These categories group representation-like phenomena in order of increasing sophistication; only the first two will be relevant to the simple signalling systems modelled later in the thesis. As we shall see, Millikan's typology applies to communication without adjustment. In her theory, communication is simply the exchange of representations between organisms.

Tacit suppositions

Millikan starts by arguing that some of the phenomena we might want want to dignify with the label "representational" are not worthy of it. Specifically, she suggests that in the many instances in which organisms are adapted in such a way that they resemble or reflect some aspect of their environment, this is only a "tacit supposition" and not any kind of representation. Tacit suppositions occur when the design of an organism meshes so neatly with a feature of the environment that it is tempting to say the design "represents" that feature. For example, if a biological clock, i.e., a mechanism controlling circadian rhythms in an animal, produces a cycle close to 24 hours then we may be tempted to say that the clock mechanism somehow represents the length of the terrestrial day. Similarly, the patterning of the Viceroy butterfly (*Basilarchia archippus*) might be said to represent its model, the Monarch (*Danaus plexippus*).

Millikan refers to these adaptations as tacit suppositions because they presuppose certain facts about the environment in order that their proper function may be fulfilled. Biological clocks presuppose that the day is 24 hours long, while the body patterning of a Viceroy presupposes that there are Monarchs in the area and that predators are familiar with their unpleasant taste. These presuppositions can be false—think of a moth with dark-coloured camouflage that lands on a light-coloured tree—and that is why the temptation exists to think of them as representations. However, Millikan argues that they are best seen as assumptions about the environment in which the organism will typically find itself.

Intentional icons

Millikan proposes that for a system to qualify as minimally representational, it must involve more than tacit suppositions. Firstly, there must be something identifiable as the representation itself; an "icon". Furthermore, the icon must have a "producer" and a "consumer". It must be the proper function of the producer to generate the icon in accordance with a mapping rule that relates one or more dimensions of possible variance in the icon to variance in the environment. It must be the proper function of the consumer to use or be guided by the icon in some way. If all of these conditions are met, Millikan suggests that the system involves an "intentional icon". For example, the transmission of electrical energy between the visual and motor cortex of a frog, stimulated by the approach of a fly or similar object, and resulting in the frog's launching its tongue towards the insect, is an intentional icon. The icon itself is the pulse of electrical energy transmitted down a particular neuron or neurons—note that intentional icons are often temporary. The producer of the icon is the frog's visual system, which is constructed in such a way that different relative positions of the fly (left, right, straight ahead) produce different icons—this is the mapping rule. The consumer of the icon is the motor cortex: depending on the relative position picked out by the icon, the tongue will be launched in a different direction—in this way the consumer is guided by the icon. Harvey (1996) has outlined a very similar view: he maintains that "representation" is a four place predicate, of the form "P uses Q to represent R to S". In order to use the concept of representation in a principled way, we must specify who or what plays the roles of P, Q, R, and S. In this case the visual system uses the electrical pulse to represent the presence of the fly to the motor cortex.

There is no requirement in Millikan's definition that the consumer mechanism be within the same organism as the producer mechanism. When the two are in different organisms, we have communication. For example, the waggle dance of the honeybee is a paradigm case of an intentional icon: the dance itself is the icon, the dancing bee is the producer, and a mapping rule relates the angle and duration of the dance to the direction and distance to a food source. The watching bees are the consumers of the icon, because it is the proper function of the dance to guide them to the food source.

The definition of an intentional icon makes it easier to see why tacit suppositions do not qualify as representational. In the case of biological clocks, an observable 24-hour cycle is produced by some mechanism, but it is not an intentional icon as there is no consumer that is properly guided by it. In Harvey's terms, the system is not representing the length of the day *to* anyone. Looking at camouflage and mimicry, we might view the process of growth as the producer, and the cryptic morphology itself as the icon. However, there is generally no mapping rule which relates variance in the icon to variance in the environment: the same adult form is produced regardless of environmental variation. Even if a mapping rule existed, as it does in the chameleon with its constantly adaptive camouflage, there is no proper consumer for the icon: predator visual systems may fail to register the chameleon due to its camouflage, but it is not their proper function to do so.

Intentional icons can be very simple. Millikan gives the example of tail-slapping in beavers, an undifferentiated alarm call given to a range of possible threats. The icon, producer and consumers are in this case obvious. The mapping rule, however, is more subtle: Millikan suggests that the time and place of the behaviour are part of the signal. Thus, tail-slapping is a simple intentional icon that represents danger, and the mapping rule is "here and now".

Note that intentional icons can map to something that does not exist. For instance, beavers are apparently rather nervous creatures, and produce many tail-slap alarms when there is no real danger impending. A false alarm still means "danger", however: reliability is not part of the definition of an intentional icon. As long as the icon is produced and used in accordance with the proper functions (i.e., the historically determined functions) of the producer and consumer, then it remains an intentional icon. Thus the problems of the causal and the informational approaches to communication and representation are avoided. The icon's content is not determined by its cause or trigger, so the fact that a false alarm call was caused by some innocuous movement does not threaten to make it not-an-alarm-call after all. Similarly, the content is not determined by whether or not the call transmits veridical information, and so the fact that a false alarm may not actually transmit any information does not threaten its status as an alarm call.

3.3 A functional definition of communication

We are at last in a position to defend a definition of communication, based on Millikan's notion of proper function and her split between tacit suppositions and intentional icons. The definition involves specifying how "real communication" (or *proper signalling*, as we shall call it) differs from three other related behaviours. It will therefore be easier to begin with a broad view of animal behaviour.

Animals have evolved to do the right thing: during every moment that they are alive, they face the problem of what to do next. They confront this problem in a complex environment, and the single most complex aspect of that environment will typically be the behaviour of other animals. An animal must deal with the challenges posed by both conspecifics—such as offspring, rivals, and potential mates—and heterospecifics, such as predators, prey, and competitors.

An animal's decision[6] about what to do next must be based on the states of the world that it can perceive or be influenced by. Some of these states may be internal to the animal, e.g., feedback mechanisms that report hunger or exhaustion, or the retrieval of information from memory. Some may be associated with other animals, e.g., detecting the presence of a predator, or perceiving the size and condition of a potential mate. Finally, some of the informative states accessible to the animal may be aspects of the inorganic environment, such as being aware of a high wind, or seeing that there is a large rock nearby. An animal does not, of course, have access to all the informative states it might conceivably be interested in: the animal's evolved perceptual system will constrain the set of states that can make a difference to its subsequent behaviour. This is simply to restate the ideas of von Uexküll (1928), who pointed out that every animal lives in a particular perceptual world or *Umwelt*.

Similarly, the animal's behavioural choice in response to all this information is not completely open, but is limited by the kind of evolved body it has and how that body currently stands in relation to the environment. A rabbit cannot suddenly decide to turn and gore a pursuing fox, and disappearing down a rabbit-hole will only be an option if there is a suitable hole nearby.

In this context, clearly, animals can influence each other, but influence is not synonymous with communication. Let us use the term *influence interaction* to refer to an event where one animal acts in such a way as to influence the perceived states of the world, and thus alter the subsequent behavioural response, of a second animal. Note that the first animal's action is itself a response to *its* perceived states of the world, and that the action's effect on the second animal is mediated by the environment—the first animal does something *in the world* which the second animal then perceives. In any given influence interaction, we can ask whether the actions of the first and of the second animal are fulfilling their proper functions. The possible answers to these two questions—yes or no in each case—constitute four distinct situations.

Firstly, it may be the case that influencing the behaviour of the second animal is not the proper

[6] Talk about animal decisions, as with talk about animal beliefs and desires, is usually hedged with the proviso that such intentional language does not constitute a claim for *real* intentionality in the animal concerned. However, if we take Millikan's position seriously we must accept that a term like "decision" can be applied to the operation of any biological mechanism, however simple or however complicated, that has the proper function of guiding an organism towards the most appropriate of several alternative behaviours. This time, therefore, the scare quotes have been deliberately omitted.

function of the first animal's action, and nor is the second animal's response fulfilling its proper function. For example, the vibration and noise caused by a pig rooting for truffles might prompt a mole to flee because it believed that a predator was approaching. The proper function of the pig's behaviour is to uncover truffles; the proper function of the mole's behaviour is to help it evade predators. The fact that the pig has influenced the mole in this way is in accordance with neither of these two functions. We will therefore refer to such cases as examples of *accidental influence*.

Secondly, it is possible that the first animal's action is fulfilling its proper function, but the second animal's response is not. For instance, when the lure of an angler fish attracts a smaller fish, causing it to approach and be eaten, the lure display is fulfilling its proper function. The smaller fish's approach response is not—its proper function is to guide it towards its own prey. Krebs and Dawkins (1984) call this *manipulation*; we will adopt the same terminology here. Specifically, the first animal is manipulating a response that has evolved for some other purpose.

The third possibility is that it is *not* the proper function of the first animal's action to influence the behaviour of the second, but the second animal's response *is* performed in accordance with its proper function. An example will help to make this clear: if a cheetah picks out a particular antelope because it appears to be sick or injured, then we can say that the antelope's behaviour has influenced the cheetah. However, the antelope's behaviour—walking in such a way as to reveal its poor condition—cannot be in accord with any proper function.[7] The cheetah's reaction, on the other hand, is consistent with the proper function of guiding the animal towards easy prey. In Krebs and Dawkins's (1984) terms this would be described as mind-reading on the part of the cheetah. However, a second example suggests that *exploitation* is a more suitable term: if the wind changes when a cheetah is stalking a herd of antelope, and they catch her smell and flee, then the antelope have exploited natural information about the cheetah. It is not the proper function of the cheetah's sweat glands to produce smells that will scare off antelope, but the proper function of such a response in the antelope is surely to keep them out of danger.

Finally, the behaviour of each animal in an influence interaction may be fulfilling its proper function. This is most easily seen in cases where the outcome is mutually beneficial: the dance of a returning bee and the subsequent directed foraging behaviour of its hive-mates are both fulfilling their proper functions. When both the action and the response are performed in accordance with their proper functions, the aspect of the first animal's behaviour that influences the second qualifies as an intentional icon. (The first animal is the producer, the second the consumer, and the mapping rule is determined by whatever process brings about the first animal's behavioural choice in the light of its perceptual input.) We will refer to this class of interactions as *proper signalling*.[8]

For the purposes of the thesis, communication is defined as proper signalling. Some of the advantages and implications of this definition will be discussed below. It should first be noted that the definition is not entirely original: Bullock (1997a) defines "full-blooded signalling" in a similar fashion, and Oliphant (1997) seems to be getting at much the same idea when he says that true signalling is what happens when an interaction is simultaneously exploitative and manipulative.

[7] The proper function of a trait is to do whatever it was that conferred a fitness advantage on ancestral possessors of the trait. It is thus difficult to conceive of circumstances in which bringing about the immediate death of the organism could be the proper function of any trait.

[8] The term is intended as a nod towards Millikan's notion of proper function. Given the claim that proper signals qualify as intentional icons, there is a case for using the term "intentional signalling", but this was felt to be inviting linguistic confusion.

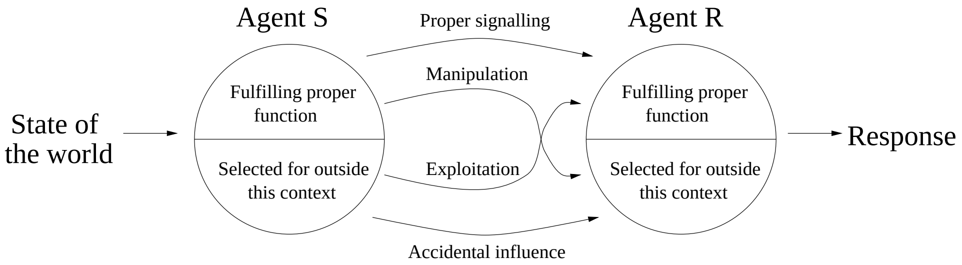
Agent S
Agent R
Proper signalling
Manipulation
Exploitation
Accidental influence
Fulfilling proper function
Selected for outside this context
Fulfilling proper function
Selected for outside this context
State of the world
Response

Figure 3.1: Possible influence interactions between two animals

The diagram in Figure 3.1 shows the possible influence interactions between two animals S and R, and contrasts proper signalling with accidental influence, manipulation and exploitation.

Equating communication with proper signalling and intentional icons satisfies the intuition that communication is not as simple as one animal influencing another. Proper signalling does require that one animal influences the behaviour of another, but it is also necessary that the behaviour of *both* animals has been selected for in this context. The idea that information should be transmitted is implicit in the definition: if proper signalling has occurred, then the second animal has gained information about the perceptual world of the first. For instance, if a beaver hears a tail-slap alarm, it has learned that the producer of the alarm believes there to be danger present. Note that the information will only be veridical if everything is proceeding in the historically normal way—as with intentional icons, it is no part of the definition of proper signals that they be reliable.

We should expect to find proper signals when it is evolutionarily stable for two animals to co-ordinate their behaviour in an interaction, i.e., when there is a mutual benefit in transmitting information. However, this does not mean that proper signals will only evolve in nakedly co-operative situations. Processes such as the handicap principle show that communication can be evolutionarily stable despite an apparent conflict of interests: poor-quality signallers at a handicap equilibrium do not honestly signal their low quality for any mystical reason, but because the excessive costs of exaggeration make it in their interests to do so. It is (apparently) the proper function of the peacock's tail to signal male quality just as much as it is the proper function of the bee dance to indicate the location of nectar.

The content or meaning of a proper signal is determined as for an intentional icon. We need to find that feature of the world that is picked out by the mapping rule and would allow the consumer to use the icon in the historically normal way. For instance, we can confidently assert that the leopard-alarm call of vervet monkeys means "leopard" because although it is produced in response to a variety of leopard-like objects, it is only those calls that were produced in response to leopards (and which brought about the proper flight response in other vervets) that would have contributed to the inclusive fitness of callers. If the icon maps to more than one feature—consider, for example, that leopard alarm calls must map to both leopards and to the shadows of leopards—then content can still be determined by looking at which mapping is more important for consumers, i.e., for receivers of the signal.[9]

[9] Readers familiar with the doctrine of the indeterminacy of translation, due to Quine and Davidson, may bridle at the suggestions in this paragraph. However, the claim is not that a single definite meaning can be picked out for any and all signals, but only that a Millikanian attention to evolutionary history is the appropriate tool for the pragmatic and imperfect job of content ascription.

In contrast to proper signals, the adaptations involved in exploitation and manipulation are merely tacit suppositions: for example, the design of the angler fish lure tacitly supposes that there are worm-seeking, edible fish around. The lure cannot qualify as an intentional icon or proper signal because it is not the proper function of the consumer, i.e., the prey fish, to respond to it. Still, as Figure 3.1 makes clear, exploitation and manipulation are intimately related to proper signalling, and because the proper functions of particular traits can be difficult to determine in practice, there is scope for confusion. Thus the difficulty in classifying the problem cases described earlier in this chapter (section 3.2.1). Incidental information transfer turns out to be an example of exploitation: in the example given—a bird that observes another flying down to feed and approaches in order to find food for itself—the approach of the second bird is performed in accordance with its proper function, but it is not the proper function of the first bird's behaviour to attract others. Camouflage, deception, and the direct causation of responses all qualify as manipulation. In each case the behaviour of the "signalling" animal is performed in line with its proper function in the interaction described, but the behaviour of the other party is not. Co-ordination without signalling is harder to classify—in the example given, if the behaviour of one ant genuinely has no effect on the behaviour of the other, then the situation is not even an influence interaction. On the other hand, if it could be shown that the proper function of the first ant's placement of larvae was at least partly to trigger the chamber-capping behaviour of the second, and the second ant had been selected to respond thus, then the example would be one of proper signalling after all.

Chapter 4

Artificial life as a method for studying evolution

The thesis uses computer simulations of evolution as a way of exploring—and perhaps even testing, proving or disproving—certain theories about the function of animal communication. But how can computer simulations qualify as science? Surely a computer program ultimately returns as output only that which has been put into it by the programmer, and could never be a means of discovering something new about the world? What could a simulation possibly tell us about the evolution of vervet monkey alarm calls or the aggressive displays of mantis shrimps?

Furthermore, computer simulations of evolutionary phenomena have lately been much in vogue under the banner of artificial life (AL), and the claim has been made that this is a new science of "life as it could be" (Langton, 1989). Certain proponents (e.g., MacLennan, 1991; Ray, 1994) have gone as far as suggesting that the entities within their computer programs are not mere *models* of living organisms, but are genuinely alive. What are we to make of this? The cautious reader may wish to know whether the simulations presented in later chapters are intended as models or instances of animal communication.

This chapter considers these questions. After clarifying some terminological points, a minimal amount of philosophy of science is introduced in order to make the case that the type of computer simulations used in AL can be acceptable research tools. The analytic-synthetic distinction is briefly resurrected in order to attack the claim that AL simulations are alternate worlds worthy of study in and of themselves. Finally, there is a discussion of AL's methodology problem: how can we tell good simulation work from bad? The use of evolutionary simulation models to test and extend existing ideas from theoretical biology is advocated as one way forward. The relationship between simulation modelling and the older tradition of mathematical modelling in biology is discussed, and connections are drawn between the AL approach and Millikan's work.[1]

[1] Readers with a taste for hard-nosed pragmatic empiricism may protest that this chapter should be subtitled "Shooting fish in a barrel": does anyone really doubt that simulation can be a scientific tool, or, on the other hand, believe that simulated organisms are really alive? The first objection is partly justified, because simulation methods in general *are* widely accepted; however, the way in which the individual-based, bottom-up simulations of AL can best be carried out requires clarification. As for the second objection, some authors believe exactly that, and worse. Certain fish in the AL barrel are crying out to be shot.

4.1 Some matters of terminology

4.1.1 Artificial life

What is artificial life? Clark (1996) calls it “a very broad church indeed”, and space precludes any attempt to fully delimit or describe the field here. The term is used in this chapter to refer to a particular approach, or family of approaches, to biological simulation modelling—and even this loose characterization may exclude some of the relevant research. We will consider both work that self-consciously styles itself “artificial life”, e.g., work associated with the journal and the conference of the same name (see Langton, 1989), and other work that perhaps deliberately avoids the label, going under such aliases as “the simulation of adaptive behavior” (see Meyer, 1994).

The term “artificial life” may not be the best or most accurate name for the practice of constructing simulation models to assist in our understanding of biological phenomena, but it seems to have stuck. It is interesting, in this connection, to note that Simon (1981) had doubts about the the term “artificial intelligence” as a name for the field that he helped create:

> At any rate, “artificial intelligence” seems to be here to stay, and it may prove easier to cleanse the phrase than to dispense with it. In time it will become sufficiently idiomatic that it will no longer be the target of cheap rhetoric.

Like Simon, we are probably better off bowing to popular usage than trying to coin a new, more appropriate label for the field.

Bullock (1997a) has provided a useful classification scheme for work in AL; he identifies three approaches that roughly correspond to philosophical, scientific and engineering perspectives. Bullock refers to these as high class, model class and working class AL respectively. Model-class AL will be most relevant to the discussion presented here, and the high-class philosophical applications of AL modelling will also be of some interest. We will not be concerned, however, with the working-class activities of engineers who wish to steal a trick or two from nature in the quest to build smarter robots or sleeker airframes.

AL is commonly associated with artificial evolution; much of the work employs some evolutionary procedure, typically a genetic algorithm (Holland, 1975; Mitchell & Forrest, 1994). The idea is that evolution will search the space of possible designs for a class of simulated organisms such that their behaviour will be optimized in a way that parallels adaptation through natural selection. However, the use of artificial evolution is not a defining characteristic of AL models. Work such as Reynolds (1987), who simulated flocking behaviour in birds, or Webb (1994), who built a robot model of cricket phonotaxis, is clearly part of the field, but neither model includes an evolutionary component.

At the risk of settling the question by fiat, it seems that the key feature of work in AL is the use of bottom-up, individual-based modelling. In AL simulations, entities at one explanatory level, e.g., the gene, cell, organism, or population level, are modelled explicitly (Taylor & Jefferson, 1994). These lower-level entities are then permitted to interact, in the hope that the phenomenon of interest will emerge[2] at a higher level of description. For example, Kauffman (1993) modelled individual logic gates connected in a random fashion, and found that a global

[2] In section 4.3.2 the notion of “emergence” will be critically considered. The term is the subject of heated philosophical debate which is not likely to end here. Suffice it to say that AL is strongly committed to the idea that there are such things as emergent properties, and will presumably stand or fall based on the ultimate coherence of that notion.

order emerged. MacLennan (1991) modelled individual organisms as finite-state machines, and described the emergence of a communication scheme at the population level.

4.1.2 Models and simulations

The term “model” will be used here in what should be an uncontentious way: a model is simply a description of some phenomenon. Models can be used for prediction or, more generally, as an aid to understanding. The paradigm case of the use of models in science is equational modelling, in which expressions like $F = ma$, or $x' = x\frac{w(x)}{\bar{w}}$, are used to describe processes occurring in the world. However, models can also be physical, computational or verbal, for instance. It will be crucial to the argument later on to observe that an equation or any other construction only qualifies as a model when its descriptive relationship with the world is made clear. $F = ma$ is just an empty expression until the meaning of each symbol is spelt out.

The term “simulation” has been used in different ways by different authors. Maynard Smith (1974a), for example, suggests that models are *simple* descriptions of the world, while simulations are at the other end of a continuum of complexity: “whereas a good simulation should include as much detail as possible, a good model should include as little as possible”. Along similar lines, Smithers (1994) has argued that simulations are models that can be validated in detail against some real system, e.g., a computer model of airflow that can be backed up by wind-tunnel testing. Smithers uses this definition to argue that “simulated robots”—a staple of work in AL—are usually not simulations at all, because their behaviour is never validated against that of a real physical robot.

A distinction between simplified models and detailed simulations seems reasonable, and certainly captures differences in modelling practice between, for example, theoretical biology and engineering. However, for the purposes of this chapter let us stipulate that models can be pitched at arbitrary levels of detail, and reserve the term “simulation” for use in a sense outlined by Bullock (1997a):

> A simulation is a model that unfolds over time. Rather than constructing static representations of the process under examination, such as flow charts or equations, and relying on human interpreters to simulate the passage of time, or determine the state of the system at some arbitrary time analytically, the simulation designer captures the dynamics of the original process by specifying dynamic mechanisms which govern how the system changes over time. The character of such a simulation’s dynamics is determined experimentally, through allowing the simulation to unfold over time.

Having defined simulation in this way, we can then look at the question of whether simulation models of evolution (i.e., the kind used in AL) have anything to offer over static modelling techniques such as game theory.

Note that a simulation need not involve a computer. A cellular automaton model, for instance, could be manually iterated by moving pieces on a chequer-board, as Conway did with the original Game of Life: this would qualify as a simulation. Of course, Life (Gardner, 1970; Poundstone, 1985) is perhaps a bad example: although it is certainly a simulation in terms of unfolding over time, it is doubtful as to whether Life is *modelling* anything; it is not a description of any aspect of the world except in an extremely abstract sense (this theme will be taken up again in section 4.4).

4.2 This thing called science

Science might be described as the art of improving our models of reality. Detailed arguments for a comprehensive position in the philosophy of science are clearly beyond the scope of this chapter; however, the central issues can be stated quite briefly and will better allow us to assess the potential of AL as a scientific tool.

The fundamental insight of empiricism was that our models of the world must be testable—models are *good* models not because they are elegant or in line with religious dogma, but because they allow us to make predictions that are borne out by experiment. Hume (1748) spoiled the fun somewhat by pointing out the problem of induction. Hume said that "all inferences from experience suppose, as their foundation, that the future will resemble the past", i.e., that just because a particular model has made correct predictions up until now, that gives us no logical reason to suppose that it will continue to do so. Hume's observation seemed to detract from the power of the experimental method to tell us which of our models were true. Popper (1968) attempted to salvage the situation by suggesting that science was not about proving particular models correct, but that scientists should seek to *falsify* as many models as possible. In Popper's view, scientific knowledge consisted of those models that had not yet been falsified by experiment.

However, Popper should have known that the game was up. Quine (1951), among others, had shown that models do not stand or fall alone, but in combination with each other. As Quine put it, "our statements about the external world face the tribunal of sense experience not individually but only as a corporate body." To give a contrived example, if an astronomer predicts a solar flare and builds an instrument that should allow her to detect it, and then fails to do so, she can either blame her model of solar flares or her theory of instrument construction. It is thus impossible to neatly sort our models into the categories "falsified" and "not yet falsified". Any one model or statement may be maintained as true despite all manner of apparently contradictory evidence, as long as one is prepared to make adjustments elsewhere in one's overall conceptual scheme. Our astronomer might doggedly believe in her model of solar flare occurrence, despite its apparent failure, by rejecting increasingly sophisticated theories of instrument construction.

We therefore find ourselves in a position where no one experiment can provide incontrovertible evidence in favour of one competing model over another. Our criteria for choosing between models become pragmatic. We invoke Occam's razor and choose the simplest of two models that can explain a phenomenon; we respond to new or surprising experimental evidence by adjusting our conceptual scheme to the minimum degree that will accommodate it. Quine's (1951) own words on the subject are again apt:

> Each man is given a scientific heritage plus a continuing barrage of sensory stimulation; and the considerations which guide him in warping his scientific heritage to fit his continuing sensory promptings are, where rational, pragmatic.

4.3 Artificial life as science

4.3.1 Does artificial life qualify?

How does artificial life fare with respect to this Quinean view of science? We should first deal with a trivial sense in which AL could be said to be part of the scientific project. Few positions in the

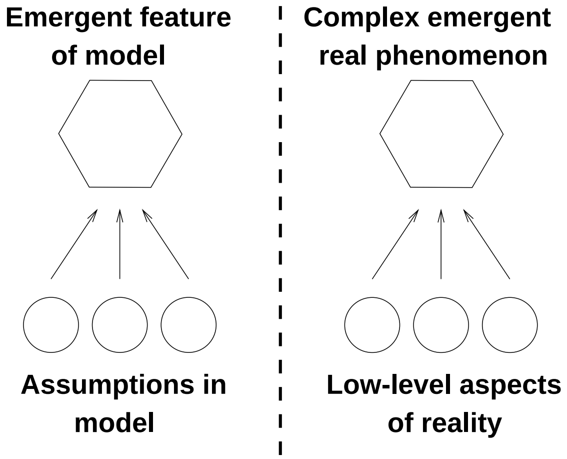
Emergent feature of model
Complex emergent real phenomenon
Assumptions in model
Low-level aspects of reality

Figure 4.1: The proposed isomorphism between artificial life models and reality.

philosophy of science have much to say about the *generation* of new models; this usually remains a mysterious process. The much mythologized scientific method in fact suggests that hypotheses can come from anywhere, and that what matters is how they are put to the test. In this modest way, then, AL as much as astrology might function as a source of testable models.

However, we can do better than this. Quine's account of science suggests that the process of generating new models (i.e., theory building) is not random but is done in an attempt to accommodate new information while requiring only small changes in our overall conceptual scheme. We want our new models to be testable, but we also want them to be largely consistent with what we already believe. Let us consider one way in which AL could provide such models (see Figure 4.1).

Imagine that we have observed some real-world phenomenon in animal behaviour, call it E_R, and are interested in using an AL simulation to generate a model or theory that might account for it. Recall that AL simulations model a low level of description explicitly; thus, there will be a set of assumptions or axioms A_M that exhaustively specify the lower-level description of the model. When the simulation is run, let us assume that the assumptions A_M give rise to some higher-level emergent phenomenon; call it E_M, and thus $A_M \rightarrow E_M$. The critical point is this: if E_M, the emergent outcome of the simulation, is sufficiently similar to E_R, the real world phenomenon ($E_M \approx E_R$), then it seems reasonable to advance the empirical claim that there exist A_R, real-world analogues of the low-level simulation assumptions, and that, through a similar emergent process, these factors give rise to the real-world phenomenon E_R. That is, $A_R \rightarrow E_R$. The theory is, of course, not assumed to be true, but referred back to empirical biology as a working hypothesis. To the extent that the low-level assumptions are in line with what we already know about this particular organism or system, this is a paradigm case of AL functioning as a scientific tool: a conceptually economical model of the high-level phenomenon is produced that requires us to posit nothing new beyond the background assumption that an emergent process is at work.[3]

The work of Reynolds (1987) will serve as an illustrative example. In Reynolds's simulation,

[3] The notation used here makes it plain that there could be other ways to develop an empirical claim from an AL simulation: for instance, $A_M \rightarrow E_M$ and $A_M \approx A_R$ would suggest a prediction of E_R. Alternatively, $E_M \approx E_R$ and a minimal set of assumptions A_M such that $A_M \rightarrow E_M$ could be used to predict a minimal set A_R such that $A_R \rightarrow E_R$, if the set of candidates for A_R was large.

flocking behaviour (E_M) emerges from the interaction of simple agents. At the level of the individual, there are three rules that the agents follow: maintain a minimum distance from obstacles and other agents, match velocities with nearby agents, and move towards the collective centre of mass of nearby agents. These are the assumptions, A_M. At the group level, the actions of individuals give rise to behaviour that substantially resembles flocking in real birds ($E_M \approx E_R$): when confronted with an obstacle, for instance, the flock splits into two sub-flocks and then re-forms on the other side.

Reynolds has thus shown that the three assumptions are sufficient to generate flocking at a global level: $A_M \rightarrow E_M$. To be certain that only the three assumptions are responsible for the flocking behaviour, we might insist on a careful examination of Reynolds's code, and there might also be debate over whether the flocking is "close enough" to the real thing. These are important but not crippling objections. If Reynolds was to translate the assumptions A_M into real-world terms A_R, and assert them *as empirical claims*, then he would be advancing a scientific theory of flocking in real birds: that $A_R \rightarrow E_R$. What was previously just a computer program would then become a model of reality.

The testability of such a model lies in the fact that we can look at real birds and determine whether they are behaving in accordance with Reynolds's assumptions. Is it plausible that they're moving towards a local centre of mass, for instance? If the evidence suggests that flocking birds really do behave like that, then Reynolds's model is entitled to stand among our rationally maintained beliefs. The theory would be adopted subject, as always, to the proviso that it might be replaced or adjusted as new evidence comes to light.

It should be stressed that Reynolds has never actually asserted the three assumptions as a theory of bird flocking, i.e., he has never proposed that his program might describe reality. His flocking algorithm has been put to more working-class use, e.g., in producing realistic animal movement in computer-generated film sequences. The model would probably not be a very good one anyway: Zaera, Cliff, and Bruten (1996) have argued that real animals in real flocks do not have access to the sort of information that Reynolds's model requires them to have. In Quinean terms, assuming that Reynolds's low-level assumptions are true of real birds would require the revision of too much of what we already believe. Nevertheless, Reynolds's work has been developed as an example because the simplicity of his assumptions serves to illustrate the way AL *could* make empirical claims, and thus qualify as science. In fact, it is disturbingly difficult to find examples of AL projects that *do* make testable claims. For instance, Toquenaga, Kajitani, and Hoshino (1994), who look at foraging behaviour in egrets, and Prusinkiewicz (1994), who describes models of morphogenesis, certainly base their work on the relevant biology, but stop just short of presenting their conclusions in the form of testable predictions. The work of Kitano, Hamahashi, Kitazawa, Takao, and Imai (1997) on morphogenesis in *C. elegans* and *Drosophila* stands out as an example of work within AL that has led to concrete predictions concerning real-world entities.

4.3.2 The concept of emergence

It will be clear from the discussion in the previous section that AL models involve a commitment to the concept of emergence, an idea that has been variously defined. Bhaskar (1978) is a leading proponent of emergence among philosophers of science; he invokes the term when:

> . . . the operations of the higher level cannot be accounted for solely by the laws governing the lower-order level in which we might say the higher-order level is 'rooted' and from which we might say it was 'emergent.'

In the strongest sense, then, an emergent property is a high-level property that arises from low-level constituents, but is neither predictable from nor reducible to those constituents. A standard example in the philosophical literature is the water molecule, H_2O. No matter how much one knew about the properties of hydrogen and oxygen alone, it is claimed, one would not have been able to predict the properties of water (e.g., transparency, liquid state at room temperature); furthermore, statements about the properties of water cannot be reduced to statements about hydrogen and oxygen, and therefore molecular chemistry is emergent with respect to atomic chemistry. Kim (1992) is sceptical about such a notion, referring to it as "magical emergence," and suggests that it violates a basic physicalist assumption that no special causal powers should appear at higher levels of description. In contrast, Davidson (1970) has argued for the irreducibility of the mental to the physical, and thus that mind, at least, must be an emergent phenomenon. Dennett (1991b) and Clark (1996), among others, show that a weaker sense of emergence can be justified on an epistemological basis. That is, given our limitations as knowers, absolute reductionism (in the sense of the Laplacean ideal) is not possible and therefore we need the concept of emergence in order to make sense of things. Clark (1996) discusses the use that the term is put to in AL, and argues for a weak, functional conception: emergent phenomena are simply those that arise as a result of collective activity as opposed to the action of a dedicated component or control system.

There is thus disagreement about in exactly what sense, if any, we should say that a result like flocking in Reynolds's simulation is *emergent* from lower-level assumptions. However, let us assume that some notion of emergence will ultimately prevail: it is certainly a pragmatic concept, without which it would be difficult to make sense of the apparently hierarchical relationship between the sciences, i.e., physics → chemistry → biology → cognitive science, etc. Even if our commitment to emergence is only in Clark's weak sense, the function and value of AL modelling is clear. The function of an AL simulation is to establish the plausibility of the central emergence hypothesis: that if the low-level assumptions hold, then *this* should happen. The unique value of AL modelling lies in the fact that such a theory would be difficult to formulate any other way. There is, of course, a danger of circularity here, in that AL simulations and the concept of emergence could become caught up in a mutually justifying but empirically empty loop. Emergence needs clarification independently of its use in AL.

4.4 Artificial life and the return of the analytic-synthetic distinction

A method has been outlined by which AL could function as a mode of scientific inquiry, but it is essential to this method that AL refers its theories back to empirical biology (or some other empirical domain). What then of Langton's (1989) enthusiasm for AL as the science of life as it could be? It is difficult to see how to test an AL model that, to the extent that it can be said to make predictions at all, only makes predictions about some state of affairs utterly alien to terrestrial biology. Is such speculative AL science? The short answer is no, it is not. An old and disreputable distinction between analytic and synthetic statements will now be revived in order to elaborate on this.

Unfortunately, the terms “analytic” and “synthetic” have an existing usage in AL that is different from the sense intended here. “Analytic” is used by Langton (1989) to describe the traditional biological approach, in the sense of analyzing a system in terms of its component parts. Langton uses “synthetic” to refer to the AL approach of building a complete artificial or synthetic system.

The terms will be used here in roughly the same way that they were used by the logical positivists: mathematical and logical statements are paradigm cases for analytic statements. Analytically true statements are supposed to be true by virtue of the meanings of the terms contained within them, e.g., “seven is prime,” or “a sow is a female pig.” A statement is analytically true if and only if it is true and there is no conceivable empirical observation that would render it false. For example, it is analytically true that the knight in chess cannot reach the opponent’s back row from its starting position in fewer than four moves. Once we have established that the knight can move at most two squares forward per move, and that there are seven steps to the opponent’s back row, it should be clear that there is no point in observing a great number of chess games to see whether the knight ever gets there any faster. (It is not being suggested that the rules of chess are necessary truths, but that given the rules of chess, the result is analytic).

Synthetic truth is defined in opposition to analytic truth: something is synthetically true if its truth value is to be found not in the meaning of the terms used to express it but by examining the world. Statements are synthetically true if they are true but the world could have been otherwise. It is a synthetic truth, for instance, that nitrogen constitutes 78% of the Earth’s atmosphere, or that pigs are omnivorous. The statement “Everest is the world’s tallest mountain,” is generally agreed to be true, but that does not mean that it always has been or always will be. It is a matter to be settled by empirical investigation, and thus a synthetic truth.

The analytic-synthetic distinction has, of course, been deeply suspect since Quine (1951) argued that because our theory-statements stand or fall not alone but in combination, there is no sensible dividing line between statements that can be held true come what may, and statements whose truth is contingent upon the state of the world. (Indeed, given the enthusiasm with which a Quinean view of science has been endorsed, it may seem positively perverse to resurrect the distinction.) As argued in section 4.2, any statement can be held true in all circumstances, by more or less radical adjustments in the rest of one’s belief system. Similarly, no statement is immune to revision in the face of new evidence. However, although Quine established that there was no binary distinction to be made between the analytic and the synthetic, he did not argue that human knowledge was doomed to a hopeless state of flux. Some statements are much less likely to be revised in the face of experience than others. Quine used the metaphor of a “field of force” that represented the whole of scientific knowledge, and suggested that statements at the periphery were vulnerable to change in the light of new experience, whereas statements at the core were relatively invulnerable. These “core” statements are precisely those that have been characterized above, with deliberate naïveté, as analytic. For example, it is easy to imagine a single observation that would lead us to revise the statement “Jupiter has sixteen moons,” and harder to imagine a set of observations calling for revision of “Crocodiles generally eat other animals.” It becomes very hard indeed to conceive of the observations that would cause us to revise statements of a more analytic nature, such as “Right now it is either Monday or some other day of the week,” or “$2 + 2 = 4$.” Quine’s argument implies a *continuum* between analytic and synthetic statements, and it is with this proviso firmly in mind that the distinction is invoked. In what follows, the reader should prepend the modifier “relatively” to instances of the terms “analytic” and “synthetic”.

The distinction can be usefully applied to AL. A bald statement of the results of an AL simulation—i.e., a description of the simulation’s output, without the assertion of a possible descriptive relationship with the world—is effectively an analytic statement. Thus such simulations, in and of themselves, cannot lead to synthetic truths, i.e., they are not tools for empirical discovery. The argument runs as follows: AL simulations are computer programs, and a computer program is the conjunction of a collection of mathematical and logical statements. Conjunctions of two or more analytic statements are still analytic.[4] Therefore any result of an AL program—such as the difference in mean fitness scores between two conditions, or the qualitative character of some dynamic behaviour pattern—is an analytic result. The result is implicit in, and determined by, the way the program is put together.

It might be objected that the programs that instantiate AL simulations are such complicated combinations of logical and mathematical statements that we cannot analyze or predict their behaviour without actually running them, and therefore that in running an AL program we may discover things about the world, which means in turn that AL programs give us synthetic truths. Certainly it is true that in practice our mathematical skills are not up to predicting the outcome of even moderately complex AL simulations; indeed in section 4.1.2 we *defined* simulations as models that must unfold over time, the implication being that there is no known mathematical short-cut for predicting the state of the simulated system at an arbitrary time t. However, appealing to the poverty of our collective mathematical ability is no argument. When the young Carl Friedrich Gauss came up with the general formula for the sum of any series in response to a lengthy summation problem, he arrived at an analytic truth. His less gifted classmates, who presumably added up the long list of numbers in a conventional fashion, were not gaining access to synthetic truth. They were simply dealing in analytic statements the long way round; there is a sense in which AL programmers are doing the same. That we cannot *analyze* a problem statement does not mean that the statement is not *analytic*: $10^{100}+1$ is either prime or not prime, and analytically so, despite our lack of ability to say which.

Recall that analytic statements are, in Quine’s terms, the statements we would be least likely to revise in the face of new empirical evidence. Now imagine that an AL-inspired theory—let us say Reynolds’s flocking model—is the received wisdom in some area of animal behaviour. If new observations of real bird flocks show that they deviate from the Reynolds model in the way they approach a certain kind of obstacle (for example), which statements are we likely to revise? Certainly we may doubt the new empirical reports, but if further field studies corroborate the initial findings then we will presumably accept them. We will then revise a central claim of Reynolds’s theory, $E_M \approx E_R$: that there is a correspondence between the emergent flocking phenomenon in the simulation, and the flocking phenomenon in real birds. We would be extremely unlikely to revise $A_M \rightarrow E_M$: that Reynolds’s three assumptions lead to “Reynolds-flocking” (now distinguished from the real thing) in his simulation. The new data, and indeed (almost) all possible new data, simply do not have any bearing on the truth of $A_M \rightarrow E_M$. That $A_M \rightarrow E_M$ is true is something we

[4] For instance, $1+1=2$ and $2+2=4$ are both analytic statements. Putting the two statements together in some way, e.g., by constructing $(1+1=2)\wedge(2+2=4)$ or $(1+1)+(2+2)=6$, still produces a statement whose truth or falsity is an analytic matter.

can demonstrate on any computer, or by hand if necessary, no matter what real birds are doing in any possible world. $A_M \rightarrow E_M$ is analytic if anything is.

To say that AL simulation-building is an analytic exercise, then, is simply to say that it involves using a computer program to see what logically follows from a certain set of assumptions. Simon (1981) points out that this is no trivial task:

> . . . even when we have correct premises, it may be very difficult to discover what they imply. All correct reasoning is a grand system of tautologies, but only God can make direct use of that fact. The rest of us must painstakingly and fallibly tease out the consequences of our assumptions.

A discipline dealing with analytic statements is hardly without value: consider mathematics. Nor is AL limited to the analytic. It has been argued in section 4.3.1 that, if the authors of a simulation make clear its descriptive relationship with the world, AL can be a source of empirical claims (i.e., synthetic statements). Unfortunately, AL cannot also function as a proving ground for those claims; they must refer back to the world. We should remember that in this respect AL is in no worse a position than any other theoretical science. The theories offered by AL will sometimes be difficult to test, but that in itself does not count against them.

It may also be the case that an AL project makes no empirical claims, but nevertheless uses a set of assumptions that are strongly suggestive of some real-world problem. For example, an AL simulation of the evolution of sex could offer some useful analytic results, perhaps that trisexual reproduction is stable under certain conditions, but the authors might feel that any specific empirical claims would be premature, i.e., they choose not to present their work as a *model*. There is nothing inherently wrong with such an approach. That particular analytic result says nothing about the world, but may be useful later: if a biological theory comes along that asserts the contrary, the AL result shows the incoherence of the biological theory. If trisexual reproduction is one day observed in some exotic species, the AL result *might* form the basis of a theory to explain it.

While analytic AL cannot, of itself, uncover empirical truth, it can certainly be used as a tool for assessing the *logical coherence* of pre-existing theories. For example, Hinton and Nowlan (1987) used an extremely simple simulation based on a genetic algorithm in order to show that the Baldwin Effect was a plausible force in evolution. Their work was not asserted as a model of any real system, but as an abstract demonstration that Baldwin's theory could work. Binmore (1992), commenting on the similar use of game-theoretic models, has remarked that "the use of such unrealistic, oversimplified formal models for testing the *internal consistency* of theories is very important indeed, although widely misunderstood." To give a more extended example, Franceschini, Pichon, and Blanes (1992) were interested in the question of how flies avoid obstacles. Through neuroscientific investigations, they had already developed a theory that a certain neural structure was extracting information from the motion flow-field such that the fly could be said to know how far away an object was. This theory was based on the principle that if one is travelling forward at a constant speed, closer objects will move across the retina at a higher angular velocity. The theory could be reasonably well-supported based on the neurological evidence alone, but in order to strengthen their case, Franceschini et al. built a robot along the postulated design principles—the robot's task was to navigate through a field of obstacles without collision. The robot in fact did this successfully. The authors thus demonstrated that their theory was analytically consistent:

there was no logical contradiction in proposing that the postulated neural structure could extract range-to-target information. If the robot had *failed* to navigate the field of obstacles, and if we could assume that the robot really had been built in accordance with the postulated design principles, then the exercise would have effectively falsified the theory on analytic grounds alone. There is much useful work for AL to do along these lines.

It should be admitted at this point that the analytic-synthetic distinction cannot be made so easily in the case of robotic AL. A robot is not a simulation or a mathematical formula, and the question of how a particular robot will behave in a particular environment does seem to be a matter for empirical investigation. Possibly we have to admit that the robotics partisans who insist on the primacy of physical realizations over (computer) simulations—notably Brooks—have a point. However, in work with robots there will still be statements like $E_M \approx E_R$ and $A_M \rightarrow E_M$. If we consider Webb's (1994) work on phonotaxis in a cricket-like robot, and imagine future empirical data inconsistent with her model, we would be, as before, more likely to revise the hypothesis of a correspondence between robot and cricket than we would be to revise hypotheses about how the robot functions. In this sense the correspondence assumption is "less analytic" than the theory of the robot's architecture.

4.4.1 Confusion between the analytic and synthetic modes

Our mathematical intuitions being what they are, the implications of an AL model are usually not obvious to us. We have to run the simulation to find out what happens. Simulations are usually stochastic procedures, and thus we run the simulation more than once, with different random seed values, in order to assess the range of variation in the results. We may also be interested in the effects of different parameter values or initial conditions on the emergent outcome. This suggests the use of statistics and such traditionally experimental concepts as control conditions, 2×2 designs, etc. We have discussed emergent results in simulations as if they were always as easy to see as flocking behaviour, but of course the emergent result of a simulation may not be easily recognizable. For this reason, statistics and the experimental method are useful in analytic AL. However, we should not be fooled by the fact that we are using some of the props of empirical science into thinking that we are in fact doing empirical science.

AL work, to its detriment, often ignores the distinction between the analytic and the synthetic. This typically occurs when an analytic result is asserted as if it were an empirical finding. In the extreme case, researchers such as Ray (1994) believe that their simulations are in fact alternate universes, and thus that the results of a simulation constitute empirical data: "digital life exists in a logical, not material, informational universe." Langton (1989) has made the related claim that AL "will be interested in whatever emerges from the process [of simulation], even if the results have no analogues in the 'natural' world." But Langton and Ray want AL to have its cake and eat it too: whether or not simulation results appear relevant to terrestrial biology, they are to be regarded as empirical science. This is just wrong. As has been argued, an AL simulation that produces an emergent phenomenon analogous to something in the world might prove the inspiration for a theory that can be fed back to empirical biology. An AL result that apparently has no parallels in the world may well be interesting, and it might be put aside until the day a parallel is discovered, but it is not a scientific model. An AL result, considered alone, is simply the analytic finding that a certain set of assumptions imply a certain outcome. This confusion may be rooted in the tendency of some AL researchers to use the terms “world” and “universe” literally when referring to their simulations (Helmreich, 1995)—a little metaphor can be a dangerous thing.

4.5 A sound methodology for artificial life

4.5.1 In the footsteps of theoretical biology?

In AL simulations a complex emergent phenomenon is sometimes presented as being of interest in its own right, despite the arbitrary, theory-free nature of the simulation that gives rise to it. As we have seen, this will not do. Nor is it fruitful to talk about vague similarities between simulated and real phenomena: Bullock (1997a) points out that post-hoc parallels drawn between an AL simulation and reality are either “merely accidental (and thus not interesting), or merely purposed (and thus not interesting).” AL needs some way of ensuring that it is in the business of constructing *models* rather than just abstract simulations. Miller (1995) has suggested that the solution to this methodological problem is for AL to attach itself to theoretical biology. Miller points out that models in theoretical biology, usually implemented as a system of differential equations, contain many simplifying assumptions, e.g., random mating and infinite populations. This is done in order to make the mathematics tractable. AL, Miller says, could take such models, re-implement them in the bottom-up simulation style, and gradually relax the simplifying assumptions. An AL model constructed in this way would thus perform the analytic task of making clear the implications of the substantive core of the original theoretical biology model. This, in turn, would help to make it clear which (if any) empirical predictions might be warranted.

It will be clear from the time devoted to biological theory in chapter 2 that Miller’s suggestion has been taken up in this thesis. But what might Miller’s program add to existing research in theoretical biology? After all, biologists have been using computer simulations for years. However, simulations are often used in biology as simply a quick and dirty way of finding approximate solutions for equational models that are too complex to be solved analytically. In AL work, the simulation *is* the model, and not a surrogate for it. There are some instances in biology in which simulations have been presented as models in themselves; however, it is usually the case that these simulations do not stray outside the boundaries of more traditional conceptual tools such as game theory. For example, Maynard Smith and Price (1973) and van Rhijn and Vodegel (1980), in studying animal contests, both used simulations in which all possible combinations of the strategies under consideration were played against each other. The results from the simulations were then used to construct payoff matrices such that potential ESSs could be determined in the conventional manner.

More recently, simulations have been done that blur the boundary between AL and biology—work by biologists, published in biology journals, that nevertheless uses such staples of the AL approach as genetic algorithms and artificial neural networks. To give two examples related to communication, Arak and Enquist (1993) looked at the way the evolution of pattern recognition might lead to “hidden preferences” for certain types of novel signals (see section 2.6.1), and Krakauer and Johnstone (1995) described a model of the evolution of honest signalling in which a level of exploitation or cheating was tolerated. There seems little point in bickering over whether this sort of work constitutes biologically influenced AL or AL-influenced biology.

Granted that there has been this convergence of methods by some authors, we must still ask what distinguishes AL from the traditional game-theoretic and population-genetic models used in theoretical biology. More particularly, what can we achieve with the former that we cannot achieve with the latter? One advantage of AL modelling is that it gives us some tools for coming to grips with the dynamics of evolution: the use of genetic algorithms allows us to look at the trajectories of populations moving through the space of possible genotypes or strategies. This allows AL simulations to examine competing hypotheses that would be hard to distinguish using conventional methods. For example, both Zahavi's handicap principle and the signalling arms race theory of Dawkins and Krebs might predict that costly signalling would evolve in a certain context. Thus, each theory predicts the same state as an end-product of the evolutionary process, and they differ only in their predictions about how and why the population would arrive at this point. AL methods are natural tools for assessing theories that stand in this relationship to each other.

Dynamically oriented simulations can also help with what is known in game theory as the *equilibrium selection* problem. This refers to the fact that game theorists have no universally agreed way of deciding which of several equilibria is the "correct" or "natural" solution to a game. As Binmore (1992, p. 395) puts it:

> When the processes studied converge, they always converge to an equilibrium of the underlying game. But, when the game has several equilibria, the particular equilibrium to which the process converges will depend on the historical accident of where the process started from.

Starting an AL model from a theoretically justifiable initial state (or from a series of possible states) and observing the subsequent evolution of the population goes some way towards solving the problem; this idea is further developed in chapter 6.

The advantages of AL in looking at evolutionary dynamics should not be overstated, however. Whereas evolutionary game theory necessarily focuses on equilibria, it is certainly possible, in sufficiently simple games, to construct a map of the strategy space with vectors indicating the direction that an evolving population would take—see Gomulkiewicz (1998) for a discussion of this approach, and Maynard Smith (1982), among others, for examples. Population-genetic models are themselves explicitly dynamic; the equations with which they are expressed describe changes in the frequency of different genotypes over time.

Having said that, a significant problem with population-genetic models is that they do not really involve a population; there may be selective effects, due to the interactions between individuals, that cannot be captured with an equation describing genotype frequencies in the abstract case of an infinite population. AL models, of course, genuinely instantiate a population of simulated organisms. In this way they allow us to investigate questions about the adaptive value of particular individual strategies down to arbitrary levels of detail. For example, we can look at the effects of space and mobility—the fact that an animal is not static but moves about in space and encounters other animals in a non-random way. In equational models space can be captured to some extent, e.g., by imagining that animals are arranged in some abstract topology, but even this becomes mathematically complicated and cannot be taken very far. AL-style models, as noted in section 4.3, are possibly the only way to get to grips with theories that propose that certain effects might *emerge* from the low-level details of space, time, and interactions between organisms.

AL simulations are not without their problems. In section 4.4.1 it was mentioned that AL models often require pseudo-experimental methods and the use of statistics. Interpreting the results or implications of a particular simulation is therefore complicated. In particular, the fact that a simulation involves setting many parameters to real values means a sacrifice in generality. With an equational model the "message" can simply be read from a formula: if $V > 2C$ then strategy X will be evolutionarily stable. The results of an AL simulation will be more equivocal: for example, the programmer might set $C = 1$, and observe that strategy X evolves in the "high-V" condition ($V = 3$), but not in the "low-V" condition ($V = 1$).

There is also a grave danger of building artefacts into a simulation. Because writing a computer program involves being explicit about every imaginable detail of the "lives" of the simulated organisms, including many details that do not seem to be important with respect to the theory under investigation, it is easy for a programmer to make choices that will unduly affect the results of the simulation. For example, Nowak and May (1992) simulated a population of organisms that were arranged on a grid and whose lifestyle consisted of playing the well-known Prisoner's Dilemma game (see Axelrod, 1984) against their neighbours. Over time, remarkable fractal patterns appeared on the grid, made up of regions inhabited by co-operators and defectors. Nowak and May suggested that their model provided clues about the importance of spatial arrangement in the evolution of co-operation. However, the simulation was criticized by Huberman and Glance (1993): when a more plausible asynchronous updating method was applied to the game-players in Nowak and May's model, the results were completely different—and not nearly as interesting.[5] Problems with artefacts in a simulation will be explored in more detail in chapter 5, in which a model by MacLennan and Burghardt (1994) is replicated and critiqued.

On the other hand, artefacts can easily be found in game-theoretic analyses too. A game-theoretic model is only as good as the strategies that the author has elected to include. It may well be that strategy A is an ESS when considered with strategies B and C, but would not be one if strategy D were included in the model. The use of AL techniques does not make this problem go away, but, especially in the case of such powerfully expressive architectures as artificial neural networks, the problem is eased because very many strategies are accessible to evolution.

The simulation methods of artificial life thus complement the older traditions of modelling in theoretical biology; they are not inherently better or worse. It therefore seems reasonable to pursue Miller's (1995) program of extending and testing models from biology using the methods of AL. Sympathy should be noted, however, for objections raised by Di Paolo (1996) concerning Miller's ideas. Di Paolo points out that the approach advocated by Miller is quite conservative: to say that the *only* sensible course for AL is to become an appendage of theoretical biology places unnecessary limits on the scope of the field. There is no reason why AL, considered as a style of simulation, should not find applications in disciplines beyond biology.[6] Secondly, if AL follows theoretical biology, then AL models will necessarily be pitched at the genetic level characteristic of theoretical biology; it seems foolish to tie AL work to this mode of explanation when it can also

[5] In fact Nowak and May have offered a reply of sorts to the criticisms levelled by Huberman and Glance: see May, Bonhoeffer, and Nowak (1995). The point remains that phenomena observed in a simulation can be highly dependent on seemingly minor implementation details.

[6] See, for example, the work in computational economics by such authors as Vriend (1995) and Tesfatsion (1997).

address others. Finally, Di Paolo points out that Miller's suggestion relieves AL researchers of the burden of model and theory construction only by handing it over to theoretical biologists. On the face of it, there is no reason to suppose that theoretical biology has cornered the market in theory construction. Essentially, Di Paolo is saying that Miller has described one way in which principled AL work could be carried out, but it is surely not the whole story. Di Paolo is almost certainly right; however, Miller's program will nevertheless be adopted here because it ensures that the AL simulations presented will have a descriptive relationship with the world, i.e., they will be *models* and not free-floating curiosities.

4.5.2 Connections with Millikan

The Millikanian perspective suggests that paying attention to the evolutionary history of an AL simulation is a good idea. In an evolutionary simulation, if it can be shown that a certain behaviour or trait has been selected for in the past—i.e., that simulated organisms possessing the trait experienced greater reproductive success than those that did not—then it can be said to have a proper function. For example, we have defined proper signalling, in chapter 3, in terms of proper functions. It should therefore be possible to show that particular interactions between simulated organisms count as real communication, because the evolutionary history of the simulation allows the existence of the appropriate proper functions to be determined.

On the other hand, a historical analysis of the data may be misplaced in very simple simulations. For instance, in a simple model in which only two behavioural strategies are available to the simulated organisms, we know that the ancestors of an individual were selected because of their propensity to do X rather than Y, because there was no other dimension on which they could vary. In this sense, the proper function of an evolved strategy may be immediately apparent from "snapshot" data that describe the population at the end of a simulation run; the historical course of evolution in sufficiently simple models may be ignored because it can not hold any surprises.

Sometimes the notion of proper function can be difficult to hang on to when the results of a simulation are examined in detail. For instance, consider communication: we may have a particular game being played by a population over evolutionary time, with a particular set of payoffs and certain starting conditions. Communication then either evolves or it does not. Or perhaps it evolves in a certain proportion of simulation runs but not in others. Where has function gone in this picture? The answer is simple. When communication reliably *does not* evolve, that means that no coherent functional story can be told for communication in that context. The associated prediction is that no communication will be observed under those circumstances in the real world. (Of course, if we know that there is in fact a great deal of communication out there in the world under those conditions, then we have not constructed the model properly.) If communication *does* evolve, that means that the functional story for the system is described by the parameters and structure of the game, e.g., that vervet monkeys warn their conspecifics of predators, despite an apparent cost to themselves, in order to maintain reciprocal relations with those around them—a vervet monkey that does not signal will lose out because others will not warn him.

Finally, it should be noted that a commitment to Millikan's concept of proper function means recognizing that evolved behaviours in simulations have proper functions as much as anything in the real world does, as the behaviours have persisted over a history of selection; Millikan takes pains to specify that she does not want the notion of proper function limited to genetic evolution. However, this curious result should not be taken to license the worst excesses of speculative AL, because the idea is that the evolved functions in the model are isomorphic to similar evolved functions in the world.

Chapter 5

Artificial life and communication

We have seen in chapter 4 that the simulation methods of artificial life provide us with a tool for examining the adaptive value of different behavioural strategies. Evolutionary simulation models potentially have much to tell us about situations, such as communication, in which the strategic choices of individuals give rise to a global phenomenon in a way that is difficult to capture with traditional mathematical methods. The purpose of this chapter is therefore to review work that already exists, within the literature of artificial life and related fields, in which computer simulations are used to model the evolution of communication. Some of this work is only superficially relevant to our concerns: for instance, some of the extant models are pitched at a linguistic level, looking at the evolution of specific features of human language and assuming the existence of a prior, more primitive form of communication. However, there are a number of authors who have used simulation models in much the same way as is being proposed in this thesis: i.e., to look at the function of communication systems and examine the conditions necessary for the evolution of simple signalling schemes from non-communicative origins. Unfortunately this body of work is of limited value in helping us to understand how and why communication evolved in the real world. Given the argument presented in chapter 4—that work in artificial life must be integrated with the broader project of scientific inquiry—it will be seen that much of the earlier research consists of isolated proofs of concept, and could be improved upon by closer attention to links with existing biological theory.

In order to illustrate this point, a simulation model of the evolution of communication by MacLennan and Burghardt (1994) is replicated, critiqued, and extended. MacLennan and Burghardt's study and its replication will be described in some detail; this is intended to serve several additional purposes. Firstly, detailed consideration of a specific model introduces the reader to the issues involved in constructing an evolutionary simulation, which will be useful when the time comes to justify the models presented here in chapters 7, 8 and 9. Foremost among those issues is the problem of artefact, and a recurring theme in the discussion of MacLennan and Burghardt's work will be the presence of factors that cannot be comfortably mapped to reality but which nevertheless have a great effect on the simulation's outcome. Finally, it follows from the pragmatic view of science that has been adopted in the thesis (see section 4.2) that replicating a simulation from its written description is itself a contribution to a principled artificial-life approach: because simulations can be large and complex computer programs, and because they require the programmer to make many arbitrary decisions about details that may or may not be important, it is essential that several versions of the same basic model be independently constructed by different authors. This should allow us to decide whether a surprising or puzzling result (i.e., one that has implications for the survival of other beliefs we may hold) should be taken on board or rejected as an artefact.

5.1 Previous attempts to simulate the evolution of communication

5.1.1 Overview

The body of published work describing artificial-life models of communication is not particularly large; however, much of the research is novel and idiosyncratic, and this makes the task of grouping and categorizing it difficult. Nevertheless, we must impose some sort of classification scheme on the literature if we are to avoid simply listing the work that exists.

Simple signalling systems

Most important for our purposes are those papers that look at the evolution of a communication system in a population of agents who do not initially possess one. These simulations typically define a game that the agents participate in; for instance, their lifestyle may consist of repeatedly being grouped into randomly assigned pairs, and then being rewarded with “fitness points” to the degree that they co-ordinate their behaviour with that of their partner. Communication is made possible but not inevitable because one agent can observe the behavioural choice of the other before responding, for instance. The genetic algorithm or other evolutionary procedure is set up in such a way that a reasonable number of possible behavioural strategies are available to the agents. Evolution is allowed to proceed for a certain number of generations and then the success or failure to evolve communication is investigated. This can be done either by measuring the level of behavioural co-ordination achieved, or by blocking the “communication channel”, i.e., preventing one agent’s behaviour from influencing another, and observing whether the agents’ success rate drops. Two early and often-cited papers that fit this description are MacLennan (1991; later expanded in MacLennan and Burghardt, 1994) and Werner and Dyer (1991). MacLennan’s work involved agents who were rewarded for successfully communicating to each other a local state (accessible only to the signaller) that could take one of eight values; this simulation will of course be described in detail in section 5.2. Werner and Dyer looked at agents that lived on a two-dimensional grid, and demonstrated the evolution of a signalling protocol that allowed immobile females to guide blind, mobile males toward them for mating opportunities.

This kind of simulation work varies on two dimensions: firstly, the lifestyle of the agents can be more or less elaborate and complicated. Thus, we find elaborate simulations such as de Bourcier and Wheeler’s (1994), in which the agents exist in a continuous two-dimensional world (i.e., not just a simple grid-world) and possess quite complex visual systems which they use to find food and to detect the approach of other agents. The authors argue that territorial signalling evolves (see also de Bourcier & Wheeler, 1995, 1997; Wheeler & de Bourcier, 1995); the evolved agents forage in loosely-defined territories and move threateningly towards others who encroach on their territory.[1] In contrast, some simulation work sets up extremely simple lifestyles for the agents

[1] However, we should note in passing that the bulk of the complexity in de Bourcier and Wheeler’s model is built

involved, such as Oliphant (1996). Oliphant models an abstract situation in which one agent, the sender, gains access to a hidden state that can take two possible values and must then make one of two "signals", while another agent, the receiver, must try to choose a behaviour that matches the hidden state. (This is known as the minimal signalling game and is illustrated in Figure 7.2; see Hurd (1995) for an example of its use in theoretical biology.)

Secondly, the work may have strong or weak links to biology. Bullock (1997b), for example, constructed a general model of sexual signalling, in which males of varying quality solicited females for a favourable response. Bullock's model was explicitly designed to test the ideas of Zahavi, and specifically to see whether a game-theoretic model of the handicap principle developed by Grafen (1990a) was as general as its author believed it to be (see section 2.4.3). On the other hand, artificial life has produced a lot of work like that of Ackley and Littman (1994), who built an elaborate scenario for their agents that involved movement along parallel tracks and the possibility of signalling to each other about the presence of food or predators at each end of the track. The agents who participated in this game were also involved in a higher-level population structure; they lived in small local colonies, and there was periodic random migration between the colonies. Ackley and Littman found that altruistic signalling evolved in this context and assumed that it was due to the effects of kin selection: agents within a colony were likely to be related to each other and thus it was in their genetic interests to co-operate. While the work borrows certain concepts from biology, such as the idea of migration between local populations, and kin selection, no attempt is made to test specific biological theories.

Other work on the origin of signalling systems includes Collins and Jefferson (1991), who attempted to evolve communication via pheromones between ant-like agents, and Werner and Dyer (1993) who covered communication tangentially in a model of herding behaviour. Robbins (1994) added parasitism to Werner and Dyer's (1991) model and argued that the resulting communication systems were more robust; Saunders and Pollack (1996) examined the evolution of communication over continuously-valued channels (as opposed to the discrete signals used in most other work) and found that agents evolved to use the less noisy of two available channels. Werner (1996) used a model of sexual signalling to examine the plausibility of the runaway process (see section 2.7); Werner and Todd (1997) simulated the evolution of male advertisement signals such as bird song that do not index quality but serve solely to manipulate female responsiveness. Di Paolo (1997b, 1997a) looked at the evolution of co-ordinated action in a spatially arranged population.

Linguistic models

There also exists a body of research with chiefly linguistic goals: some authors have used artificial-life methods as a way of building simple models of features of human language. For instance, Batali (1994) looked at the interaction between evolution and learning in the acquisition of syntactic communication, and Steels (1995) built a model in which the vocabulary of a population of linguistic agents could "self-organize", spontaneously building up initial linkages between symbols and their referents, and absorbing the influx of new, novice speakers (see also Steels, 1996a, 1996b; Steels & Vogt, 1997). In some of these models it is the communication scheme itself

into the agents. For example, the way they respond to visual stimuli is specified by the designers; the agents also have a motivational system that imposes periodic aggression. Thus not much of their behavioural strategy is actually open to variation and selection. Indeed, the only parameter that *was* allowed to evolve was the degree to which an agent exaggerated its level of aggression.

that evolves, rather than the abilities or strategies of the agents—and sometimes both, as in Kirby and Hurford's (1997) simulation, which the authors use to support an argument that postulating a Chomskyan "language acquisition device" is not a necessary step in explaining human linguistic competence. Other work in this class includes Hashimoto's (1997) model of the evolution of grammar, and de Boer's (1997) work on vowel systems.

Work of this kind is certainly interesting, and its existence helps to show that the bottom-up simulation methods of artificial life can have a wider application than biology. However, to the extent that these models are pitched at learned, syntactic communication schemes, and assume a prior motivation for the agents to communicate with each other, they are not relevant to our inquiry into the selective pressures that shaped the origin of simple signalling systems.

Communication in the service of robot co-operation

Finally, there has been some work on the engineering of optimal communication schemes for use by co-operating social robots, e.g., Yanco and Stein (1993), Mataric (1994), Dautenhahn (1995), and Moukas and Hayes (1996). Mataric, for instance, discusses the problems involved in communication between mobile robots that must co-operate in a puck-collection task. Some of this work does not actually involve the *evolution* of a communication scheme; rather, one is deliberately engineered by the designer. However, the aims and methods are consistent with the artificial life approach: global co-ordination emerges from simple local (i.e., single-robot) behaviours.

Like the linguistic work, this material is not strictly relevant to our investigations. However, it is mentioned here because the attempt to look at how one robot might observe the communicative behaviours of another, and through this learn the other robot's "language" (see Moukas & Hayes, 1996), is suggestive of some important issues that we will not be able to do justice to in this thesis. That is, such work forces us to question our assumption that we can model communication by looking at the evolution of strategies for participating in fixed, game-like interactions. How is one robot to know that the other's behaviour should be interpreted as a signal, for instance? How long does a signal last? When does one end and another begin? The behaviour of real-world agents like animals and robots occurs continuously, and is not neatly broken up into question and answer or signal and response. The decision to nevertheless use fixed, game-like scenarios as models for the evolution of communication will be defended in section 6.2.

Exceptions

There are some exceptions to the broad classification scheme outlined above. MacLennan and Burghardt (1994) and Oliphant (1997) are certainly concerned with the evolution of simple, genetically specified signalling systems, but both papers also attempt to investigate learned communication. Saunders and Pollack (1996) are apparently interested in engineering applications as well as in simulating an aspect of animal communication. Di Paolo's (1997a, 1997b) goals are also more ambitious than simply presenting a model in which communication evolves; he criticizes the idea that communication should be defined in terms of information transmission and instead argues for a definition in autopoietic[2] terms. For Di Paolo, communication is synonymous with co-ordinated behaviour, and his model demonstrates the evolution of something like signalling in a context in which both agents share all the relevant information. Finally, Steels and Vogt (1997)

[2] Autopoiesis is a very general biological theory of self-organization and self-maintenance; see Maturana and Varela (1980) for an introduction.

employ real robots rather than the simulation models more typical of linguistically oriented work.

5.1.2 Problems with the earlier work

It will be obvious from the lengthy arguments put forward in chapter 4 that the simulation dimension suggested above—that of greater or lesser contact with biological theory—is not intended to be value-neutral. The claim is that those simulations that build on models and theories from theoretical biology will be much more likely to result in useful findings. However, artificial-life enthusiasts may still be sceptical. What is wrong with an original model of the evolution of communication? Might it not lead us in novel directions that biological theory has failed to anticipate? Let us consider the case of Ackley and Littman's (1994) simulation. One of their conclusions is that what they called "festival" migration was more effective in promoting kin-selected altruism[3] than "wind" migration: festival migration involves the mutual exchange of individuals between groups of four adjacent colonies, whereas wind migration means that, across the grid, some individuals from every colony move to the next colony downwind. The question that Ackley and Littman must answer boils down to "so what?". The elaborate local scenario of movement on parallel tracks, the arrangement of colonies of agents on a grid, the decision not to incorporate genetic mutation, and the mechanics of the two alternative migration methods are all so far removed from reality that it seems disingenuous to even try to invent any implications or predictions about animal behaviour from this work. We may take Ackley and Littman at their word when they tell us that one migration paradigm was more conducive to kin selection than another (although strictly speaking we should probably replicate their simulation in case this was an artefactual result) but they have shown only that using one procedure over another leads to a different outcome in their program. In this sense their result is a mere "proof of concept"; they establish that in the case of their particular simulation environment, migration methods can make a difference. Showing that anything follows from this is another matter entirely.

Compare this with the conclusions we can draw from Bullock's (1997b) more biologically oriented work. Bullock tells us that if signallers and receivers are in a situation in which signallers always want positive responses and receivers want to respond positively only to high-quality signallers, then a handicap signalling equilibrium can be achieved if certain relationships hold regarding the cost and benefit functions for signallers. Other, non-costly signalling equilibria may also be possible under certain circumstances. The application of Bullock's result to sexual signalling in a real species would be far from easy, because of the general problem of assessing fitness costs and benefits in complex real-world ecologies, but it is clearly telling us something about the world. It has a potential for falsification that Ackley and Littman's (1994) simulation seems to lack. It is also true that Bullock's work contributes to a broader theoretical picture; as Bullock himself makes clear, the results of his simulation help to correct and contextualize earlier work by Zahavi (1977), Grafen (1990a) and Hurd (1995).

It is hoped that the advantages of a connection with biological theory will by now be obvious.

[3] Di Paolo (1997a) points out that Ackley and Littman do not formally demonstrate that kin selection has affected the course of evolution in their models. They simply assume that kin selection can be equated with spatial arrangement, in which an agent interacts with neighbours who are likely to be relatives. Oliphant (1996) makes the same unsupported assumption. Di Paolo presents an argument that spatial arrangement alone can explain altruistic communication in some circumstances.

Note that the critique of MacLennan and Burghardt's work presented in the second half of this chapter is in some ways a cautionary tale about what can go wrong when artificial life drifts too far from the project of modelling some aspect of reality.

5.1.3 Common and conflicting interests

There is actually a third dimension on which we can classify simulation work on simple signalling systems: the degree to which the agents in the simulation have congruent or conflicting interests. Many of the earlier models have been constructed such that signalling is in the interests of both signallers and receivers—any communication systems that evolve can therefore be described as co-operative. For example, Werner and Dyer (1991) postulated blind, mobile males and sighted, immobile females: the evolution of a signalling system was in the interests of both parties as it allowed mating to take place at better-than-chance frequencies. In MacLennan and Burghardt's (1994) model, signallers and receivers were rewarded if and only if they engaged in successful communicative interactions.

Other models (Ackley & Littman, 1994; Oliphant, 1996) have looked at the special case where communication would benefit receivers, but the potential signallers are indifferent. Oliphant argues that this is a good way to model the evolution of alarm calls: it captures the idea that the potential signaller already knows about the danger of the approaching predator, and tests the stability of a strategy of sharing that information. In fact the models suggest that signalling will not evolve in these cases unless a mechanism such as reciprocal altruism or kin selection is in place. (Note that such mechanisms have no mystical effect: they simply shift the expected long-term inclusive-fitness payoffs for particular strategies such that communication is mutually beneficial.)

Finally, some work considers the evolution of communication in situations where the two parties appear to have conflicting interests, e.g., de Bourcier and Wheeler's (1994) model of aggressive territorial signalling, or Bullock's (1997b) model of insistent signallers and choosy receivers.

It will be argued in chapter 6 that simulations in which communication evolves when it is clearly in the interests of both parties to exchange information do not necessarily add much to our understanding. This is because they are entirely in keeping with the basic finding from game theory that in a co-ordination game (which is what most signalling games essentially are) a co-ordinated equilibrium will likely be achieved if the payoffs for both players favour it (Binmore, 1992). In the "which side of the road shall we drive on?" game, for example, drivers in particular countries settle on a stable strategy in which they all stay on one side of the road. However, this fact is not particularly interesting or surprising, because the payoff to all players for arriving at a co-ordinated strategy is much more attractive than the cost of constant head-on collisions. Similarly, we should not be surprised to observe that simulated animals come to co-ordinate their behaviour when the payoffs for co-ordination are high. Therefore the simulation models presented in this thesis will generally reflect contexts like aggressive and sexual signalling in which the signallers and receivers have conflicting interests.

It should be noted that we skate on the edge of a contradiction here: the definition of communication as proper signalling (see section 3.3) strongly suggests that communication will only occur when it is in the interests of both parties. Having adopted the adaptationist perspective, i.e., viewing evolution as a long-term optimizing or satisficing process, it is difficult to maintain that it could ever be evolutionarily stable for an organism to participate in a communication scheme that was against its genetic interests. This logic, in its turn, starts to suggest that the ultimate message of the thesis must be the banal “communication only happens when it is in the interests of both parties to communicate”. In a sense, this *is* and must be the message of the thesis—the adaptationist stance demands it.

However, there are two defences against a charge of circularity or theoretical emptiness. Firstly, the chequered history of the handicap principle, for example, and the current consensus that handicap signalling equilibria exist in nature, show that the phrase “in the interests of” can be cashed out in highly counter-intuitive ways—the fact that so many eminent biologists now appear to have been wrong in their initial rejection of the handicap principle should stand as a salutary warning for the rest of us. Thus, a program of looking for evolved communication in contexts where the two agents in an interaction *appear* to have conflicting interests is not paradoxical; we are seeking potentially complicated and surprising routes by which a communicative strategy might turn out to be to their mutual advantage. Secondly, the potential that evolutionary simulation models have for tracing the dynamics of evolution allow us to investigate hypotheses about *unstable* communication: the possibility that a population of animals could get temporarily stuck in communicative strategies that are not in their long-term interests, but from which they cannot easily escape.

5.2 MacLennan and Burghardt’s work

MacLennan and Burghardt’s (1994) paper has been singled out for such close attention here because it is representative of work in artificial life on the evolution of communication. The paper on which it is based (MacLennan, 1991) came relatively early in the recent explosion of work in this area, and has therefore influenced much of what has been done since.

5.2.1 Description of the simulation

Justification

MacLennan and Burghardt describe their method as *synthetic ethology*, contrasting it explicitly with simulation in Maynard Smith’s (1974a) or Smithers’s (1994) sense of detailed modelling. They state that:

> Our goal in these experiments was to design a synthetic world that was as simple as possible while still permitting communication to evolve. (MacLennan & Burghardt, 1994, p. 165)

MacLennan and Burghardt repeatedly emphasize that their “synthetic world” is not supposed to reflect any real environment, nor are their simulated organisms like any actual species. Inspired by the synthetic psychology of Braitenberg (1984), they hoped that, in comparison with empirical ethology, their stripped-down approach would be “more likely to suggest behavioral laws of great generality” (MacLennan & Burghardt, 1994, p. 163). While the term “synthetic” is being used here in the sense of “constructed”, it is clear from this quote that MacLennan and Burghardt are evasive about whether and how they are modelling anything. Finding “behavioral laws of great generality” is a consummation devoutly to be wished, but observing a synthetic world degenerates all too easily into what Fontana, Wagner, and Buss (1994) called “digital naturalism”—the pointless amassing of facts concerning the behaviour of a simulation.

MacLennan and Burghardt were aware of the difficulty of defining communication, and of the problem of imputing intentionality. They adopted Burghardt’s (1970) definition of communication, which “finessed the issue of intent by the requirement that the behavior be likely to influence the receiver in a way that benefits, in a probabilistic manner, the signaler or some group of which it is a member” (MacLennan & Burghardt, 1994, p. 163). We have seen in chapter 3 that such behaviour-based definitions of communication are ultimately unsatisfactory. To be fair, however, in the circumstances of MacLennan and Burghardt’s simple synthetic world, this inadequate definition does them no great harm.

MacLennan and Burghardt chose to investigate co-operative communication. They reasoned that for communication to be selected for, some of the simulated organisms must have access to information that the others in the group do not, otherwise communication would be unnecessary. The non-shared information must also be of environmental significance; it must be worth talking about. In line with their definition of communication, they designed the synthetic world such that communicating this non-shared information would tend to confer a selective advantage.

Method

MacLennan and Burghardt used populations of simulated organisms that they refer to as “simorgs”. The simorgs all have access to a shared global environment, and each individual has access to a private local environment. The global environment provides a medium for communication, and the local environments are a source of significant information that the simorgs may evolve to communicate *about*. Each of the environments is as simple as possible, represented by a single variable that can take on a finite number of values. It is emphasized that “there are no geometrical relations among [the simorgs]. . . they are not in a rectangular grid, nor are some closer than others” (MacLennan & Burghardt, 1994, p. 166).

MacLennan and Burghardt suggest, by way of analogy, that the global environment can be thought of as the air, capable of transmitting only one sound at a time, and the local environments can be considered exclusive hunting grounds, into which different species of prey may wander. In other words, states of the global environment have the potential to be exploited as signals, and states of the local environment are particular circumstances that it will pay simorgs to signal about. The global environment thus provides the potential for intentional icons and proper signalling.

Simorgs have only two classes of behavioural choice open to them: they can *emit* a signal (into the global environment), or they can *act* in an attempt to respond to the signal of another. The state of the global environment can be changed by any of the simorgs if that simorg emits a signal when its turn comes; the states of the simorgs’ local environments are not under their control, and are periodically reset to random states.

In the synthetic world, simorgs achieve fitness by successfully co-operating with another simorg: specifically, by responding to a signal with an action that matches the local environment state of the signaller. When this occurs, both the signaller and the respondent are rewarded with a point of fitness. Continuing their analogy, MacLennan and Burghardt suggest that this is to be regarded as two hunters bringing down a prey animal that neither could bag alone. Assuming that successful communication has taken place, note that the signal does not mean "I've got some prey here", but "I've got prey of type λ here; would you mind helping out with action-λ?" The state of another simorg's local environment is not directly knowable, and successful co-operation can only come about through a lucky guess or the employment of communication.

In order to implement their ideas in a computer program, MacLennan and Burghardt had to make a number of somewhat arbitrary practical decisions. Thus, time in the synthetic world is discrete. Once each time step, the simorgs respond (i.e., act or emit) in a fixed order; effectively they are arranged in a ring. The program keeps track of the "owner" of the symbol currently occupying the global environment. It is possible, for example, for one simorg to emit and then earn several fitness points consecutively as a series of other simorgs act in response to the same persistent signal.

Every five time steps (one environment cycle) each local environment is reset to a random value, ensuring that the simorgs must react to changing circumstances if they are to succeed. Every fifty time steps there is a breeding cycle: two fit simorgs are stochastically selected as parents and, using two-point crossover with a small chance of mutation, a new simorg is generated. An unfit simorg is stochastically selected to be replaced by the child, keeping the population size constant. This arrangement more or less constitutes a steady-state genetic algorithm.

The experiments reported were run for 5000 breeding cycles, populations were of size 100, there were eight local environment states (L) and eight global environment states (G)—"just enough possible sounds to describe the possible situations" (p. 175)—and the mutation rate was a 0.01 probability of one mutated allele per birth.

Finite state machines (FSMs) serve as the internal architecture of the simorgs. MacLennan and Burghardt could have used any number of architectures, and considered using neural networks, but settled on FSMs because they "are both readily understood intuitively and easy to represent in genetic strings for simulated evolution" (p. 167). In the experiment described[4] the FSMs were of only one state, which reduces to a look-up table. The response a simorg would make at any one time step was completely determined by the state of the global environment and the state of its local environment. The content of each of the 64 (8×8) entries of the look-up table was a flag indicating *act* or *emit*, and an integer representing the action-type or the emitted symbol respectively. The genetic coding of the simorg was a direct mapping of this structure; i.e., there was no distinction between genotype and phenotype.

MacLennan and Burghardt included in the program a mechanism to (optionally) prevent communication from occurring: the global environment could be overwritten with a random symbol after the response of each simorg. Their logic was that if fitness increased more rapidly when communication was permitted, compared with when it was blocked, then "true communication... involving a sender" (p. 172) was taking place. In a similar fashion they were interested in exploring the effect of a simple learning rule, whereby a simorg that makes an incorrect action (i.e. an action that does not correspond to the local environment state of the last emitter) in response to a signal has the appropriate entry in its look-up table altered so that it *would have* given the correct response. Thus, they report the results of subjecting the same randomly generated

[4] MacLennan and Burghardt actually conducted two experiments; we will focus entirely on the first. Experiment 2 was an attempt to evolve multiple-symbol communication and the results led them to conclude that "making the step to multiple-symbol syntax is evolutionarily hard" (MacLennan & Burghardt, 1994, p. 183).

initial population to each of the following experimental conditions:

$C^{-}L^{-}$ communication blocked and learning disabled;

$C^{+}L^{-}$ communication permitted and learning disabled;

$C^{+}L^{+}$ communication permitted and learning enabled.

In each of the conditions, they collected data on mean fitness over time. They also constructed a "denotation matrix", which recorded the number of successful communication events, arranged in a table by local and global environment states. They found that these matrices were most useful when tallied over the last 50 breeding cycles of a 5000-cycle experimental run. Under these circumstances, the matrix was interpreted by MacLennan and Burghardt as describing the evolved "language" of the simorgs. The degree of structure present in the matrix was indexed by co-efficient of variation and entropy statistics.

Results and conclusions

MacLennan and Burghardt report that communication did indeed evolve in their synthetic world. The results reported are for a single random initial population subjected to each of the three conditions. MacLennan and Burghardt assure us that these results are typical, although clearly the presentation of results averaged over a number of runs would have been an improvement. In the $C^{-}L^{-}$ condition, there was only a very slight increase in fitness over the length of an experimental run, whereas in the $C^{+}L^{-}$ condition the rate of fitness increase was an order of magnitude greater. In the $C^{+}L^{+}$ condition, the rate of fitness increase was higher still. MacLennan and Burghardt conclude that, when it is not suppressed, communication is selected for and leads to higher levels of co-operation. The provision of the single case learning rule further increases the effectiveness of the communicative strategy.

Analyses of the denotation matrices showed that in the $C^{-}L^{-}$ condition, the pattern of symbol use was almost random. When communication was permitted the matrices were quite structured, as measured by the entropy statistic. Visual inspection of the denotation matrices made it clear that certain symbols had evolved to (almost uniquely) represent certain local states. There was ambiguity in two senses: sometimes a symbol would be used to represent two or more states, and sometimes a state was represented by two or more symbols. MacLennan and Burghardt suggest that the ambiguity is either due to two subpopulations using different symbol dialects, or to individual simorgs using one symbol to represent two different states.

That there should be any fitness increase at all in the $C^{-}L^{-}$ condition is not obvious. MacLennan and Burghardt refer to this phenomenon as "partial cooperation through co-adaptation", and regard it as a "low-level effect" (p. 185). They explain it by noting that simorgs can do better than chance if they emit a symbol only in a subset of their local situations, and guess actions within that same subset.

5.2.2 A replication

MacLennan and Burghardt's experiment was replicated; the program was based on the published descriptions of their procedure (MacLennan, 1991; MacLennan & Burghardt, 1994). The replication gave qualitatively similar results, in that fitness improved over time when communication

		MacLennan & Burghardt	Replication results		
			Mean	*SD*	*p*
Fitness increase	C^-L^-	0.37	0.99	1.16	n.s.
	C^+L^-	9.72	14.6	6.54	n.s.
	C^+L^+	37.1	10.6	10.6	0.025
Final mean fitness	C^-L^-	≈ 6.6	6.74	0.43	n.s.
	C^+L^-	10.28	12.71	2.68	n.s.
	C^+L^+	59.84	46.13	4.02	0.004

Table 5.1: Rates of fitness increase (determined by linear regression and measured in units $\times 10^{-4}$ breeding cycles) and final mean fitness scores. Note that mean fitness data was a moving average smoothed over 50 breeding cycles, and that final mean fitness in the C^+L^+ condition is much higher because the simorgs had four chances per environment cycle to respond after correction by the learning rule: fitness scores in this condition start at 40+ rather than the usual chance level of 6.25. Rates of increase are thus a better comparison across conditions.

was enabled, and structure developed in the denotation matrices, but the specific results in the three conditions were not reproduced. Table 5.1 contrasts MacLennan and Burghardt's results with those of the replication; the rate of fitness increase per 10^4 breeding cycles and the mean final fitness are shown. MacLennan and Burghardt's results refer to the single run they presented as the typical case. The replication results show the mean and standard deviation across 20 runs with different random seed values. For each condition, the column labelled "p" shows the statistical significance of a two-sample *t*-test of the null hypothesis that MacLennan and Burghardt's result could have come from the same distribution as the replication data ("n.s." means not significant, i.e., $p > 0.05$).

The C^-L^- and C^+L^- conditions showed slightly higher rates of fitness increase in the replication. More importantly, the rate of fitness increase in the C^+L^+ condition was more than three times *smaller* than in MacLennan and Burghardt's data, and this was statistically significant. The replication results do not support MacLennan and Burghardt's finding that the C^+L^+ condition, i.e. communication with learning, leads to the highest rate of fitness increase. The replication suggests that communication with learning to is inferior to communication alone, in terms of the rate of fitness increase.

Table 5.2 shows the entropy of the denotation matrices over the last 50 breeding cycles of the experimental runs. Again, MacLennan and Burghardt's figures are taken directly from their paper and describe a single run, while the replication results summarize 20 different runs. In the C^-L^- condition we find significantly *more* structure to the denotation matrices than did MacLennan and Burghardt, and in the C^+L^+ condition we find significantly *less*. Instead of the lowest entropy being associated with C^+L^+, it is associated with C^+L^-. In other words, the most structured communication conventions develop in the communication-only condition, and the addition of the learning rule only reduces that structure.

The differences between these findings and those of MacLennan and Burghardt should not be

		MacLennan &	Replication results		
		Burghardt	*Mean*	*SD*	*p*
Entropy	C^-L^-	5.66	4.96	0.15	< 0.001
	C^+L^-	3.95	3.36	0.50	n.s.
	C^+L^+	3.47	4.45	0.36	0.015

Table 5.2: Entropy statistics, calculated on the denotation matrix of the final 50 breeding cycles of the experiment. An entropy value of 6 would indicate a completely random matrix. A value of 3 indicates a perfectly structured matrix, with one symbol per situation.

exaggerated. In all measurements, across all conditions, the replication data was well within an order of magnitude of MacLennan and Burghardt's figure. The interpretation given here of their experimental method may not reflect *exactly* their actual procedure, but the nature of any discrepancy has proved elusive. MacLennan and Burghardt's central result was successfully replicated: that communication, when enabled, leads to relatively high rates of fitness increase, and to the evolution of a structured "language" as evidenced by the denotation matrix. However, the failure of the replication to produce more closely matched results highlights an inter-subjectivity problem with simulations: it is easy to make alternative interpretations of the written description of a simulation, and that is what seems to have happened here. This could leave us wondering which of the two programs—original or replication—to take more seriously, except that in this case the results from an independent replication performed by the second author of Noble and Cliff (1996) agree with the work presented here in disagreeing with MacLennan and Burghardt's figures.

Nevertheless, it is important to recognize that attempting to circumvent the issue by simply obtaining MacLennan and Burghardt's original simulation code would not have solved anything, as the original code may contain artefacts that we would not wish to reproduce. This kind of problem mitigates against the claims of simulation partisans like Taylor and Jefferson (1994) and even Miller (1995)—see chapter 4—that simulations, with their requirement that every last detail must be specified in order to create a running computer program, are more explicit than equational models.

5.2.3 Extension and critique

Having described the methods used by MacLennan and Burghardt, and noted the degree to which the replication results match those of the original work, we can now comment critically on certain aspects of their experiment.

No geometry?

MacLennan and Burghardt claim that there are "no geometrical relations" (1994, p. 166) among the simorgs. This is in keeping with their goal of constructing a synthetic world that is as simple as possible while still permitting communication to evolve. If the simorgs were arranged on a toroidal grid and could communicate only locally, for example, this would certainly complicate things.

However, in the current set-up, the simorgs are effectively arranged in a ring. As MacLennan and Burghardt put it, "The simorgs react one at a time in a fixed order determined by their position

		Mean	*SD*	Effect
Fitness increase	C^-L^-	0.94	1.52	-4.5%
	C^+L^-	18.6	7.05	$+27.4\%$
	C^+L^+	33.7	13.8	$+218\%$
Final mean fitness	C^-L^-	6.76	0.53	$+0.23\%$
	C^+L^-	14.47	2.83	$+13.9\%$
	C^+L^+	22.24	5.21	-51.8%

Table 5.3: Effect of random-order updating. Rate of fitness increase $\times 10^{-4}$ breeding cycles (determined by linear regression), and final mean fitness scores are shown, with means and standard deviations across 20 runs. The “effect” column compares the random-order results with the standard updating results from the replication (see table 5.1); note that if the updating method was not influencing the results, we would expect this value to be close to zero.

in a table.” Thus there is at least a topology, if not a geometry: simorgs will tend to receive signals from their immediate neighbours in one direction, and send signals to their neighbours in the other direction. The experiment could have been performed without this modest topological assumption if the simorgs were updated in a different random order at each time step. The replication program was therefore modified to use just such an updating procedure. Table 5.3 shows the rates of fitness increase and final fitness scores under this method.

There is a dramatic difference between the two updating methods. In the communication only (C^+L^-) and no communication (C^-L^-) conditions, performance does not differ greatly across the two updating methods. The effect of the learning rule, on the other hand, depends very much on the updating method used: under random-order updating, the rate of fitness increase is much higher. This recalls the similarly drastic difference observed when Huberman and Glance (1993) applied an asynchronous updating method to Nowak and May’s (1992) spatial Prisoner’s Dilemma model. The use of fixed-order updating is common in artificial life simulations, possibly because it is an idea that comes naturally to computer programmers. However, random-order or asynchronous updating is quite easily implemented, and, all things being equal, is more likely to be an effective model of a real-world situation: many events in the natural world can be modelled as Poisson processes, in which the likelihood of an event occurring is unrelated to the time that has elapsed since its previous occurrence. Applying this to communication, it is much more plausible to suggest that senders and receivers will encounter each other at random time intervals than to build a model in which everyone waits their turn.

The difference in results across the two updating procedures demonstrates that MacLennan and Burghardt’s results are dependent upon such apparently minor assumptions built into their program. Their goal is to uncover general laws that can be translated back into the realm of real biology, but if the effect of learning on the evolution of communication is dependent on the updating method used, it is difficult to know what biological conclusions should be drawn. Does learning facilitate the development of a communicative system, or doesn’t it?

Dialects or sub-optimal look-up tables?

MacLennan and Burghardt, noting the ambiguous symbol use evident in the denotation matrices, comment that “we cannot tell from [the denotation matrix] whether this multiple use of symbols results from two subpopulations or from individual simorgs using the symbol to denote two situations.” (1994, p. 179). The idea that there could be subpopulations using different dialects seems quite plausible, especially given that the topology of the simorgs’ environment ensures that they will only be communicating with near neighbours. One can imagine a series of simorgs using variant communication scheme A in one section of the ring, shading gradually into variant B in the opposite section, and back again.

MacLennan and Burghardt claim that the facts of the matter could easily be uncovered: given that the underlying finite state machines are available in computer memory, “there need be no mystery about how the simorgs are communicating, because the process is completely transparent” (p. 179). However, they do not follow their own advice, making no attempt to analyze their data in this regard. They make no clear statement as to whether they in fact believe there are two or more subpopulations using variants of the evolved “language”. MacLennan, in his earlier paper, is less conservative: “the differing use of symbols in various contexts makes it quite possible for every simorg to be using a different dialect of the ‘language’ manifest in the denotation matrix.” (MacLennan, 1991, p. 653).

In an attempt to resolve this question, a convergence statistic was used in the replication. Each position on each simorg genome was examined in turn, and the mean percentage of identical entries across the population was calculated. Thus, a convergence statistic of 100% would indicate a population of simorgs with identical genomes and, thus, identical FSMs. In runs of 5000 breeding cycles duration, the final convergence statistic was typically between 75% and 85%. This is not conclusive: it means that up to 25% of the simorgs could have been different from the norm, or that 25% of the genetic material of each simorg could be unique, and so leaves plenty of room for the possibility of different dialects. However, when the runs were extended to 2×10^4 breeding cycles or more, final convergence statistics in the C^+L^- condition were approximately 99.5%, and denotation matrices were qualitatively similar, i.e., they still showed ambiguous communication. It is implausible to suggest that there might be different dialects when the simorgs in a population are 99.5% identical to each other. We must conclude that the suggestive ambiguity in the denotation matrices is nothing more than the net effect of the whole population using a single, inefficient “language” that sometimes represents a state by more than one symbol, or uses one symbol to denote more than one state.

Despite their stated wariness about adopting any sort of intentional stance towards the simorgs, MacLennan and Burghardt are not immune to the temptation to think of them as intentional agents. In this case, that temptation has led them astray. Language without the scare quotes is undoubtedly the exclusive province of sophisticated intentional agents (argued in, e.g., Bennett, 1976; Dennett, 1987, see section 3.2.4), but having drawn the analogy between human language and simorg communication, MacLennan and Burghardt were too ready to suspect that, like real language users, simorgs might have dialects.

Consequences of the FSM look-up table approach

Imagine for a moment that you are a simorg. Disregarding the fact that simorg "decisions" are entirely determined by the look-up table, imagine that you have decided to emit a symbol. The only context that is important to you is the state of your local environment: you need to choose the right symbol to describe your situation, according to the "language" conventions that have developed. The identity of the symbol in the global environment is unimportant, because you're going to overwrite it anyway. Similarly, if you're going to act, you don't care about the state of your local environment; you only want to interpret the global symbol in such a way as to correctly match the environment of the last emitter, and thereby score a point of fitness.

For the real simorgs of MacLennan and Burghardt, things are not this simple. There is no prior decision to emit or to act, only the consultation of a table with an entry for every possible combination of local and global environment states. As MacLennan and Burghardt put it, "finite-state machines have a rule for every possible condition." (1994, p. 168). The FSM architecture therefore makes evolving a communication system harder for the simorgs than it might be given some other control architecture. For example, if during a particular run it became advantageous to reliably perform action-2 in response to symbol-7, FSM-controlled simorgs would have to ensure—through evolution or learning—that eight distinct entries in their look-up table came to be identical. That is, they would need to perform action-2 in response to symbol-7 in the context of eight different possible local environment states. By contrast, a simorg that was controlled by, for example, a classifier system (Holland, 1975; Wilson, 1995) would need only to generate a single production rule: "perform action-2 in response to symbol-7". Although it has not yet been investigated, we must suspect that simorgs controlled by classifier systems would evolve significantly faster than FSM-controlled simorgs.

MacLennan and Burghardt did not believe that FSMs were the only architecture open to them, and adopted them for pragmatic reasons. In fact they appear to have been under the impression that the very unwieldiness of FSMs in this context would help to shield them from claims that communication had come too easily to the simorgs. However, if an arbitrary choice of control architecture is influencing the results in unexpected ways, it is again difficult to see how MacLennan and Burghardt's conclusions can be reliably translated back to biology.

Counter-intuitive optimal strategy

The optimal strategy for the simorgs, at least at the population level, must be to act as often as possible, and to emit as infrequently as possible. This is because emitting scores no fitness points directly. The best way for the simorgs to achieve this is to build up a link between a *single* global symbol γ and a *single* local state λ. A situation develops where simorgs always "blindly" respond with action-λ, unless they are in state λ themselves, in which case they emit γ. Imagine all the simorgs acting in this way: it is clear that they would no longer need to be concerned about the particular identity of the symbol in the global environment. They would *know* that it will always be γ, and that it will always reliably indicate that the last emitter (whoever that may be) is in state λ.

Assuming 8 symbols and 8 local states, this means that an episode of successful communication will take place $7/8$ of the time[5]. This would translate as a mean fitness of 87.5—a very high

[5] Discounting for a moment the unfortunate simorg who acts in response to an out of date symbol immediately after the local environments have been randomly reset; in the *Umwelt* of the simorgs, this is an infrequent event.

value relative to the results presented in tables 5.1 and 5.3. In general the maximum fitness will be equal to $100 \times \frac{L-1}{L}$, where L is the number of local states. In exploratory simulations using the replication code, this phenomenon has been observed to evolve spontaneously only for $L \leq G \leq 4$, but the principle remains.

The trouble with this result is that one presumably does not want to call it an evolved communication system or "language", even though the simorgs are ostensibly fitter than ever before. If the global environment is (almost) always in the same state, it is difficult to describe it as carrying any information. The simorgs in such a situation appear to be exploiting a loophole in the experimental design. Their behaviour certainly does not qualify as communication because there is no mapping relation between signals and any state of the world. The behaviour reduces, in Millikan's terms, to a tacit supposition that others in the environment will be following the same strategy.

MacLennan and Burghardt were aware of this possibility (see section 5.2.1). They saw it as most relevant to the C^-L^- condition, in that it provided an explanation for the otherwise mysterious increase in fitness observed. MacLennan (1991, p. 653) felt that "in most cases [it] is a low level effect that is unintrusive and can be ignored". But the existence of this strategy is either an unexpected finding or, more likely, a rather alarming artefact: if we translate it back into real-world terms, the suggestion is that communication schemes will evolve to use fewer and fewer symbols before settling on only one.

Fewer symbols: faster improvement

The point outlined above has a number of implications. Given that the optimal strategy involves the utilization of only one symbol, it can be hypothesized that giving the simorgs progressively *fewer* symbols to work with should steer them towards that strategy and thus *improve* their performance. This contrasts with the intuitive hypothesis that n local states will require simorgs to use n symbols to denote them. MacLennan and Burghardt seem to have assumed the truth of the intuitive hypothesis: they speak of the ideal denotation matrix as having one symbol to denote each situation, and refer to the fact that $L = G$ as meaning that "there were just enough possible sounds to describe the possible situations." (p. 175).

To test the fewer-symbols hypothesis, the C^+L^- condition was employed, with the number of local environment states held constant at $L = 8$. The number of global environment states G, i.e., the number of possible symbols, was systematically varied from eight down to one. For each case, 20 runs of length 5000 breeding cycles were conducted. The results are shown in Figure 5.1. Overall, higher rates of fitness increase were associated with smaller numbers of symbols. These results certainly contradict the intuitive view. In a *post hoc* effort to make a connection with real-world biology, one might argue, for instance, that the result demonstrates the principle that small signal repertoires enhance the detectability of ritualized signals (Wiley, 1983). The argument would be without merit, however. A small signal repertoire is a means of enhancing signal detection *in a noisy environment*. The simorgs' environment has no noise, and their perception of symbols is direct, immediate and reliable. Again, there is no easy biological translation of this observation concerning MacLennan and Burghardt's synthetic world.

Symbol use over time

If the simorgs were evolving a "language", with an eventual one-to-one correspondence between global symbols and local states, we should observe a fairly even distribution of the G symbols.

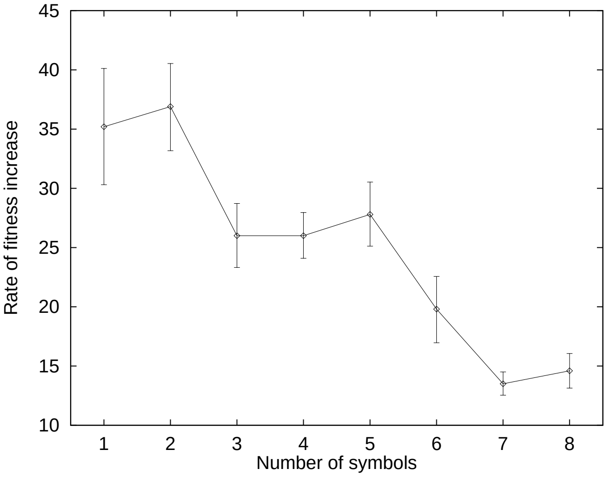

Figure 5.1: Mean effect on rate of fitness increase of varying G, the number of global symbols, while L, the number of local states, $= 8$. Error-bars represent ± 1 standard error. Rate of fitness increase measured in units $\times 10^{-4}$ breeding cycles and determined by linear regression.

That is, simorgs should use each symbol about equally often. This follows from the fact that the distribution of local states is random and therefore uniform. This is not observed, however. Popular symbols tend to get more popular; in the 20 runs of the $C^{+}L^{-}$ condition reported in table 5.1, the mean usage of the most popular symbol at the end of the run was 41.45%. When longer runs were conducted, the popularity of the most popular symbol was even higher. Infrequently used symbols often dropped out of use altogether. This seems to indicate that the simorgs are drifting towards the optimal strategy described above.

5.2.4 Lessons to be learned

MacLennan and Burghardt clearly succeed in establishing that communication can evolve in a computer simulation, at least given their particular simulated environment and specific architectures for the simulated organisms. They express the hope that their work will suggest general laws or principles concerning animal communication, but they are aware that “if the synthetic world is too alien, we may doubt the applicability to our world of any observations made of the former.” (1994, p. 166). In section 5.2.3, it has been shown that, in various ways, their synthetic world is indeed alien. Regrettably, it *is* difficult to see how certain aspects of MacLennan and Burghardt’s results could be translated into real-world biology. In other words, their simulation is not a model.

MacLennan and Burghardt were trying to do a number of things at once. Primarily, they were attempting to provide an existence proof that communication can evolve in simulation, and they make no secret of having constructed the synthetic world so that the simorgs will be likely to reproduce only if they co-operate (i.e., communicate) in the specified way. Given the historical context of MacLennan’s (1991) original work, coming as it did in the early days of the recent wave of interest in artificial life, seeking to provide a simple existence proof was not foolish. As MacLennan has said (personal communication) “You have to remember that at this time we were not sure that communication would evolve *at all!*” However, in terms of future work, it is surely time to accept that, given appropriate co-operative payoffs, communication will evolve in a simulation without too much trouble, and move on to building models of specific phenomena.

Beyond their proof-of-concept ambition, MacLennan and Burghardt were also examining a process by which arbitrary symbols can evolve to denote something in a simple "language". As they put it, "beyond merely detecting the presence of communication, we are also interested in studying its structure." (p. 173). Furthermore, because the simorgs must come to know not only the correlations between symbols and local states, but also when to act and when to emit, MacLennan and Burghardt were effectively looking at the evolution of turn-taking. Finally, they were interested in the effect of learning on the evolution of communication. This conjunction of goals seems to have worsened their problem with artefacts. Each of these phenomena are poorly understood, and each is worthy of a separate, narrowly-focused simulation experiment. When all of these questions of interest are thrown in together, they interfere with each other and make the extraction of general principles impossible. For instance, in trying to push the simulation towards communication, MacLennan and Burghardt chose to reward both the sender and the receiver of a message, and in an effort to leave things open-ended enough for spontaneous symbol meanings to develop, they used the FSM architecture. But what is the relative importance of these two factors in accounting for the observed results? MacLennan and Burghardt allowed spontaneous strategies for emitting vs. acting to develop amongst the simorgs, presumably to leave them as unconstrained as possible, but this decision created the loophole described in section 5.2.3. Would the same type of communication have developed if the simorgs were constrained to be senders and then receivers in turn?

In principle, it may be that communication between simorgs is entirely dependent on their internal architecture, or on the fitness reward structure used, or some other quirk of the methodology—MacLennan and Burghardt themselves note that when the method for selecting parents was deterministic rather than stochastic, communication did not develop. It is not possible, from MacLennan and Burghardt's results alone, to determine any necessary or sufficient conditions for the evolution of communication.

Of course, the claim is not that if only the various factors bearing upon the behaviour of MacLennan and Burghardt's simorgs could be isolated, then the general principles governing naturally evolved communication would be laid bare. It is quite likely that there are complicated, non-linear interactions even in their small system. However, if we do not understand the effect of each factor alone (e.g., costs and benefits of communication, updating method, control architecture) then it would seem optimistic to hope to understand the complex case. We have to learn to walk before we can run.

Chapter 6

A program of simulations

The thesis has covered a range of topics up to this point. We have reviewed biological theories on signalling, and considered more closely the concepts of communication and of function. We have also explored the possibility of using evolutionary simulations as scientific models, and looked at previous simulation work on the evolution of communication. The aim of the current chapter is to briefly integrate all of this material and to come up with a concrete program of simulation research. Given the perspective that has been outlined, there are a number of questions about what to do next. For example, which questions in the theoretical literature are most amenable to a simulation approach? How can we be sure that proper signalling, in the sense described in section 3.3, has evolved in a simulation? What kind of simulation model will best allow us to avoid the artefactual pitfalls of artificial-life methods?

6.1 The question of what to model

A good starting point can be found in Quine's view of the nature of science (see section 4.2). Quine points out that the project of science is not as simple as sorting theories into the categories true and false. The truth of a particular theory or idea can always be maintained despite apparently contradictory evidence by making adjustments elsewhere in one's conceptual scheme. Therefore those of us who wish to construct evolutionary simulations should accept that the results of a simulation are never going to be the final word on the truth of a scientific theory. Evidence from simulations must be interpreted in the context of what we already believe to be true, and if everything goes well the results of a simulation should lead to our favouring one model over another, rather than standing as definitive proof that any one theory is correct. It is important to recognize that simulations are in no worse a position, in this respect, than any other research technique—Quine's philosophy shows us that there is no royal road to truth.

Quine's view of science also implies that the most useful simulation will be one pitched at a contentious issue. In other words, the models we construct should deal with cases where some piece of theory is at stake, so that the results of the simulation have the potential to push us towards one theory or another. There is no point in modelling a situation in which the result is a foregone conclusion and will not surprise anyone nor upset anyone's favourite theory. For example, an evolutionary simulation that established that fecundity was favoured by natural selection would only be a re-statement of something that is already accepted. As Sober (1996) puts it, artificial-life models will be of particular value if the results are surprising. Incidentally, this is another way of looking at the criticisms that have been levelled at MacLennan and Burghardt's (1994) work in chapter 5: showing that communication will evolve in their model, when positive payoffs to signallers and receivers mean that it is strongly favoured by selection, is no surprise. To use Quine's own metaphor, the models we construct must have the potential to occasion a shift in the "field of force" that is the whole of scientific knowledge.

It follows that in terms of Bullock's (1997a) classification scheme for work in artificial life (section 4.1.1), the simulations presented here will definitely be "model class". We will not be dealing with radical new ways of looking at the whole issue of animal communication; the thesis stands within the established adaptationist paradigm. Nor will we use simulated evolution as a tool for building communication systems that have some practical utility. The idea is to construct simulations that will help to refine our scientific picture of how and why animals come to communicate.

Several contentious points in the theoretical literature on communication were discussed in chapter 2. A central theme, and an issue that connects all of the simulation models that will be presented in the following chapters, is the problem of honesty. Darwin and the ethologists thought that honesty would be its own reward, and did not see its existence as problematic. However, a more considered view of the logic of selection suggests that in a very wide range of cases there will be room for cheats, liars and free-riders to invade a population of honest signallers—game theory often predicts that the only evolutionarily stable outcome will be poker faces and non-signalling equilibria. But there is an obvious tension between this conclusion and observations of the natural world: when we look around the animal kingdom, we find what appears to be signalling almost everywhere. Simulation models might help us to reconcile this discrepancy.

The problem of honesty can be approached by thinking about conflicts of interest. As has been argued, the evolution of honest signalling when both agents have a clear common interest is not very interesting, at least not from a functional or strategic point of view. Moreover, the "selfish gene" view of evolution suggests that such co-operative situations are not the norm but represent a phenomenon in need of special explanation—reciprocal altruism, mutualism and kin selection have been proposed for this purpose. On the other hand, much of the signalling we see in nature involves, at least potentially, a conflict of interests. For example, we need to be able to explain why alarm-calling monkeys do not quietly slip away from the approaching predator and save themselves. We need to know whether we should ever expect two animals engaged in a contest to exchange honest signals of strength or aggressive intent. We need to explain why males should use sexual ornaments as honest signals of their underlying quality, if indeed that is what such ornaments are. In each of these cases, there appears to be some benefit in signalling dishonestly or in not signalling at all, and thus a conflict of interests exists between signallers and receivers. The program of simulations that will be pursued here looks at just these questions.

To some extent the specific topics that will serve as the bases for simulation models have been arbitrarily chosen. In the space available, it is not possible to look at all possible signalling scenarios involving conflicting interests. However, one virtue of the selection of issues is that it covers a wide range of ecological contexts. All of the “four F’s” of animal behaviour—feeding, fighting, fleeing and reproduction—are represented. Thus, in chapter 7 we will look at situations such as food calls and alarm calls, in which one animal can potentially inform another about some state of the world, but perhaps at some cost to itself. That such calls have evolved in many species is not in dispute, but this basic signalling scenario allows us to investigate the claim by Krebs and Dawkins (1984) that signals will be costly in situations where a conflict of interests exists, and cheap when the situation is truly co-operative. In chapter 8 we will consider signals exchanged during contests over a resource. Building a simulation allows us to deal with a somewhat more realistic model than would be tractable given a game-theoretic approach. We will look at the success of the standard game-theoretic prediction that competitors will maximize ambiguity concerning their strength and intentions rather than signalling honestly. There exist some game-theoretic accounts that make the opposite prediction: that cost-free signals of strength and intent can be evolutionarily stable. The results of the simulation should help direct us towards one or the other of these two views. Finally, in chapter 9 we will look at sexual signalling. Several recent models have shown that, given certain assumptions, it can be evolutionarily stable for males to signal their underlying quality to females. However, these models have almost all assumed that male quality was environmentally determined; the simulation will deal with the more complicated case in which a male’s quality level is a heritable trait. Furthermore, whether or not sexual ornaments have a communicative function has been the subject of a long-running debate, because Fisher’s process of runaway sexual selection provides a compelling alternative explanation for their existence. The effects of the Fisher process will therefore also be considered in the simulation.

It should be noted that some of the questions we might want to ask about animal signals are not going to be addressed by the type of simulation model presented here. Guilford and Dawkins (1991) and Johnstone (1997) argue that potential signallers face two kinds of problems, and thus that there are two complementary ways of understanding the function of animal signalling (see section 2.6). One way is to look at problems of *efficacy*: for example, a signal must be noticeable above background noise and the noise of other signallers, it must be noticeable at the range that potential receivers are likely to be encountered, and it must be sufficiently distinctive so as not to be confused with any other signals. On the other side of the coin are problems of *strategy*: given the signaller’s internal state and external circumstances, given the behaviour of other signallers, given the likely behaviour of receivers (both intended and unintended), what signal (if any) would it be most profitable to make right now? The style of evolutionary simulation that has been advocated in chapter 4 is clearly more appropriate for addressing the second set of concerns than the first. The sort of functional explanation we are going to be able to support with these models is much more along the lines of “signallers produce this costly signal as a way of proving to receivers that they are being honest” rather than “signallers produce a loud and costly signal because that particular volume and frequency of sound maximizes the range of transmission in their environment.”

6.2 When is it proper signalling?

A definition of proper signalling has been defended at length in chapter 3. How will we decide whether or not proper signalling has actually evolved in a particular simulation? The simulations described in the following chapters will deal with situations in which one individual has access to information, typically in the form of a hidden state variable, that another cannot perceive. For example, in chapter 8 each animal knows it own fighting ability but cannot perceive the fighting ability of its opponent. In chapter 9 males know their own quality but females cannot detect this value directly. The possibility exists in each model that a signalling system might evolve, whereby some action of the signaller's informs the receiver as to the value of the hidden state variable. The evolution of communication can therefore be indexed by the degree to which receivers come to behave as if they could perceive the hidden state. To put it another way, if receivers behave differently for different values of a certain variable, and if there is no way that receivers can perceive this variable for themselves, then signallers must be informing them of the variable's value, and thus we have communication. For example, in chapter 7 a simple signalling game is described (see Figure 7.2). The hidden state can take one of two values, and we can think of these values as referring to the presence and absence of a predator, or the presence and absence of food. Only signallers can perceive the hidden state. If receivers behave in one way when a predator is present, and in another way when there is no predator, then signallers must actually be sending an informative signal. A simple χ^2 statistic applied to a cross-tabulation of predator presence or absence and receiver response (stay or flee) will tell us whether or not communication is occurring.

This choice of modelling strategy, in which statistical evidence of information transmission is equated with proper signalling, may seem like a cheapening of Millikan's ideas. After all, it has been argued in section 3.2.3 that communication cannot be effectively defined solely in terms of information transmission. The whole point of the definition of proper signalling is that it is not enough to show that information is being transmitted; one has to show that there is a history of selection for both the production of the signal and the performance of the response. However, the reason we can get away with the simple statistical approach described above is that in a sufficiently simple communication game, proper signalling reduces to information transmission. This is because, in the extremely limited world of the simulation, we can see that there is no other possible function for the "signal" than to function as a signal. Similarly, the only possible function for a response strategy that specifies different behaviours based on the signaller's behaviour is as a response to information. Due to the receiver's inability to perceive the hidden state, any statistical link between receiver behaviour and this state would not exist unless both parties had been selected to participate in a communication system. Therefore such a statistical link counts as evidence of proper signalling. We will see that in the somewhat more complex simulation described in chapter 8, in which both parties are at once signallers and receivers, and in which the potential signal may indeed have another function, the identification of proper signalling becomes a much more difficult task.

Although this is not the place to re-open the debate on what counts as communication—that bridge has been crossed in chapter 3—it should be acknowledged that there are critics (e.g., Di Paolo, 1997b) of the idea that the essential features of communication are one sender, one receiver, and the transmission of information between the two via some token in the environment. Di Paolo prefers to define communication as any kind of social co-ordination. He acknowledges that some communicative scenarios can be modelled using the sort of simple signalling game that will be introduced in chapter 7, but points out that we should first explain how co-ordination itself evolved. How do animals manage to pair up as signaller and receiver? How do they come to perceive some kinds of sensory input as signals? Di Paolo's points are well made, but there are grave practical and theoretical difficulties involved in modelling the evolution of the basic co-ordination that must occur before things like communication games are possible. We have good reason to believe that such evolution occurred in the natural world, but attempting to mimic it in the computer brings us squarely up against the problem of epistemic autonomy (Pattee, 1995; Prem, 1997). If we give simulated agents too much in the way of existing sensors, effectors, internal architecture, etc. then it is difficult to claim that any resulting co-ordination truly "evolved from scratch". On the other hand, if we give them too little—as an extreme example, if we simulate primal chemical processes in detail and wait for metabolism to emerge—then simulated evolution will take a dauntingly long time on even the fastest computers. Furthermore, even if co-ordination and communication did evolve in such a minimal simulation, it is not clear what the implications would be. What exactly would we need to revise, among our scientific beliefs, if communication did or did not evolve in such a case? The appeal of models that look for evolutionarily stable strategies in fixed communication games is that they do not suffer from these problems.

6.3 Getting the most out of simulations

In chapters 4 and 5, a great deal of discussion was devoted to the potential pitfalls of the artificial-life approach. The reader might legitimately ask what has been done to ensure that the work described here will not suffer from these problems. The first point that should help to ensure the usefulness of these simulations is that they are designed to be models of real-world phenomena rather than abstract exercises in computer programming. Thus the outcome of each simulation has potential consequences for the way we look at the world. The simulations also build on earlier work, largely from the theoretical biology literature. This is important given the Quinean picture of science that has been argued for: a simulation effort that attempted to establish a completely novel conjecture might be revolutionary, but far more often it would be easily dismissed as irrelevant. The value of the evolutionary simulations in chapters 7, 8 and 9 lies in the fact that they can help us to select one theoretical account over another in cases where traditional models and methods of argument have proved inconclusive.

Some care has been taken to avoid introducing artefacts into the models. It would of course be foolish to claim that the simulations were *free* from artefacts, and indeed making sure that a computer model includes only the salient features of the modelled situation is something of a black art. However, the fact that the models have strong theoretical connections means that many of the potentially arbitrary decisions that must be made during the simulation construction phase, such as the exact value of a particular constant, are not wild shots in the dark but have some basis in prior work. In addition, variations on each model have been explored in as much depth as space and time constraints have allowed. Caryl (1987) has expressed dismay at a tendency in the theoretical-biology literature for those who build mathematical or simulation models to engineer them solely in order to support a favoured hypothesis, and to fail to consider the broader implications and predictions of such models. Caryl's point is that it is very easy to judiciously choose parameter values in order to get a desired result, but harder to construct a model that makes sensible predictions in a range of contexts. It is hoped that the presentation of minor variants on each simulation will increase the reader's confidence in the validity and general applicability of the models. Finally, the models have been kept as simple as possible. As Maynard Smith has said (personal communication) “it’s no good replacing a phenomenon we don’t understand with a model we don’t understand”.

The attempt to keep the simulations simple must be balanced, however, with the need to exploit the potential benefits of the artificial-life approach. For example, there is no point in constructing a simple simulation to show that the ESS in the war of attrition involves random waiting times drawn from a negative exponential distribution, because the existing mathematical models make this point in an even more concise way. If we are going to construct simulations for some reason other than mere fear of algebra, then those simulations should model aspects of the world that are difficult or impossible to capture with mathematical methods. There are all sorts of possible ways to achieve this; some of the most obvious are to construct simulations that include more realistic treatments of evolutionary dynamics, of space, and of time. Unfortunately it has not been possible to build all of these features into all of the models.

Nevertheless, in chapter 7 attention will be paid to evolutionary dynamics: we will see that evolution takes a different course depending on whether the initial population of simulated organisms consists of honest signallers or of non-communicators. Chapter 7 will also look briefly at the effects of spatial arrangement on the tendency of the simulated organisms to exhibit altruistic communication. In the work on animal contests in chapter 8, time will be treated in much more detail than is possible in a game-theoretic model. In addition, the genetically specified control architectures for the simulated organisms will be recurrent neural networks; effectively, this allows us to consider a much larger space of possible strategies than in the typical game-theoretic approach, in which only a small group of “representative” strategies are used. Finally, in the model of sexual signalling in chapter 9, the use of an actual population of simulated organisms each with their own genotype, rather than the merely abstract population of a population-genetic model, means that we can allow important genetic correlations to vary naturally. This would normally result in mathematical intractability.

Chapter 7

Co-operative and competitive signalling

This chapter has a threefold purpose. Firstly, it introduces a simple signalling game that can be used to model situations such as food and alarm calls, in which one animal informs another about some state of the world. Secondly, it is an attempt to test Krebs and Dawkins's (1984) theory that two kinds of signal co-evolution should be expected in nature (see section 2.5.2): expensive signals resulting from manipulative arms races when participants have conflicting interests, and conspiratorial whispers that evolve when the interests of the participants are congruent. Finally, it is an attempt to position some of the previous artificial-life work on the evolution of communication in a broader theoretical context.

7.1 Background

7.1.1 Explaining food and alarm calls

In many social species, an individual that has discovered a supply of food may, under some circumstances, produce a signal that serves to alert conspecifics to the presence of the resource. For example, chimpanzees *Pan troglodytes*, on discovering a fruit tree, will make loud hooting sounds that attract others (Reynolds & Reynolds, 1965; Sugiyama, 1969). Male domestic chickens *Gallus gallus* give a distinctive call in response to food; they are more likely to produce the call if a hen is present, and the calls attract other chickens (Evans & Marler, 1994). The elaborate dances of bees (von Frisch, 1967) can be considered a particularly sophisticated food signal. Some social animals also produce alarm calls, in which an individual that has detected a predator alerts other group members: the calls of vervet monkeys are an excellent example and have already been much discussed. Alarm calls are also given by other mammalian species (see e.g., Sherman, 1977) and by many birds (Klump & Shalter, 1984; Hauser, 1996). Sometimes alarm calls even serve to recruit conspecifics to mob (i.e., to jointly attack or distract) the approaching predator.

The function of these kinds of signalling systems seems transparent: the signal serves to alert others, and the response of a receiver (i.e., approaching the food or running away) is likely to have positive fitness implications given the context. Barring misidentification, as could occur when what appears to be a food call turns out to be an aspect of sexual advertisement signalling for instance, the adaptive significance of food and alarm calls looks obvious. However, as discussed in sections 1.2 and 2.3.3, the problem of altruism means that food and alarm call systems are not so easily explained. It is easy to see where the benefit lies for receivers of the signal; being informed of the approach of a predator or the location of food is clearly useful. It is not so easy, however, to determine why the signaller should share the relevant information. In many contexts there will either be no benefit in doing so, or, more likely, costs involved. These costs may be due to, for example, energy expenditure in the production of the signal, an increase in personal risk for the signaller, or the loss of food that might have been consumed alone. There is thus a degree of altruism in such signalling, and a conflict of interests between the signaller and the receiver. With mobbing calls, the altruism runs in the other direction: why should receivers of the signal risk their own lives by assisting in a group attack on the predator?

The problem of accounting for honesty becomes even more acute when we consider communication that occurs with a more explicit conflict of interests between signallers and receivers. For example, in aggressive or territorial signals, each animal would prefer that the other respond by retreating. In many sexual advertisement signals, it is in the interests of the average male to convince any female he meets to copulate with him, but it is in the average female's interests to be difficult to persuade. In these cases and in the apparently co-operative context of alarm and food calls, what prevents the invasion of free-riders who gain the benefit of others' honest signals, but do not pay the costs of honesty themselves? How can honest signalling be an ESS? Furthermore, how might communication have evolved in the first place—why, against an initial background of non-communication, would the first proto-signallers have been selected for their behaviour?

Reciprocal altruism (Trivers, 1971), kin selection (Hamilton, 1964), and the handicap principle (Zahavi, 1975, 1987) are among the mechanisms that have been proposed to explain the evolution of stable, honest signalling, and each of these ideas has spawned a vast literature of its own—particularly that on reciprocal altruism and the Prisoner's Dilemma. However, these three mechanisms will only be treated briefly if at all in this chapter. Our goal is instead to consider a prediction arising from Krebs and Dawkins's (1984) account of animal signalling.

7.1.2 Expensive hype and conspiratorial whispers

Krebs and Dawkins (1984) provide another way of looking at the problem of honesty. As we have seen in section 2.5.2, Krebs and Dawkins challenge the default notion that animal communication is about information transmission; they suggest that propaganda and advertising make better metaphors for animal communication than does the co-operative use of language to share information. They predict two distinct varieties of signal co-evolution. On the one hand there will be evolutionary arms races between manipulative, exploitative signallers and sceptical receivers. This will occur when there is a conflict of interests between the two parties, and the result will be increasingly costly signals. On the other hand, there are some situations in which—to use Krebs and Dawkins's terminology—it is to the receiver's advantage to be manipulated by the signaller. When the two parties share a common interest in this way, there will be selection for signals that are as cheap as possible while still being detectable: "conspiratorial whispers".

The aim of this chapter is to construct a model of food and alarm call situations and then to ascertain whether, given appropriate manipulation of the degree to which the participants have common or conflicting interests, these two types of signal evolution in fact take place. If so, Krebs and Dawkins's theory may turn out to be a sufficient explanation for "honest" signalling in nature: signalling systems in contexts of common interest are not subject to invasion by dishonest, free-riding mutants, while signalling systems that exist despite conflicting interests are likely to involve much more costly signals and to be ultimately unstable.

In contrast to the handicap principle, few mathematical or simulation models of Krebs and Dawkins's theory have ever been constructed. Presumably, their ideas were accepted without detailed modelling because the argument followed so naturally from the dominant selfish-gene paradigm—models of the handicap principle were constructed because there was fierce debate over whether it would or would not work. In order to test Krebs and Dawkins's prediction, it will first be necessary to determine whether communication should be expected *at all* when signallers and receivers have a genuine conflict of interests.

7.1.3 Putting artificial-life models of communication in perspective

We have seen in section 5.1.3 that previous artificial-life work on the evolution of communication has considered situations in which signallers and receivers have common interests (Werner & Dyer, 1991; MacLennan & Burghardt, 1994), conflicting interests (de Bourcier & Wheeler, 1994; Bullock, 1997b), and intermediate cases in which signallers are ambivalent about the response of receivers (Ackley & Littman, 1994; Oliphant, 1996). A secondary goal of the current chapter is to position this earlier simulation work in an over-arching context. Section 7.1.4 below describes a classification scheme for common and conflicting interests between signallers and receivers; investigating the course of signal evolution across a range of contexts will allow us to incorporate the earlier findings in a unified picture.

While on the subject of previous artificial-life work, it should be noted that the model developed in this chapter postulates a single environmental variable that animals might come to communicate about. This state can take on one of two values, corresponding to, for example, the presence or absence of food. Earlier work—notably MacLennan and Burghardt (1994)—considered "multiple meaning" situations in which a number of environmental states came to be paired up with a number of potential signals. However, if MacLennan and Burghardt's simulation shows us anything, it shows by existence proof that positive payoffs all round for successful communication can transform initially random token-meaning relationships into a workable communication system. The same can be said for work by Steels (1995). The current chapter is limited to the simple one-meaning case in order to more clearly study the effects of different payoff and signal-cost values on the evolution of signalling.

7.1.4 Conflicts of interest

The first requirement in constructing a general model of communication is a classification scheme for determining when a conflict of interests exists between signallers and receivers—Figure 7.1 shows such a scheme, adapted from Hamilton (1964). Assume that a successful instance of communication in a particular scenario has fitness implications for both participants. The fitness effect on signallers, P_S, and the fitness effect on receivers, P_R, together define a point on the plane in Figure 7.1. For example, consider a hypothetical food call, by which one animal alerts another to the presence of a rich but limited food source. By calling and thus sharing the food, the signaller in-

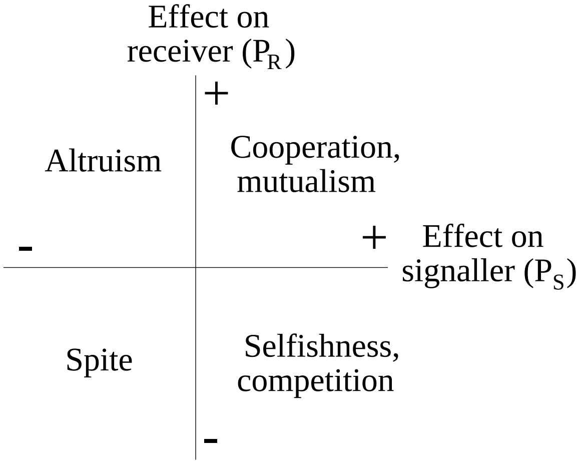

Figure 7.1: Possible communication scenarios classified by their effects on the fitness of each participant.

curs a fitness cost; by responding to the call, the receiver benefits through obtaining food it would otherwise have missed. Thus, the call would be located in the “altruism” quadrant. The situations modelled by Ackley and Littman (1994) and Oliphant (1996)—discussed in section 5.1.3—in which receivers benefit but signallers are ambivalent, can be thought of as points on the positive vertical axis, i.e., where $P_S = 0$ and $P_R > 0$.

Conflicts of interest can be defined as interactions in which natural selection favours different outcomes for each participant (Trivers, 1974), or in which participants place the possible outcomes in a different rank order (Maynard Smith & Harper, 1995). Conflicts of interest therefore exist when P_S and P_R are of opposite sign, i.e., in the upper-left and lower-right quadrants. Selection will, by definition, favour actions that have positive fitness effects. In the upper-left and lower-right quadrants, one individual but not the other will be selected to participate in the communication system: their interests conflict. The “spite” quadrant does *not* represent a conflict of interests because agents will be mutually selected not to communicate.

If the specified fitness effects of participating in a communicative interaction are truly *net* values, and already include such factors as the cost of signalling and the cost of making a response (as well as inclusive fitness considerations and costs due to exploitation of the signal by predators, etc.), then predicting the evolution of the communication system is trivial. Proper signalling requires that it be in the interests of both signallers and receivers for the communication system to exist, and so presumably will only develop when $P_S > 0$ and $P_R > 0$, i.e., when individuals in both roles are selected to participate. However, real animals sometimes appear to communicate despite conflicts of interest, as in signalling during contests (chapter 8) and sexual signalling (chapter 9). Recent models (Grafen, 1990a; Bullock, 1997b) have established that, in certain situations where communication would otherwise be unstable, increasing the production costs of the signal can lead to evolutionarily stable signalling. The costs of signalling (and responding) have therefore been separated from the cost or benefit associated with the outcome of the interaction. In other words, P_S and P_R refer to gross fitness effects before the specific costs of producing the signal, C_S, and making the response, C_R, have been taken into account. Assuming for the sake of the

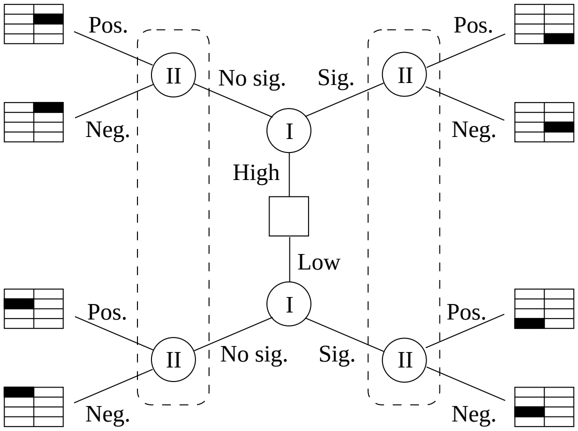

Figure 7.2: Extended form of the simple signalling game. The shaded cell in each chart icon indexes the relevant payoff value in Table 7.1.

argument that Krebs and Dawkins are correct in predicting two kinds of signal co-evolution, this separation makes it possible to identify the two regimes based on variations in C_S, the cost of signal production.

7.2 A simple signalling game

If the signalling interaction is to involve information transmission, and allow for the possibilities of proper signalling, deception, and manipulation, it must be modelled as a game of imperfect information, in which the signaller knows something that the receiver does not. Figure 7.2 shows the extended form of a simple action-response game that captures the structure of the alarm- or food-call context, and arguably other contexts besides. The game begins with a chance move (the central square) in which some state is randomly determined to be either “high” or “low”. The signaller has access to this state, and we can suppose that it represents either a feature of the environment that only the signaller has detected (e.g., noticing an approaching predator), or a hidden internal state of the signaller (e.g., ovulation). Based on this state, the signaller (player I) must decide whether or not to send an arbitrary signal of cost C_S. The receiver (player II) is ignorant of the hidden state and only knows whether or not a signal was sent—the dashed rectangles show the receiver’s information sets. The receiver can respond either positively, i.e., perform some action “appropriate” to the high state, or negatively, i.e., not respond at all. Positive responses incur a cost, C_R. If and only if the hidden state is high, a positive response results in the payoffs P_S and P_R to the signaller and receiver respectively. Table 7.1 specifies the payoff matrix.

Hurd (1995), Oliphant (1996), and Bullock (1997b) used similar games with different payoff structures. In each of these earlier games, the receiver was explicitly rewarded for accuracy in determining the hidden state. In contrast, in the current game accuracy is not a goal of the receiver *per se*; the receiver simply wants to maximize its average payoff. This is in keeping with Bullock’s point about the information requirements of receivers, discussed in section 2.6.2. Depending on the precise payoff values, the best way to maximize one’s payoff might be to respond in a blanket way, i.e., responding negatively or positively whatever the signal. This is meant to reflect the fact

	State of environment	
	Low	High
No signal		
Neg. response	0 , 0	0 , 0
Pos. response	0 , $-C_R$	P_S , $P_R - C_R$
Signal		
Neg. response	$-C_S$, 0	$-C_S$, 0
Pos. response	$-C_S$, $-C_R$	$P_S - C_S$, $P_R - C_R$

Table 7.1: Payoff matrix for the simple game. Entries in the table represent the payoff to the sender and receiver respectively.

that receivers in natural contexts can presumably opt out of the communication system if it is to their advantage to do so; there is no force compelling them to pay attention to the signaller.

The game models a range of possible communicative interactions. For example, suppose that the high state represents the signaller's discovery of food. Sending a signal might involve emitting a characteristic sound, while not sending a signal is to remain silent. For the receiver, a positive response means approaching the signaller and sharing the food, whereas a negative response means doing nothing. Various possibilities exist besides honest signalling of the high state: the receiver might *always* approach the signaller in the hope of obtaining food, regardless of whether a signal was sent. The signaller might be uninformative and never signal, or only signal when food was *not* present. One important feature of the game is that the signaller is ambivalent about the receiver's response in the low state—in terms of the example, this represents the assumption that when no food has been discovered, the signalling animal does not care about whether the receiver approaches or not.

The strategies favoured at any one time will depend on the relative values of P_S, P_R, C_S and C_R, as well as on what the other members of the population are doing. (Another parameter of interest is the relative frequency of high and low states; in the models presented here each state occurred 50% of the time.) Allowing the base fitness effects P_S and P_R to vary across positive and negative values will allow the payoff space of Figure 7.1 to be explored, and thus determine whether changes in signal and response cost can produce stable signalling in situations that would otherwise involve conflicts of interest. Note that in the simple game, there is no potential for signals of varying costs, and thus no room for costly signalling arms races. Variable-cost signalling will be considered later on in the chapter; this initial game is only a first step towards assessing Krebs and Dawkins's conspiratorial whispers theory.

7.2.1 Stable strategies in the simple game

A signalling strategy in the simple game specifies whether to respond with no signal (NS) or a signal (Sig) to low and high states respectively. Likewise, a response strategy specifies whether to respond negatively (Neg) or positively (Pos) when faced with no signal and when faced with a signal. A complete strategy is the conjunction of a signalling and a response strategy; e.g., (NS/NS, Pos/Pos) is the strategy that specifies never signalling and always responding positively.

The strategy (NS/Sig, Neg/Pos) specifies signalling only in the high state, and responding positively only to signals—call this the “honest and trusting” strategy. Evolutionary stability depends on a strategy being the best response to itself; i.e., a strategy must be uninvadable in order to be an ESS. Honest and trusting players meeting each other can expect an average payoff per interaction of:

$$\frac{P_S - C_S + P_R - C_R}{4}$$

This will be higher than the expected payoff for any possible invading strategy (i.e., honesty and trust will be an ESS) if:

$$P_S > C_S > 0$$
$$P_R > C_R > 0.$$

That is, honest signalling is stable if the costs of signalling and responding are both positive, and if the payoffs in each case outweigh the costs. The requirement that P_S and P_R must both be positive means that the honest strategy is only expected to be stable when the interests of the parties do not conflict: positive values of P_S and P_R place the interaction in the upper right “mutualism” quadrant of Figure 7.1. For the derivations of these results and others presented in this chapter, the reader is referred to appendix A.

Of the 16 possible strategies, there are three besides the honest strategy that involve the transmission of information, in that the receiver responds differently to different hidden states. None of these three are ESSs if C_S and C_R are both positive; these two values represent energetic costs and so cannot sensibly be negative. If $C_S = 0$, i.e., if giving a signal is of negligible cost, then the reverse honesty strategy (Sig/NS, Pos/Neg) can be stable, although P_S and P_R must still be positive. It is also worth noting that a population consisting entirely of individuals playing (NS/NS, Pos/Pos) or (NS/NS, Pos/Neg), both non-signalling strategies where the receiver always responds positively, cannot be invaded by any other strategy if the payoff to the receiver is large enough, i.e., if:

$$C_S > 0$$
$$P_S > -C_S$$
$$P_R > 2C_R > 0.$$

The analysis indicates that while the cost of signalling plays some role in stabilizing the honest strategy, there are no circumstances in which stable communication is predicted when a conflict of interests exists. This is despite the fact that we have separated the costs of signalling and responding from the base fitness payoffs of a communicative interaction.

7.2.2 Evolutionary simulation model

An evolutionary simulation model of the simple game was also constructed in order to determine whether communicative behaviour might sometimes be found outside the range of identified ESSs. A straightforward genetic algorithm (GA) was used. Each individual could play both signalling and receiving roles; a strategy pair was specified by a four-bit genotype as shown in table 7.2.

	Bit value	
	0	1
If low state…	No signal	Signal
If high state…	No signal	Signal
Response to no signal	Negative	Positive
Response to signal	Negative	Positive

Table 7.2: Genetic specification of strategies.

The population size was 100, the mutation rate was 0.01 per locus, and, due to the trivially small genome, crossover was not used. Each generation, 500 games were played between randomly selected opponents. An individual could therefore expect to play 5 games as a signaller and 5 as a receiver. The fitness score was the total payoff from these games. For breeding purposes, the fitness scores were normalized by subtracting the minimum score from each. Proportionate selection was then applied to the normalized scores. The genetic algorithm was run in this manner for 500 generations. In the results presented below, the games played in the final, i.e., 500th, generation have been used as a snapshot of the evolved signalling strategies.

An attempt was made to investigate evolutionary dynamics, in that the initial populations were not determined randomly but started as either "honest" or "non-signalling". Honest initial populations were made up entirely of individuals who played the honest and trusting strategy, i.e., a genome of "0101". Non-signalling populations underwent 100 generations of preliminary evolution in which their receiving strategies were free to evolve but their signalling strategies were clamped at "00", i.e., no signalling. For each class of initial conditions, a simulation run was performed for all combinations of integer values of P_S and P_R between -5 and +5, making 121 runs in all. Each run was repeated 25 times with different random seeds. The values of C_S and C_R were fixed at 1.

Communication was indexed by cross-tabulating the hidden state value with the receiver's response and calculating a chi-squared statistic. The receiver has no direct access to the hidden state, so any reliable correspondence between state and response indicates that information has been transmitted and acted upon. Values of the χ^2 statistic close to zero indicate no communication, and values close to the maximum (in this case $\chi^2_{max} = 500$, due to the 500 games played in the final, snapshot generation) indicate near-perfect communication.

Figure 7.3 shows the average values of the communication index for honest initial conditions. Seeding the population with honesty tests the stability of honest signalling given a particular payoff pair, much as a game-theoretic analysis does. The results are compatible with the conditions outlined in the previous section: honesty is stable when the payoffs to signaller and receiver are positive and greater than their respective costs. However, there is some suggestion of intermittent or imperfect communication when $P_R = C_R = 1$, indicating that ambivalent receivers may occasionally co-operate.

Figure 7.4 shows the average values of the communication index for non-signalling initial conditions. Starting the GA with a non-signalling population tests the likelihood that communication

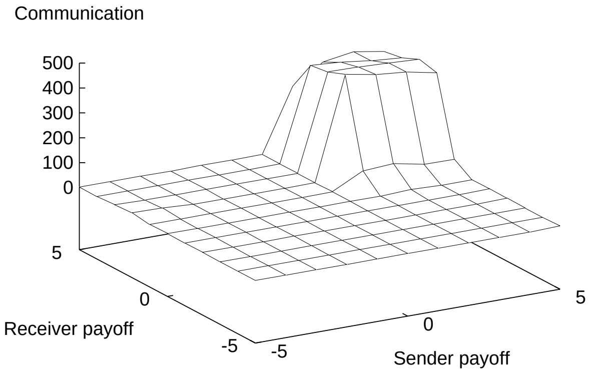

Figure 7.3: Mean communication index by P_S and P_R; honest initial conditions. Each point is a mean calculated over 25 runs. Mean standard error = 2.96.

will emerge, given a particular payoff pair. Clearly the conditions for emergence and stability-once-present are not the same. If $P_S > 1$ and $P_R = 2$ communication develops but when $P_S > 1$ and $P_R > 2$ it does not.

In the latter region $P_R > 2C_R$ and the population remains at the non-signalling equilibrium described in section 7.2.1. Despite the fact that communication would result in a higher average fitness, the high value of P_R keeps the receivers responding positively all the time, removing any incentive for the signallers to bother signalling. This response strategy could be called “blind optimism”, as receivers always respond positively. It should be noted, however, that the condition $P_R > 2C_R$ is dependent on the 50% frequency of high states; if high states occurred 10% of the time, for instance, then $P_R > 10C_R$ would be required to make blind optimism a stable strategy.

The difference in results between the two classes of initial conditions is interesting, but should not obscure the fact that no communication was observed under conditions of conflicting interests. We must conclude that, at least in the simple model discussed so far, stable communication is only to be expected when it is in the interests of both parties.

7.3 A game with variable signalling costs

In the simple signalling game, signallers can choose between a costly signal or no signal at all. The model does not allow for a range of possible signals with differing costs, and in this respect it is unrealistic. It may be that Krebs and Dawkins’s implicit prediction, that signalling can occur when a conflict of interests exists, is in fact true, but can only be demonstrated in a more complex game with a range of signal costs. The simple signalling game (see Figure 7.2) was therefore extended to incorporate signals of differing costs.

7.3.1 Stable strategies in the variable-signal-cost game

In the extended game, the signalling player has three options: not signalling, which costs nothing; using the “soft” signal, which costs C_S, and using the “loud” signal, which costs $2C_S$. Strategies in the extended game require specifying the signal to give when the hidden state is low, the signal

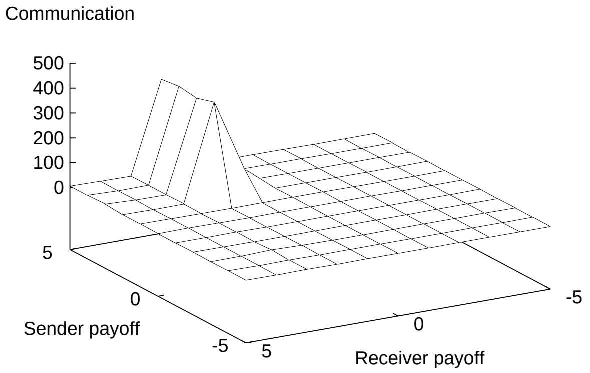

Figure 7.4: Mean communication index by P_S and P_R; non-signalling initial conditions. Each point is a mean calculated over 25 runs. Mean standard error = 1.67. Graph rotated for clarity—co-operative quadrant appears at top left.

to give when it is high, and the response to give to each of no-signal, soft and loud. The two strategies representing conspiratorial whispers or cheap signalling are (NS/Soft, Neg/Pos/Pos) and (NS/Soft, Neg/Pos/Neg). Both strategies call for the soft signal to be used in the high state, and for positive responses to the soft signal; the strategies differ only in the response to loud signals. Neither of these strategies can strictly be considered an ESS on its own (because neutral drift can take the population from one to the other) but it is shown in appendix A that the set of all mixed strategies involving these two is an ESS under the familiar conditions:

$$P_S > C_S > 0$$
$$P_R > C_R > 0.$$

Costly signalling would involve the use of the loud signal for the high state, and either the soft signal or no signal to denote the low state, with a corresponding response strategy. None of the four strategies in this category can be an ESS. For example, (NS/Loud, Neg/Pos/Pos) cannot be an ESS assuming positive costs of signalling and responding. The similar strategy (NS/Loud, Neg/Neg/Pos) is almost stable if $P_S > 2C_S$, but can drift back to the previous strategy which can in turn be invaded by the cheap strategy (NS/Soft, Neg/Pos/Pos).

Analysis of the extended game indicates that if signalling is favoured at all, then at equilibrium the signallers will always use the cheapest and the second-cheapest signal available (i.e., no signal and the soft signal). Extending the game by adding ever more costly signalling options, until we have approximated a continuous range of signal costs, does not alter this conclusion. None of the costly signalling strategies can even be an ESS, let alone support communication in the face of a conflict of interests. The possibility of expensive signalling arms races starts to look remote. However, it may be that an evolutionary simulation model will reveal signalling strategies that, while unstable in the long term, nevertheless lead to transient communication under conditions of conflicting interest.

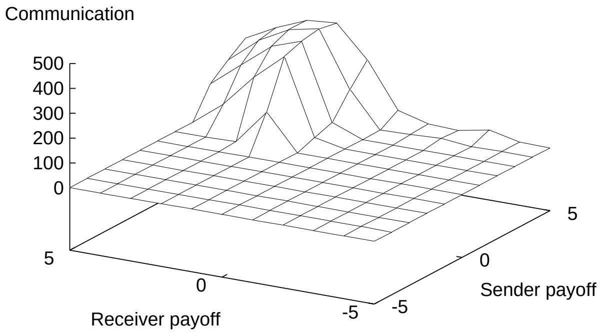

Figure 7.5: Mean communication index by P_S and P_R in the continuous simulation; honest initial conditions. Each point is a mean calculated over 25 runs. Mean standard error = 4.22. Graph rotated for clarity—co-operative quadrant appears at top.

7.3.2 Evolutionary simulation model

A second evolutionary simulation was constructed, in which the cost of signalling was continuously variable. Signalling strategies were represented by two positive real numbers C_{low} and C_{high}: the cost of the signals given in the low state and in the high state respectively. Response strategies were represented by a real-valued threshold T; positive responses were given to signals with costs greater than the receiver's threshold value. Note that threshold value could be negative, indicating a positive response to any signal.

A real-valued GA was used to simulate the evolution of strategies over time. Generally, the same parameters were used as in the previous simulation model, e.g., a population of 100. Mutation was necessarily a different matter: each real-valued gene in each newborn individual was always perturbed by a random gaussian value, $\mu = 0$, $\sigma = 0.05$. If a perturbation resulted in a negative cost value the result was replaced by zero. In addition, 1% of the time (i.e., a mutation rate of 0.01) a gene would be randomly set to a value between 0 and 5 for signal costs, or between -5 and +5 for the threshold value. This two-part mutation regime ensured that offspring were always slightly different from their parent, and occasionally very different.

The C_S parameter was no longer relevant, but C_R, the cost of responding, remained fixed at 1. Honest initial conditions were implemented by setting $C_{low} = 0$, $C_{high} = 1.0$ and $T = 0.5$. Non-signalling initial conditions were implemented by setting T to a random gaussian ($\mu = 0$, $\sigma = 1$) and then clamping $C_{low} = C_{high} = 0$ for 100 generations of preliminary evolution.

Figures 7.5 and 7.6 show the average values of the communication index for honest and non-signalling initial conditions respectively. The results are qualitatively similar to those of the discrete simulation model: communication occurs in both cases, but in a more limited range of the payoff space for non-signalling conditions. In neither case does communication occur outside the "co-operative" quadrant.

However, there is some evidence that transient communication can occur when the conflict of interests between the two agents is not too extreme. For example, consider the payoff pair

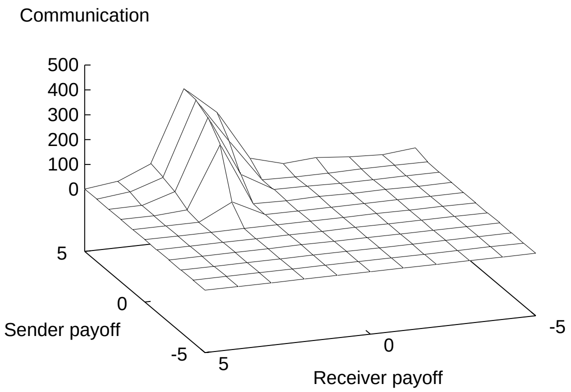

Figure 7.6: Mean communication index by P_S and P_R in the continuous simulation; non-signalling initial conditions. Each point is a mean calculated over 25 runs. Mean standard error = 3.61. Graph rotated for clarity—co-operative quadrant appears at top left.

$P_S = 5$ and $P_R = 0$. This defines a point on the boundary between mutualism and selfishness, although when the constant cost of responding ($C_R = 1$) is taken into account, the net payoffs indicate that communication under these circumstances would be selfish (from the point of view of the signaller). Nevertheless, as Figure 7.7 shows, unstable communication evolves, even from non-signalling initial conditions.

The continuous model also allows investigation of the cost and threshold values over the payoff space. C_{low}, the cost of the signal given in response to the low state, always remained close to zero—this was unsurprising as signallers are ambivalent about the receiver's response to the low state. However, the value of C_{high} varied both inside and outside the region where communication was established: Figure 7.8 shows the mean values of C_{high} for honest initial conditions. The signals given in response to the high state are most costly when P_S, the payoff to the sender, is high and when the receiver's net payoff is marginal, i.e., $P_R \approx 1$. In order to study this effect more closely, additional simulation runs were performed, with P_S fixed at 5 and P_R varied between -5 and +5 in increments of 0.1. These runs can be thought of as exploring the cross section through $P_S = 5$ in Figure 7.8. Figure 7.9 shows the cross-sectional mean values of C_{high}. Note that the "energy" devoted to signalling is at a maximum around $P_R = 1$ and drops off as P_R increases—it can be seen from Figure 7.5 that $P_R = 1$ is approximately the point where significant communication is established. The same pattern was observed for non-signalling initial conditions (not shown for reasons of space).

The threshold values showed corresponding variation. Figure 7.10 shows the mean value of T across the payoff space. The threshold values are typically very high (a "never respond" strategy) or very low (an "always respond" strategy), but in the region where communication evolved, receivers become progressively less demanding, i.e., T gets lower, as P_R increases. Figure 7.11 shows the cross-sectional results for $P_S = 5$.

Figure 7.12 plots the mean cost of high and low signals and the mean reception threshold all on one graph. This makes the relationship between costs and threshold clear: at approximately

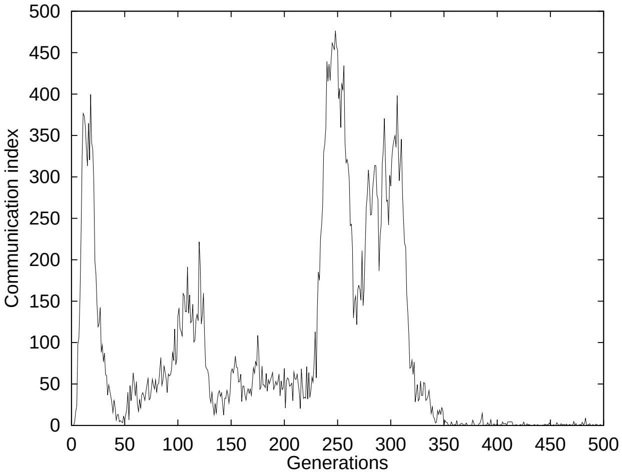

Figure 7.7: Mean communication index plotted over generational time. A typical run with $P_S = 5$, $P_R = 0$ and non-signalling initial conditions.

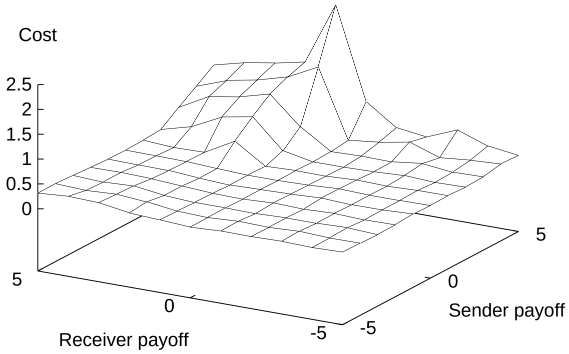

Figure 7.8: Mean cost of high-state signals by P_S and P_R; honest initial conditions. Each point is a mean calculated over 25 runs. Mean standard error = 0.032. Graph rotated for clarity—co-operative quadrant appears at top.

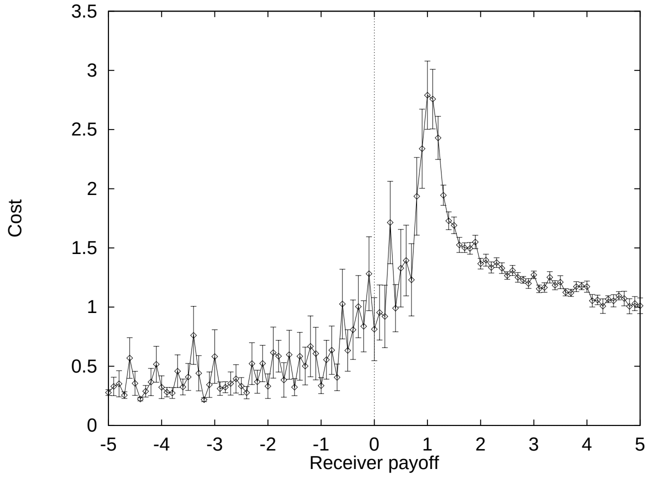

Figure 7.9: Cross-sectional means (± 1 s.e.) for high-state signal costs with $P_S = 5$; honest initial conditions. Each point is a mean calculated over 25 runs.

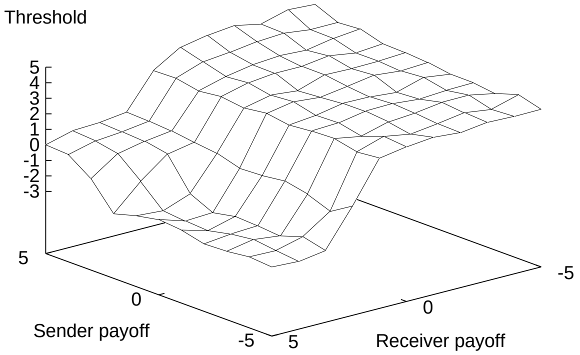

Figure 7.10: Mean threshold value by P_S and P_R; honest initial conditions. Each point is a mean calculated over 25 runs. Mean standard error = 0.19. Graph rotated for clarity—co-operative quadrant appears at top left.

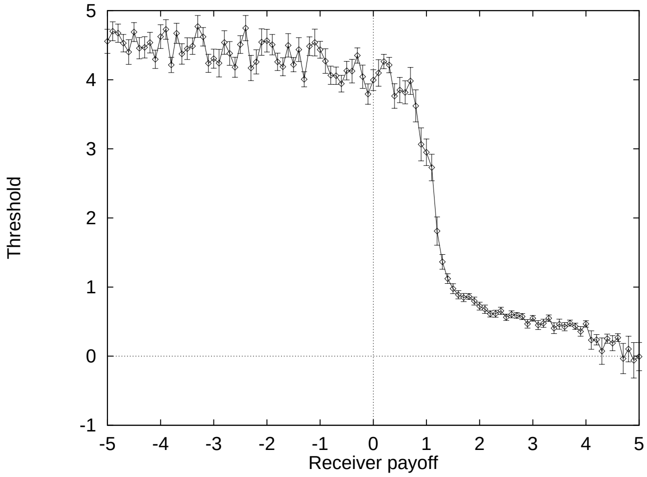

Figure 7.11: Cross-sectional mean threshold values (±1 s.e.) with $P_S = 5$; honest initial conditions. Each point is a mean calculated over 25 runs.

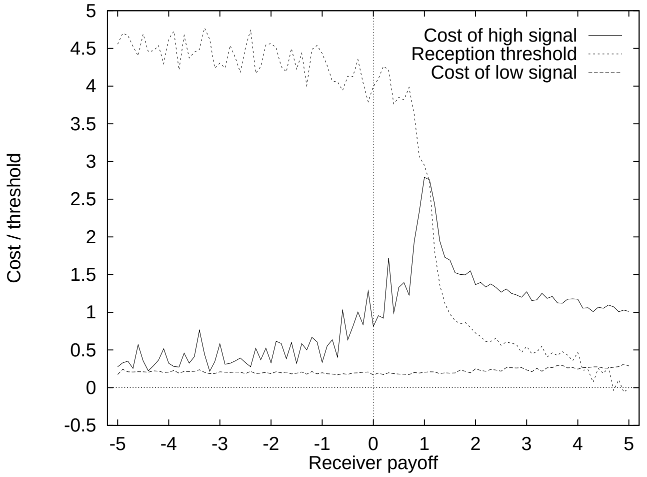

Figure 7.12: Cross-sectional means: cost of high and low signals, and reception threshold. $P_S = 5$, honest initial conditions. Each point is a mean calculated over 25 runs.

$P_R = 1$, the threshold falls to a level where the mean high-state signal will generate a positive response. As P_R increases, i.e., as the two players' payoffs approach each other, the signallers become less extravagant and the receivers less "sceptical". This is *contra* the game-theoretic result of the previous section, which implies that when signals of varying costs are available, either the cheapest pair of signals will be used, or no signalling will occur—something like Figure 7.13 would be expected if the soft-loud signalling game accurately modelled the continuous case.

Note that the initial values of C_{high} and T under honest initial conditions were 1.0 and 0.5 respectively. For all but the highest values of P_R, C_{high} has increased on average over the 500-generation run. This rules out any explanation of the results of Figure 7.12 in terms of there having been insufficient evolutionary time for a cheaper signalling equilibrium to have been reached when the profit for receivers ($P_R - C_R$) was marginal. Evolution has taken the populations *away* from the cheap signalling solution.

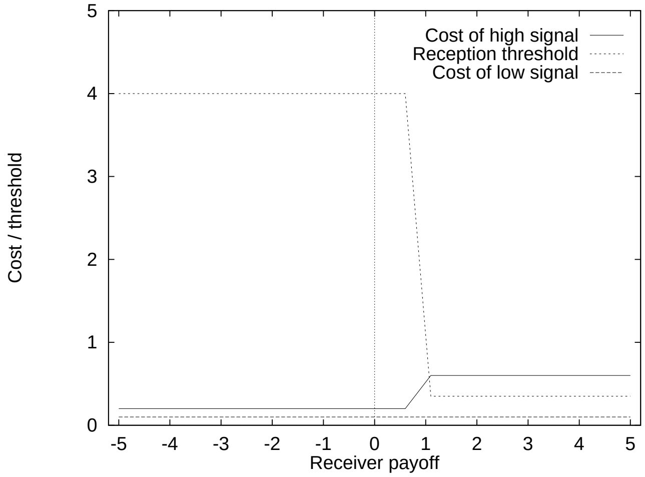

Figure 7.13: Approximate predicted results for Figure 7.12 according to discrete-cost game-theoretic model.

7.3.3 Discussion

In all of the models presented so far, stable communication evolved or was predicted to evolve only within the co-operative region of the signaller-receiver payoff space. This means that no signalling at all (costly or otherwise) was observed when the signaller and the receiver were experiencing a conflict of interests, apart from transitory communication on the boundaries of the co-operative region as shown in Figure 7.7.

The second game-theoretic model, in which discrete signals of varying costs are available, suggests that communication, if selected for, will involve the cheapest pair of signals available. However, the second simulation model, incorporating the more realistic assumption that signals can vary continuously in cost, implies that cheap signals will only be used when both parties stand to gain a high payoff from effective communication. When the net payoff to the receiver is marginal, evolved signals will be more costly than strictly necessary to convey the information. The relationship is not symmetrical: when the net payoff to the signaller is marginal, a non-signalling equilibrium, in which the receiver always responds positively, is likely to occur.

Krebs and Dawkins (1984) predicted that signalling would be costly if a conflict of interests existed; strictly speaking the results do not support nor contradict their prediction, as no signalling occurred in the conflict-of-interest cases. It might be the case that conflicts of interest in the context of a different signalling game would indeed result in costly signals. However, the failure to evolve communication given conflicts of interest in this simple game strongly suggests that in many natural contexts (e.g., food calls, alarm calls) reliable signalling should not be expected unless it is in the interests of both parties. This conclusion is not altered by separate consideration of the specific costs of producing a signal and of making an appropriate response to that signal.

The results from the second simulation model do not confirm Krebs and Dawkins's conspiratorial whispers theory, but they definitely suggest a modification of it. As Figure 7.12 shows, when the net payoff to the receiver is marginal, receivers will be sceptical and express “sales-resistance” by responding only to costly signals; signallers in turn will be prepared to invest more energy in “convincing” receivers to respond positively. When communication is unambiguously good for both parties, signals are cheaper and response thresholds lower. Therefore both expensive hype and conspiratorial whispers are expected to evolve, but in a much smaller region of the payoff space than Krebs and Dawkins's theory suggests, i.e., within the co-operative region. Expensive hype is what happens when honest signalling is highly profitable to the signaller, but only marginally so to the receiver. For example, suppose that a juvenile benefits by honestly signalling extreme hunger to its parent, because the parent responds by feeding it. If the net inclusive-fitness payoff to the parent is only slight, perhaps because the parent is the ostensible father and the species has a high ratio of extra-pair copulations, then costly signals by the juvenile are expected. Thus the model predicts that chicks should beg more loudly to their fathers than to their mothers, for instance.

7.4 Variations on the continuous-signal-cost game

In line with the reasoning presented in section 6.3, a number of variations of the evolutionary simulation model with continuous signal and threshold values will now be presented. In order to avoid any further profusion of graphs, the variants will incorporate only non-signalling initial conditions. Rather than requiring the reader to constantly compare each figure with Figure 7.6—the mean communication index data for the continuous-signal-cost game with non-signalling initial conditions—the communication index results in each variant will be presented as *differences* from that graph. That is, Figure 7.6 will be used as a reference level of communication; positive results for a variant will indicate a greater relative level of communication and not an absolute measure.

7.4.1 Noise and uncertainty

The use of continuous values for the cost of signals and for the response threshold suggests the possibility of random noise in the signalling channel. In the real world signals will not always be accurately perceived, and Johnstone (1994) found that modelling noise or perceptual error in a signalling game in fact altered the predictions about which strategies were expected to be stable. It was thought that perhaps the inclusion of noise would alter the region of the payoff space in which communication evolved.

Noise was implemented by adding a random gaussian value ($\mu = 0$) to the energy level of the signal before it was perceived by the receiver. Thus, signals will sometimes be heard as "louder" or "softer" than they in fact are. When the random gaussian value had a standard deviation of 0.2, noise made very little difference to the communication index data, i.e., communication evolved much as in Figure 7.6. When the standard deviation was set to 2.0, on the other hand, communication was entirely disrupted. Presumably intermediate levels of noise would have led to a progressive degradation of communication. However, there was no evidence that the addition of noise could lead to honest signalling in regions of the payoff space where it would otherwise not have occurred.

Randomness was also applied to the payoff values P_S and P_R in order to investigate the effects of realistic uncertainty. The payoff values, as in all game-theoretic accounts, are intended to be average expected payoffs. However, computer simulation allows us to assign payoffs in a particular interaction that are drawn from a random gaussian distribution. Thus the long term mean will be as specified, e.g., $\overline{P_S} = 2$ and $\overline{P_R} = 2$, but the rewards for successful communication in any one game will be somewhat unpredictable. When the standard deviation of the random gaussian was 0.2, the

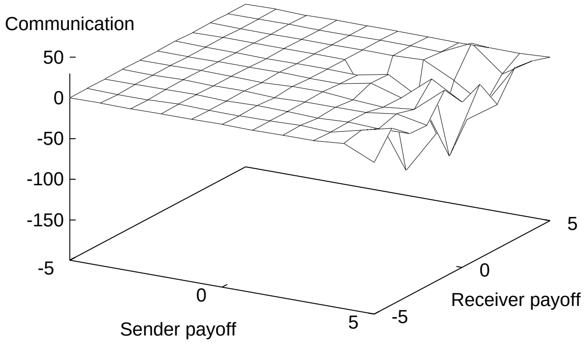

Figure 7.14: Difference in mean communication index between uncertain payoff variant (σ = 2.0) and standard continuous-signal-cost game; non-signalling initial conditions. Each point is the difference between two means, each calculated over 25 runs. Graph rotated for clarity—co-operative quadrant appears at top right.

evolution of stable communication was unaffected. When the standard deviation was increased to 2.0, communication started to degrade as shown in Figure 7.14. However, there was again no suggestion that the modelling of uncertainty in payoff values could lead to communication where it would not have otherwise evolved.

7.4.2 Exploitation of sensory biases and mutational lag

The simple games and simulations described here are in one sense an unfair way to test Krebs and Dawkins's (1984) conspiratorial whispers hypothesis. Krebs and Dawkins discuss the likely evolution of signals in complex real-world cases, and can therefore appeal to the exploitation of response patterns that had originally been selected for other purposes, the effects of differing mutation rates in signallers and receivers, etc. Communication in their predicted costly signalling arms races was not necessarily expected to be stable. For example, in a real-world situation where it was not in the interests of receivers to respond positively to a particular signal from a predator, they might nevertheless continue to do so for some time if the signal was structurally similar to a mating signal made by members of the same species. The manipulative signalling system would break down as soon as an appropriate sequence of mutations resulted in organisms that could distinguish between the predator's signal and the conspecific mating signal. In the signalling models presented all this complexity is abstracted into the base fitness payoffs for signallers and receivers.

In an attempt to investigate these issues, two simple modifications were made to the standard continuous-signal-cost game. In the first of these, we suppose that the receivers have some other ecological reason for having a low threshold value, e.g., that the same sensory mechanisms are involved in food detection. This opens up an opportunity for signallers to exploit a "sensory bias" (Guilford & Dawkins, 1991; Ryan & Rand, 1993) in the receivers. Selection pressure for low thresholds (T) was implemented by giving receivers in each game an energy bonus (b) as follows:

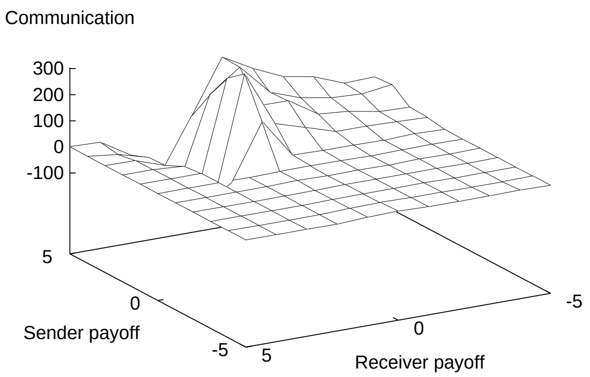

Figure 7.15: Difference in mean communication index between sensory bias variant and standard continuous-signal-cost game; non-signalling initial conditions. Each point is the difference between two means, each calculated over 25 runs. Graph rotated for clarity—co-operative quadrant appears at top left.

$$b = \begin{cases} 0 & \text{if } T > 5 \\ 1 & \text{if } T < 0 \\ \frac{5-T}{5} & \text{if } 0 \leq T \leq 5 \end{cases}$$

The results of simulation runs of this variant are shown in Figure 7.15 (using Figure 7.6 as a baseline). When receivers have other reasons for maintaining a low response threshold, communication evolves much more reliably in the usual co-operative region of the payoff space, and also occurs in the selfish region. That is, signallers are able to manipulate receivers to their own (the signallers') advantage. Furthermore, as predicted by Krebs and Dawkins (1984), the most costly signals indeed occurred when communication had been established despite a conflict of interests.

In another variant, it is supposed that response strategies might evolve more slowly than signalling strategies, i.e., there is a mutational lag on reception thresholds relative to signal cost values. Such a state of affairs could come about in the real world if the sensory equipment used to detect signals was older and affected by a larger network of genes than the organs used for signalling. It would then be possible that signallers might "out-evolve" receivers, and succeed in getting them to respond to selfish, manipulative signals. The idea was implemented by reducing both of the mutation rates for reception thresholds by a factor of 10. That is, the real-valued threshold gene in a newborn individual was perturbed by a random gaussian value, $\mu = 0$, $\sigma = 0.005$, and 0.1% of the time (i.e., a mutation rate of 0.001) a completely new threshold value was generated in the range ±5. The results are shown in Figure 7.16.

As with the sensory bias variant, communication is established more strongly in part of the co-operative region, but it also evolves in the selfish region for high values of P_S. Again, the most costly signals were also found when selfish communication had evolved. A puzzling feature of the result is that it does not appear to have come about simply because the low rate of mutation for threshold values meant that 500 generations was insufficient time for the optimal value to be reached. Mean threshold values when $P_S = 5$ and $P_R < 0$ were approximately 4 in both the

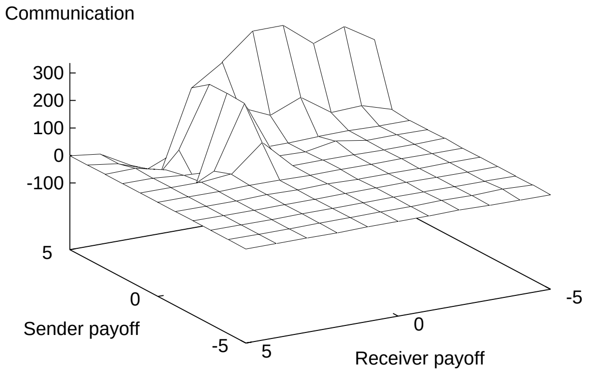

Figure 7.16: Difference in mean communication index between mutational lag variant and standard continuous-signal-cost game; non-signalling initial conditions. Each point is the difference between two means, each calculated over 25 runs. Graph rotated for clarity—co-operative quadrant appears at top left.

mutational lag variant and the original simulation data.

7.4.3 The effects of spatial arrangement

Ackley and Littman (1994) and Oliphant (1996) both found that arranging signalling populations in space led to a greater degree of altruistic signalling. In Ackley and Littman's model individuals lived in small groups, communicating and breeding only with their group-mates, but occasionally migrating to another nearby group. There was no spatial arrangement within each group, but the groups themselves were laid out on a grid. In Oliphant's model individuals were arranged in a ring, and were likely to communicate and to breed with their neighbours.

A spatial variant was implemented by arranging the population of 100 individuals on a toroidal 10×10 grid. Individuals interacted only with their 8 neighbours: in each game, a signaller was chosen at random from the population and a receiver was chosen at random from the signaller's neighbours. Breeding was also local. When one generation replaced another, the parent of the individual who would occupy a particular square was chosen, using roulette-wheel selection[1] according to fitness, from among the nine local candidates from the previous generation. That is, the parent of the occupant of a given square would either be the previous occupant or one of the previous occupant's neighbours. The results for the spatial variant are shown in Figure 7.17.

Arranging the population in space leads to an increase in the reliability of communication, but only in that section of the co-operative region where honesty has already been observed to evolve. The agents have clearly not been induced to participate in altruistic communication with their neighbours. There is no communication even when signallers are merely ambivalent ($P_S = 0$). However, it can be shown that altruism of a sort has occurred. Figure 7.18 shows the difference in

[1] Roulette-wheel selection refers to a process whereby any one individual's probability of being selected is proportional to its fitness score. The probabilities of selection can be envisaged as sectors of varying size on a roulette wheel.

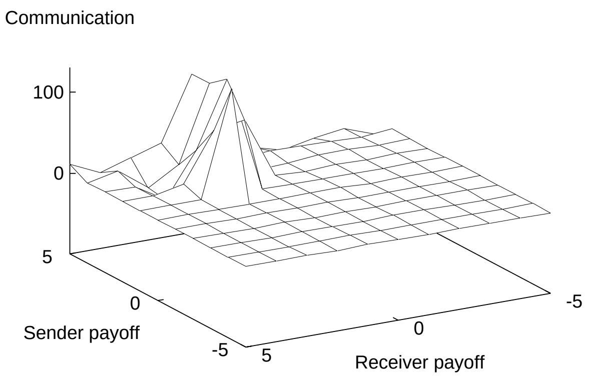

Figure 7.17: Difference in mean communication index between spatial variant and standard continuous-signal-cost game; non-signalling initial conditions. Each point is the difference between two means, each calculated over 25 runs. Graph rotated for clarity—co-operative quadrant appears at top left.

mean fitness between the spatial variant and the original simulation. There is a spike of increased fitness in the altruistic quadrant at the front of the graph: this occurs because *receivers* are refraining from constant positive responses, and thus being altruistic towards their signalling neighbours who would be penalized by a positive response because of the negative value of P_S in this area.

7.4.4 Insistent signallers

The signalling game used is not likely to be a universal model of all possible communicative interactions. In particular, and despite having the same basic structure with two signals possibly used to transmit information about a binary hidden state, the signalling game is different from those employed by Hurd (1995), Oliphant (1996) and Bullock (1997b). Hurd's game, for instance, models sexual signalling, and the male signaller is *not* ambivalent about the female receiver's response when the hidden state is low; the signaller always prefers a positive response. A low hidden state maps to low male quality, a positive response represents a copulative episode, and even low-quality males want mating opportunities. The current signalling game, in contrast, cannot model so-called "handicap" signalling, because low-state signallers do not care about what the receiver does. Furthermore, in previous games, receivers are explicitly rewarded for accuracy in discerning the hidden state, but the game presented here allows the ecologically plausible outcome that receivers simply become disinterested in the signal. The current game is a reasonable model of situations such as alarm calls and food calls, in which potential signallers have no reason to care about what receivers do when no predator has been sighted or no food source has been found. Whereas Hurd's game serves as a (discrete) model of situations where signallers vary on some dimension, the current game models situations where signallers fall into two groups, only one of which is relevant to the potential response.

However, it is a simple matter to alter the present game such that signallers are always interested in gaining a positive response. The payoff matrix is altered such that P_S, the payoff to

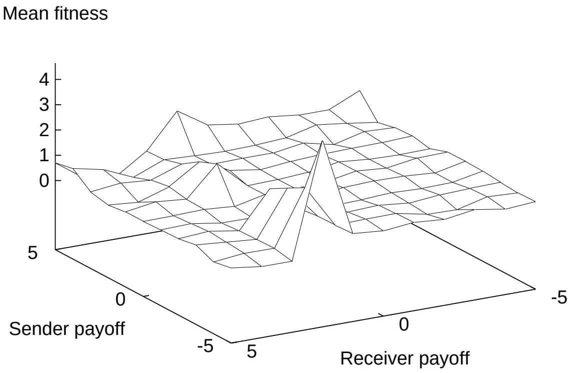

Figure 7.18: Difference in mean fitness between spatial variant and standard continuous-signal-cost game; non-signalling initial conditions. Each point is the difference between two means, each calculated over 25 runs. Graph rotated for clarity—co-operative quadrant appears at top left.

the signaller, is awarded whenever the receiver responds positively, regardless of the value of the hidden state. On the other hand, receivers are still only awarded their payoff, P_R, when they respond positively and the hidden state is high. There is thus a different kind of conflict of interests between the signaller and receiver.

Making signallers want positive replies all the time in this way almost completely breaks down communication—see Figure 7.19. There are no circumstances in which receivers can trust signallers, and extreme response strategies (always responding positively or always responding negatively) are formulated purely on the basis of the payoff to the receiver. Interestingly, communication can be salvaged if the conditions of the handicap principle are applied: that is, if the unit cost of giving a signal in the low state is greater than for the high state. The results for a run in which signals in the low state cost 5 times their normal value are shown in Figure 7.20; relative to the standard game, communication levels are only somewhat degraded.

7.5 General discussion

The results from simulations of the simple and continuous-cost signalling games suggest that communication will not evolve when there is a conflict of interests between signallers and receivers. Even when signallers and receivers share a common interest, the evolution of communication is not straightforward. Firstly, receivers may fall into blindly optimistic strategies (i.e., always responding positively) that are less efficient than the communicative equilibrium but nevertheless stable. This is particularly likely to occur when the net payoff to the receiver is high. (The expected payoff for always responding positively will of course depend on the relative frequency of high and low hidden states, a factor that was not varied in the models presented). Secondly, communication may evolve but the signals involved will be more or less costly depending on the marginal payoff of the receiver, as discussed in section 7.3.3.

Variations on the continuous-cost signalling game, while only briefly explored, suggest that

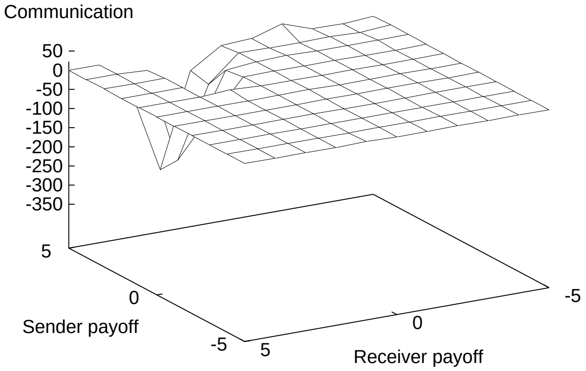

Figure 7.19: Difference in mean communication index between insistent signallers variant and standard continuous-signal-cost game; non-signalling initial conditions. Each point is the difference between two means, each calculated over 25 runs. Graph rotated for clarity—co-operative quadrant appears at top left.

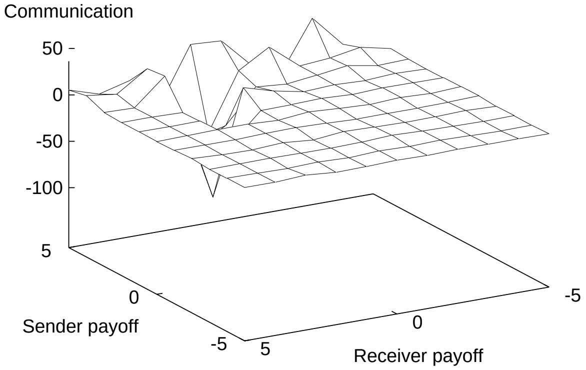

Figure 7.20: Difference in mean communication index between handicap principle variant and standard continuous-signal-cost game; non-signalling initial conditions. Each point is the difference between two means, each calculated over 25 runs. Graph rotated for clarity—co-operative quadrant appears at top left.

communication can in fact evolve under conditions of conflicting interest if receivers have a sensory bias that maintains low response thresholds, or if response strategies do not evolve as quickly as signalling strategies. In these two cases, manipulative or selfish communication can occur. Of course, in the case of a sensory bias communication that evolves is not *really* occurring under a conflict of interests, because receivers are choosing the strategy that maximizes their two sources of fitness: the communication game and the independent response bias. However, an observer unaware of the receivers' response bias would observe agents responding to signals in a way that was not in their immediate interests.

Altruistic communication (considered from the point of view of signallers) was not observed under any circumstances, including the spatial variant simulation. Spatial arrangement of the population would seem not to be a guarantee of kin-selected altruism. The occurrence of apparently altruistic food and alarm calls in nature, in circumstances where reciprocal altruism and kin selection cannot be invoked, therefore remains to be explained. In other words, this model alone cannot tell us why an alarm-calling monkey resists the temptation to quietly slip away and save itself; if an empirical study was to show that some animal gives alarm calls to non-relatives without hope of reciprocation, then we would have a genuine conundrum on our hands.

However, the model may be a step towards understanding the evolution of a different kind of non-kin, non-reciprocal altruism. Mobbing calls seem to involve a benefit for the signaller, who recruits allies to help drive off a predator, and a cost for receivers, who sustain a risk of being injured in the attack. Mobbing calls would therefore be classified as selfish under the scheme presented in Figure 7.1. In the sensory bias and mutational lag variations, this sort of selfish communication was in fact observed. From the point of view of *receivers*, this represents altruism directed towards the signaller. It might be the case that some manipulative mobbing calls are maintained despite a real cost to those who respond, because, for example, the call-production behaviour can evolve faster than the ability to distinguish between the calls of relatives and non-relatives.

The evolutionary simulation models presented are unusual in their use of non-random initial conditions. The use of non-signalling initial conditions in particular can be seen as an attempt to get at the origin or emergence of communication rather than just studying the conditions for its stability, as does orthodox game theory. Non-signalling initial conditions embody the assumption that communication must emerge from a non-communicative context—the un-clamping of signalling strategies after a period of preliminary evolution can be seen as the introduction of a mutation that allows the *possibility* of signalling. The simulation results have certainly demonstrated that the conditions for stability can be very different from those for emergence.

A final qualification must be made concerning the results: the way that conflicting and congruent interests have been defined may be too simplistic. In the simple signalling game, it is true that with positive net payoffs to the signaller and the receiver, and if the hidden state is high, both agents will benefit from a positive response, and they therefore have congruent interests. However, if we consider the moment before the hidden state has been determined, it is not clear whether the interests of the two agents conflict or not. If the signaller, for example, could somehow choose the strategy of its opponent, the receiver, it would want the opponent to play an "always respond positively" strategy—that way the signaller would always receive the payoff and would not have to expend energy in signalling. However, the receiver, if similarly allowed to determine the signaller's strategy, would prefer that the signaller used an honest strategy, precisely so that the receiver could avoid the costs of responding positively to the low hidden state. Recall that Trivers (1974) defined a conflict of interests as an interaction in which natural selection favours a different outcome for each participant. It seems that the signaller and receiver in this situation favour different strategies in their opponent, and thus have a conflict of interests, even though a high value of the hidden state would mean that their interests became congruent. If this strategy-based definition of conflicting interests were adopted, any situation in the co-operative payoff region, assuming signalling had a positive cost, would involve a conflict of interests—this would in turn mean that *all* of the signalling observed in the simulation models evolved despite a conflict of interests. The problem is perhaps that Trivers's (1974) and Maynard Smith and Harper's (1995) definitions are not specific enough about just what constitutes an "outcome" of the signalling game. The simpler definition of conflicting interests, as used in the body of the paper, is useful in isolating the co-operative region of payoff space as the place to expect signalling. It is not yet clear how the results should be interpreted if the strategy-based definition of conflicting interests was pursued.

Chapter 8

Aggressive signals as ritualized intention movements

Animal contests—disputes over resources such as food, territory or mates—are good examples of interactions in which the interests of the participants seem to be maximally opposed. This is particularly true of struggles over the control of an indivisible item: one's gain is necessarily another's loss. Nevertheless, animals contesting the possession of a resource are often observed to settle the dispute by exchanging signals or threat displays rather than engaging in an all-out fight. For example, mantis shrimps *Gonodactylus bredini* contest the ownership of small cavities in their coral reef habitat. These contests sometimes result in physical combat, but often an opponent is deterred by a claw-spreading threat display (Adams & Caldwell, 1990). Red deer stags *Cervus elaphus* compete for control of groups of females, but unless two stags are closely matched in strength, the weaker will usually retreat after a roaring contest and/or a parallel walk display (Clutton-Brock, Albon, Gibson, & Guinness, 1979).

What is happening in these cases? Are the competing animals likely to be exchanging proper signals, informing each other of their fighting ability or their intention to attack? (And if not, what is the function of their aggressive displays?) Intuitively, settling contests by signalling makes sense. We can see that an all-out fight is usually a bad idea: fighting is energetically expensive, and there is always a risk of injury or death. The early ethologists suggested that threat displays were honest signals of aggressive intent that benefited the species by preventing costly fights, but, as we have seen in chapter 2, the group-selectionist overtones of this idea mean that it is no longer taken seriously. Moreover, standard game-theoretic predictions (Maynard Smith, 1982) suggest that in contest situations, it will not be evolutionarily stable for animals to exchange signals of strength or aggressive intent because would-be honest signallers will always be less fit than bluffers. According to this perspective, there is no room in the arena of animal contests for the co-operative exchange of arbitrary signals; the aggressive displays observed in nature are either unfakeable because of physical constraints, or are the uninformative result of a manipulation arms race. On the other hand, some theorists (Enquist, 1985; Hurd, 1997b) have argued that, in effect, competing animals share enough of a common interest in avoiding serious injury that honest signalling can be evolutionarily stable. In this chapter we will attempt to decide between these two conflicting views.

8.1 Signalling in animal contests

8.1.1 Game theory: hawks and doves

It is interesting that the very first applications of game-theoretic modelling in biology were directed at the problem of animal contests (e.g., Maynard Smith & Price, 1973). In section 2.3.3 we reviewed the war of attrition model, as described by Maynard Smith (1982); from this model we can derive the basic prediction that signals of aggressive intent or high motivation will always be vulnerable to invasion by bluffers and thus not evolutionarily stable.

Another historically important model of animal contests is the hawk-dove game, also due to Maynard Smith. In this very simple game we postulate a population in which pairs of animals commonly engage in disputes over an indivisible resource—gaining the resource is worth V units of fitness. The contestants can adopt one of two strategies: hawk or dove. (Note that the model is not about competing populations of hawks and doves, but about more or less aggressive behaviour within one species.) An animal that plays the hawk strategy will fight until it wins the resource or until it is seriously injured; the latter outcome involves a cost of C fitness units. An animal that plays the dove strategy, on the other hand, tries to gain control of the resource by producing a threatening display, and will retreat if actually attacked. In the usual terminology, hawks are willing to escalate, while doves attempt to settle the contest through conventional (display) behaviour. Thus, when two hawks meet, one will gain the resource and the other will be seriously injured. We assume that when two doves meet one wins the signalling duel and gains the resource, and the other retreats without being injured. When a dove meets a hawk the dove retreats immediately, suffering no injury, and the hawk gets the resource. Note that the contests are symmetrical: no animal is a more capable fighter or a more threatening signaller than any other. This means that when hawk meets hawk, or dove meets dove, the winner is determined randomly.

Maynard Smith demonstrated that if the resource is worth more than the cost of serious injury (if $V > C$), then the only ESS is to play hawk all the time. In other words, when the stakes are high enough, constant and extreme aggression will be the order of the day. However, when being seriously injured costs more than the resource is worth—i.e., $V < C$, a reasonable condition as regards many real-world contests—then things are more interesting. It turns out that the only ESS is a mixture of the hawk and dove strategies, realized either as a polymorphism involving individuals who always play hawk and others who always play dove, or as a population of individuals who sometimes play hawk and sometimes play dove. Thus, if the risks of physical combat are greater than the rewards, then animals will be reluctant to escalate and will often be content with a threat display. Constant escalation would not be a stable strategy: in a population full of hawks, an individual can expect to win the resource half the time, but will pay the greater costs of being seriously injured the other half of the time. A single dove-playing mutant will never win the resource, but will nevertheless do better than the majority because it will avoid the high costs of injury.

It is important to recognize that the hawk-dove game, although extremely simple, provides a possible explanation for the fact that animals do not always fight to the death in situations where their interests conflict. That is, animals sometimes avoid escalation and engage in display behaviour because constant escalation does not pay, in terms of individual fitness. Furthermore, this explanation does not involve proper signalling. The two antagonists in the hawk-dove game are identical and make only one strategic choice in the course of a contest. There are therefore no properties like fighting ability or intention to attack that they might conceivably be communicating about, and the structure of the game allows no room for information transmission anyway. This tells us that just because many real-world animal contests are settled with what appear to be "threat signals" (rather than being physically fought) that does not mean that any true signalling is occurring.

8.1.2 Conventional signals of strength or aggression

The term "conventional signalling" is used here to refer to situations in which there is no physically necessary link between a signal's form and its meaning. Thus the signals used by different classes of signaller could in theory be exchanged and the system would still be stable. For example, vervet monkey alarm calls are probably conventional signals: the precise noises involved in the leopard, snake and eagle alarms are only arbitrarily connected with their referents. In contrast, low-frequency sounds as signals of large size, as apparently exhibited by species as diverse as red deer and Túngara frogs, are definitely not conventional signals. There is an obvious physical connection between large size and the ability to produce low-frequency vocalizations. In a signalling system based on this principle, reversing the meanings and having *high*-frequency signals represent large size would not be viable.

It is theoretically uncontroversial that if a signal is for some reason physically unfakeable in this way, then it will be evolutionarily stable for animals to use the signal to settle contests (see section 2.3.3). If the frequency of a threatening growl or roar gives reliable, unfakeable information about which of two animals is the larger, then it is logical that the smaller animal will retreat and avoid the costs of fighting a losing battle. Escalation is to be expected only when opponents are well-matched. This appears to be what is happening in the case of red deer stags: both the roaring contest and the parallel walk display are, arguably, ways of transmitting unfakeable information about size, strength and condition.

The similar use of conventional signals is not expected to be evolutionarily stable, because conventional signals, being arbitrarily linked to whatever it is they signify, are fakeable. If the signal for "I am strong" is something arbitrary, such as, let us say, blinking twice, then weak animals will be able to blink twice as well as strong ones, and the signalling system will collapse into the familiar cycle of bluffing by signallers and disregard on the part of receivers. However, this conclusion is apparently at odds with the ethological data: sometimes animals do seem to pay attention to conventional threat displays that could be (and are) faked. For instance, a mantis shrimp, its exoskeleton soft and vulnerable after moulting, can successfully drive off an intruder that could have defeated it in combat (Adams & Caldwell, 1990). The newly moulted shrimp achieves this through the use of a meral-spread display that can be produced by any individual regardless of condition, and is thus a conventional signal. The literature on bird behaviour provides additional evidence: despite the findings of Caryl (1979) that earlier authors were wrong to attribute informational value to the threat displays of certain bird species, subsequent work has not always backed this up. For example, Nelson (1984) showed that some aspects of the territorial threat displays of the pigeon guillemot *Cepphus columba* do in fact predict subsequent behaviour—it would be difficult to argue that the hunch-whistle, neck-stretch and trill-waggle displays used by the guillemot as threats are physically linked to high RHP or high aggressive intent, and so they must qualify as conventional signals. Hansen (1986), Dabelsteen and Pedersen (1990) and Waas (1991) describe similar instances of conventional signalling during conflicts in bald eagles, blackbirds and little blue penguins respectively. If animals really do use arbitrary signals to exchange accurate information in situations where their interests conflict, then the traditional game-theoretic models would have to be revised or abandoned.

Enquist (1985) presents a game-theoretic model which purports to account for the use of conventional signals in animal contests. As in the hawk-dove game, two animals struggle for control of an indivisible resource. Unlike the hawk-dove game, Enquist's model allows for the possibility of asymmetries in RHP: contestants are either strong or weak. Contestants are assumed to know their own strength, but they cannot perceive the strength of their opponent. The model divides contests into two stages. Firstly, cost-free signals are exchanged. There are two possible signals, A and B, and they are conventional because strong and weak animals are equally able to produce them. In the second stage, each animal decides to fight, pause-and-then-fight, or flee, based on its own strength and the type of signal sent by its opponent. The result of the contest is then assessed. Enquist concludes that, under certain conditions, it will be evolutionarily stable for the contestants to use the round of conventional signals to send honest information about fighting ability. At the equilibrium, escalated fights will occur only between evenly-matched opponents, and weak animals will defer to signals denoting strength.

The conditions derived by Enquist for the stability of this equilibrium are that $\frac{V}{2} - C > V - D$, where V is the value of the resource, C is the cost of an escalated fight between two equally matched opponents, and D is the cost to a weak animal of being attacked by a strong one. Rearranging terms, we get $D > \frac{V}{2} + C$, which means that the cost to a weak animal of being attacked by a stronger one must be appreciably greater than the cost of fighting another weak individual. In this sense we can say that Enquist's conclusion—that conventional signalling of fighting ability can be evolutionarily stable—is driven by the assumption that weak animals cannot afford to risk confronting a stronger opponent and thus must be honest about their shortcomings. Ultimately, the truth of this assumption is a matter for empirical investigation, but the simulation presented in this chapter will allow us to judge its plausibility given a semi-realistic model of animal combat.

In a second, related model, Enquist goes on to claim that the cost-free signalling of "local strategy", i.e., aggressive intent, can also be evolutionarily stable. Hurd (1997b) has recently extended Enquist's first model and also concludes that the cost-free signalling of fighting ability is possible. Furthermore, says Hurd, if only two signals are available, and if they vary in cost, it will be evolutionarily stable for weak animals to use the more costly of the two signals: not as a bluff, but as an honest advertisement of low fighting ability. These paradoxical results clearly run against the grain of most game-theoretic predictions about signalling in cases where the interests of the interacting parties conflict.[1]

[1] The implication here is that it is controversial to conclude, as Enquist and Hurd have done, that it can be an ESS to use conventional signals of fighting ability in animal contests—although clearly it is not all that controversial, as Enquist's (1985) model was endorsed in Johnstone's (1998) recent review of the literature on game theory and communication. The reader should note that we are dealing with situations in which memory-less competitors of varying RHP play out contests with randomly-determined opponents whose RHP they cannot perceive. However, if other factors are taken into account, as in van Rhijn and Vodegel's (1980) model which incorporates individual recognition and repeated interactions, stable conventional signalling of RHP or motivational state may be much easier to establish.

8.1.3 Sustainable bluffing

Orthodox game-theoretic accounts suggest that conventional signals of RHP will always be vulnerable to invasion by bluffers. Let us define bluffing as the use of a signal by a weak individual that would conventionally denote strength. Some authors have constructed models in which bluffing occurs but does not lead to the collapse of the signalling system. Both of the models that will be discussed were inspired by the case of the mantis shrimp *Gonodactylus bredini*. This animal is one of the most well-documented bluffers in nature: although the meral-spread threat display is used by high-RHP cavity owners to deter intruders, it is frequently used by vulnerable inter-molt individuals, who have the lowest levels of RHP (Adams & Caldwell, 1990).

Gardner and Morris (1989) developed a game-theoretic model of mantis shrimp contests that incorporated a dual asymmetry in information and fighting ability. Intruders were assumed to be strong, and thus cavity owners were in no doubt as to the strength of their opponent. Cavity owners were either strong or weak (inter-molt), and knew their own strength, but intruders were unable to perceive the status of the cavity owner. The contest proceeded in two stages: based on its strength, the cavity owner decides whether to produce a threat display or to flee immediately. If the owner flees, then the intruder automatically gains the cavity. But if the owner displays, then the intruder must decide whether to fight or flee. Gardner and Morris considered two costs: C, the cost of losing a fight, and S, the cost of bluffing. They established that if both C and S were low relative to V, the value of gaining possession of the cavity, then there would be no ESS, but the population would cycle through a dynamic equilibrium that included periods of bluffing and periods of relative honesty. The authors suggest that this model may explain the behaviour of the mantis shrimp.

Adams and Mesterton-Gibbons (1995) constructed a similar model, although they made RHP a continuous value and considered variation in the strength of both contestants. They argued that the use of a threat by a weak animal involves a “vulnerability cost”. This is one version of Zahavi’s handicap principle, which asserts, as the reader will recall, a necessary relationship between a signal’s cost and its reliability. A vulnerability cost exists because, if the threat does not work, a weak animal is likely to be seriously injured by its probably-stronger opponent. On the other hand, weak animals stand to gain proportionately more if the threat is successful, because their chances of winning an escalated contest are low. Adams and Mesterton-Gibbons predict that bluffing will be stable: threat displays will be produced and, depending on the RHP of the recipient, sometimes heeded; however, these displays should be expected from the very strong *and* the very weak.

Although the simulation presented in this chapter is not directed specifically at mantis shrimp behaviour, the predictions made by Gardner and Morris (1989) and Adams and Mesterton-Gibbons (1995) can be tested in at least a qualitative fashion.

8.1.4 Intention movements and ritualization

There is another motivation for the work presented here. Whilst the honest signalling of intentions looks questionable from a game-theoretic perspective, it has been cogently argued by both ethologists (Tinbergen, 1952) and game-theoretically inclined behavioural ecologists (Krebs & Dawkins, 1984) that intention *movements*—i.e., movements necessarily preceding an action, such as a dog baring its teeth in order to bite—probably function as “seeds” in signal evolution (see

	. . . plays hawk	. . . plays dove
Hawk	-1	2
Dove	0	1

Table 8.1: Payoff matrix for the hawk-dove game

sections 2.2.1 and 2.2.2). Rather than incorporating an exchange of artificial, discrete signals, the current model seeks to explore the plausibility of the intention-movements idea by using such movements as the medium of potential information transmission.

8.2 Description of the model

The current chapter presents a simulation model of contests over an indivisible resource. The results will be compared with the conflicting predictions of the models outlined above. The main aim of using a simulation, rather than a more formal approach, is to avoid oversimplification. In particular, time will be modelled in an approximately continuous fashion: game-theoretic models of aggressive signalling rarely allow for more than two time-steps—an exchange of signals followed by a choice of actions—and thus may fail to capture critical aspects of real-time interactions.

The hawk-dove game provides a useful introduction to the simulation. Maynard Smith (1982), in discussing the game, chose real values of V and C in order to simplify some of the calculations: $V = 2$ and $C = 4$. Note that setting the cost of injury higher than the value of the resource in this way means that the ESS will be a mixed-strategy equilibrium. Table 8.1 shows an expected payoff matrix for the hawk-dove game worked out using these values. The expected payoff for a hawk playing a hawk, for example, is -1, because half of the time such a player will win and gain 2 units of fitness, whereas the other half of the time they will lose, costing them 4 units.

Imagine that you are playing the hawk-dove game. Inspecting table 8.1, we can see that if you knew that your opponent was going to play hawk—i.e., that they were in some way absolutely committed to the hawk strategy—then you would do better to play dove. Your expected payoff would increase from -1 to 0. Similarly, if you knew your opponent was going to play dove, it would be more profitable for you to play hawk. The simulation poses the question as to what would happen if players of something like the hawk-dove game could each perceive the strategic choices of their opponent. As noted above, the simulation incorporates intention movements and continuous time; thus, the decision to play a hawk-like or a dove-like strategy is not an instantaneous choice but involves temporally extended action. It is not clear that the standard game-theoretic predictions will apply when there is this potential for the exchange of information about strategy choice.

The model involves two simulated animals contesting the possession of a resource. Due to a shortage of hard data on the metabolic costs of fighting (although see Riechert, 1982), the simulation does not model contests in a particular species. However, the mantis shrimp will serve as an illustration: we assume that two shrimps have discovered a desirable cavity.[2] Each shrimp

[2] Note that mantis shrimps are only used here as a rough way of describing the model. Unlike real mantis shrimps, the simulated animals only ever have RHP asymmetries. They never experience role asymmetry, such as the advantage that might accrue to the current cavity owner in a real shrimp contest. Nor do they experience informational asymmetries,

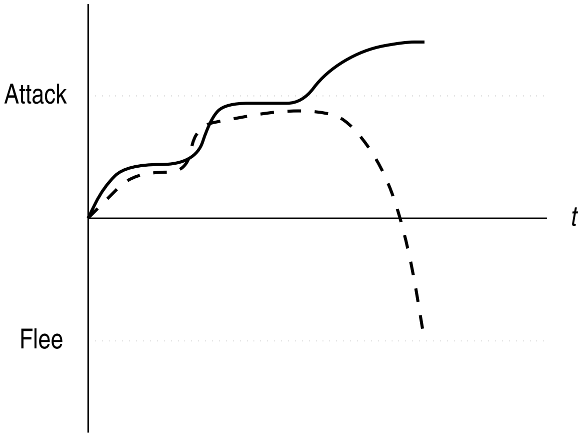

Figure 8.1: Course of a hypothetical contest: the horizontal axis is time and the vertical axis is the Θ-continuum. The weaker animal (dashed line) flees after the stronger animal shows a greater willingness to back up "threat displays" by attacking.

knows its own fighting ability, but is unable to perceive the ability of its opponent. Each shrimp can perceive movements towards aggression or retreat on the part of its opponent. Based on this information, shrimps can elect to attack or to flee or to do something in between. If the shrimps engage in an all-out fight, both shrimps will suffer costly injuries but the stronger one is likely to win and gain possession of the cavity. Over generational time, the shrimps may or may not evolve a signalling system, based on the observation of intention movements, that allows them to settle contests with a minimum of escalated fighting.

The model is based on a single behavioural continuum between attacking and fleeing. An animal is always located at some point Θ on this continuum. Contests involve two animals randomly selected from the population; animals begin the contest at $\Theta = 0$ and, each time-step, can move a maximum of δ units in either direction. Movement towards positive values of Θ is movement towards aggression, while movement in a negative direction constitutes retreat. However, values of Θ that are close to zero do not constitute a definite action of any kind, and might conceivably be used as signals. Only when an animal has $\Theta > A$ is it deemed to be physically attacking its opponent. Conversely, when $\Theta < F$ the animal flees: the contest ends and the contested resource goes to the opponent. In the runs presented here, $F = -A$ and $\delta = A/4$. The simulation captures intention movements in that animals can neither immediately attack nor immediately flee: from the starting position, it takes at least four time-steps to do either. A plausible threat display might be to "hover" with Θ just less than A, indicating a readiness to attack. A display of weakness or timidity might involve moving to a value of Θ close to F, and thus preparing to flee. Figure 8.1 shows the time-course of a hypothetical contest.

The simulated animals have associated with them an "energy" level e that is set to 0 at birth. Energy is the common currency of the model; reproductive fitness is achieved by having, at the end of the day, greater energy reserves than one's conspecifics. The animals also possess a fighting ability f $(5 \leq f \leq 15)$, which equates to RHP, and is randomly assigned at the beginning of *each*

as in Gardner and Morris's (1989) model in which owners know that intruders are strong, but intruders do not know the strength of the owner.

contest. Note that heritable fighting ability would quickly lead to an uninteresting fixation on high values; the model selects for responsiveness to varying values of f_{self}, as presumably exists in any animal that must moderate its contest behaviour according to its own condition, age or status.

The two costs and one benefit of the contest affect the value of e. The first cost is due to being attacked: at each time-step that $\Theta_{opponent} > A$, the animal suffers an injury cost of $-f_{opponent}$. In other words, high RHP manifests itself as an ability to inflict greater damage on one's opponents—consequently, the stronger animal will always win an escalated fight, assuming that both contestants attack each other simultaneously and consistently. The second cost is an energy cost for aggressive display or attack: at each time-step that $\Theta > 0$, the animal pays a cost of $-k\Theta/A$, with $k = 1$. So, for example, an animal attacking its opponent with $\Theta = 1.1 \times A$ would endure -1.1 units of energy cost per time-step. Note that the cost of attacking is always much less than the cost of being attacked, for any f; this is in keeping with Riechert's (1998) observation that the fitness costs of being injured are by far the greatest of all those associated with contest behaviour. Note also that any activity where $\Theta < 0$ involves no energy cost—this is justified on the basis that backing off, preparatory to running away, is much less energetically expensive than aggressive behaviour. The only benefit in the contest is to gain control of the resource, which is worth $V = 100$ units.

The contest can end in one of three ways. One animal may flee, as discussed above. Secondly, one animal may win the contest through brute force: if an animal loses more than $C = 200$ units of energy *during any one contest*, it has been physically overcome by its opponent. The contest ends immediately and the opponent gains the resource. Regrettably, the value of C puts an artificial cap on the amount of damage an animal can sustain in one contest; however, the values of C and V have been chosen such that, on the face of it, the resource is worth having but not worth suffering serious injury for. In the hawk-dove game, as we have seen, these relative cost and benefit values result in a mixed-strategy equilibrium in which hawk and dove are each played half of the time. Finally, the contest can end because a time limit, $t_{max} = 50$ time-steps, has been reached, in which case neither animal gains the resource. The values of t_{max}, C, k, and δ, and the range of values of f, have been co-ordinated such that it is possible for even the weakest animal to overcome an opponent within the time limit. Although a degree of arbitrariness is inevitable in setting parameter values for an abstract simulation, it is hoped that this co-ordination of values will at least prevent such disasters as, for example, enforcing honesty amongst the weak by having C too high, or t_{max} too low, for a weak animal to ever win by fighting.

The simulated animals have as sensory inputs f_{self}, Θ_{self}, and $\Theta_{opponent}$. Informally, they know their own strength, they can see what they're doing, and they can see what their opponent is doing. The animals also have access to a random input, to allow for probabilistic strategies. The animals produce a continuous output in the range $\pm\delta$ which is applied to their Θ-position for the next time-step.

The animals were implemented as five-neuron fully inter-connected continuous-time recurrent neural nets (CTRNNs), with the activity of neuron 0 taken as the output. CTRNNs are among the most general of artificial neural network architectures. The recurrent aspect of the nets makes it possible for the animals to evolve some form of short-term memory rather than being purely reactive. However, no attempt has been made to determine whether this actually occurred: the CTRNNs have been treated here as a black-box control system. All parameter values for the nets were taken from Yamauchi and Beer (1994). The evolutionary engine was a genetic algorithm with a population size of 100, run for 5000 generations. In each generation, animals were randomly selected to play out 500 contests; each animal could thus expect to participate in 10 contests in its lifetime and was guaranteed at least 5. An animal's fitness score was its cumulative energy score after all 500 contests had been played, divided by the number of contests it had actually participated in. For breeding purposes, these fitness scores were normalized as deviations from the mean: animals with negative scores were discarded, and roulette-wheel selection was applied to the remainder.

Two control conditions were devised: in the "blind" condition, the animals are denied access to $\Theta_{opponent}$ and therefore any communication is impossible. In the "unfakeable" condition, the animals are given access to $f_{opponent}$, and can be thought of as exchanging unfakeable signals of strength.[3] In order to facilitate comparisons with previous models, both continuous and discrete distributions of fighting ability were used. In the continuous fighting ability (CFA) case, f was uniformly randomly distributed between 5 and 15, while in the discrete fighting ability (DFA) case animals were either weak ($f = 5$) or strong ($f = 15$). The experimental and control conditions were crossed with the two f-distributions to make six conditions in all.

8.3 Hypotheses

If Maynard Smith's (1982) analysis of the hawk-dove game can be broadly applied to this more complex contest situation, then we would expect to find—in the experimental condition—a mixed strategy equilibrium and no communication. That is, the animals will sometimes hawkishly attack, and at other times adopt a more dove-like strategy, perhaps "displaying" with $\Theta = 0$ and then retreating if challenged. Of course, the choice of strategy is likely to be modified by the animal's perception of its own absolute fighting ability: individuals with high RHP are, on average, more likely to profit from aggressive strategies. No communication regarding fighting ability is expected, because of the standard logic that bluffers will always be fitter than honest signallers. However, in the unfakeable control condition, as noted above, the contestants have direct informational access to each other's fighting ability. Therefore the unfakeable control should stand as an index of what would happen if reliable signalling of RHP *were* occurring: escalated fights should occur only when two animals are closely matched, and at other times the weaker animal should defer immediately. Overall this would lead to lower levels of energy expenditure in fighting. In the blind condition, of course, no communication can occur, and we would expect energy expenditure to be relatively high, as contestants choosing a hawk-like strategy will have no way of knowing that their opponent is doing the same. Energy expenditure in the blind control provides an index of how well the population can do when their behaviour is completely un-co-ordinated and signalling is impossible; if energy expenditure in the experimental condition were to be equally high, then this would be evidence that no signalling was occurring.

[3] Note that the CTRNNs that made up each animal's "brain" were consistently given five inputs across all conditions. In the unfakeable condition, the five inputs were as follows: a random value, Θ_{self}, $\Theta_{opponent}$, f_{self}, and $f_{opponent}$. Rather than introducing inconsistency by having a different number of inputs in the other conditions, inputs that were not relevant were fixed at a constant neutral value. So, in the experimental condition the value of $f_{opponent}$ was set to 0.5—the inputs being scaled between zero and one—and in the blind condition both $f_{opponent}$ and $\Theta_{opponent}$ were set to 0.5.

On the other hand, if Enquist's (1985) and Hurd's (1997b) views are correct, then a system for exchanging reliable signals of RHP is likely to evolve in the experimental condition. This would be driven by the need of weaker animals to avoid the risks of an escalated confrontation with a stronger opponent. Thus, the overall pattern of behaviour is expected to be similar in both the experimental and unfakeable conditions, because a reliable communication system evolves in the former case and is enforced in the latter. More specifically, the average level of energy expenditure should be similar across the two conditions—in both cases, the signalling system means that escalation never occurs between unequal opponents. Furthermore, if Hurd is right in suggesting that cost-free signals of fighting ability will be used in preference to costly ones, then we might expect reliable conventional signals to be manifested as displays of some sort characterized by $\Theta < 0$, because behaviour in this region of the continuum carries no energy costs at all.

Predictions arising from the work of Gardner and Morris (1989) and Adams and Mesterton-Gibbons (1995) are less clear-cut. The dynamic equilibrium described by Gardner and Morris, in which bluffing plays a part, is supposed to occur only if the cost of losing a fight (C) and the cost of bluffing (S) are both low relative to the value of the contested resource (V). In the simulation this is apparently not the case: the injury cost of losing an escalated fight is 200 units of energy, compared to 100 units for gaining the resource. The cost of bluffing is not something that has been built into the simulation, but will emerge from the behaviour patterns of the population over time. Assuming for a moment that it might be reasonable to apply Gardner and Morris's results in a context different from the one in which they were derived, they would in fact suggest that if the cost of bluffing turns out to be high enough, then the ESS will involve honest signalling; this parallels Enquist's argument that it is not worthwhile for a weak contestant to pretend to be strong. If the cost of bluffing is low (relative to V) then Gardner and Morris predict a regime in which *all* animals bluff: this is equivalent to Maynard Smith's idea that it will not be evolutionarily stable to send honest signals of strength.

However, whereas Gardner and Morris postulate costs for bluffing that *all* bluffers must pay, Adams and Mesterton-Gibbons explicitly state that in their model bluffing only carries a cost when it is unsuccessful, i.e., when the opponent does not believe the false signal of high RHP. This seems a more reasonable assumption to apply in the case of the simulation. Clearly, if a signalling system evolved and bluffing occurred, the success of an attempted bluff would be judged by whether or not the opponent "bought" the bluff and fled without a fight; the bluffer would thereby escape the high costs of being injured. Given this assumption, Adams and Mesterton-Gibbons calculate that at the equilibrium threat displays will be produced by the strongest and the weakest members of the population. In terms of the simulation, this would mean that animals with very high and very low (but not intermediate) values of f perform some sort of characteristic display behaviour that tends to scare off all but the strongest opponents.

8.4 Results

Each of the six experimental conditions was run 10 times with a different random seed. The final 1000 generations of each 5000-generation run were used as a window period for statistical analysis; all of the results below refer to average behaviour within this period unless otherwise

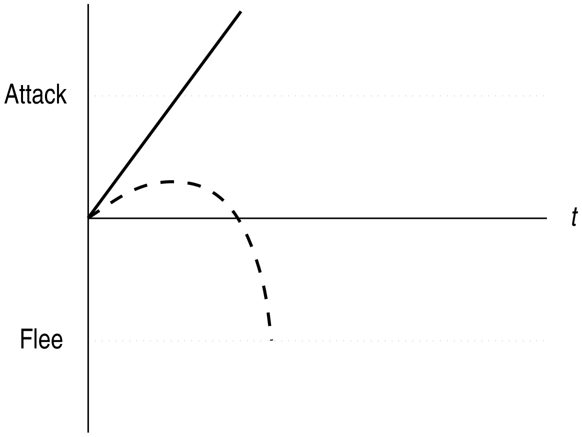

Figure 8.2: A pattern observed in some contests. The stronger animal (solid line) moves rapidly to attack its opponent. The weaker animal (dashed line) begins with a moderately aggressive move but then appears to "change its mind".

stated. Standard errors refer to error across the 10 trials.

In observing the progress of the contests on a computer monitor, the overwhelming impression was that most of the animals wasted no time in moving rapidly towards either the attack or the flee line. Contests were generally resolved quickly: the mean duration was 11.77 time-steps across the six conditions, and less than 0.3% of contests reached t_{max}. If threat displays were occurring, they did not involve sustained action.

The behaviour of the contestants was often but not always ballistic in character; whereas many animals pursued a constant gradient on the Θ-continuum over the course of a contest, others did not. One pattern that was observed is shown in Figure 8.2. In the figure, a stronger animal moves rapidly and consistently towards aggression and attack, while a weaker one starts with a moderately aggressive move but then appears to "change its mind" and flee. This immediately suggests the possibility that the weaker animal is responding to the show of strength from its opponent, and that the initial aggressive move by the weaker animal might count as a bluff or even as an honest signal of weakness. However, providing statistical justification for qualitative sketches of this kind proved difficult.

Figure 8.3 shows the mean fitness values across conditions, i.e., the average loss or gain in energy per individual per contest. As noted in section 8.3, these data are especially important in determining whether or not a signalling system has evolved. Energy usage per contest depends on the value V of the resource, which an animal can expect to obtain about half the time, less the mean costs of aggression and of being injured. In an ideal signalling system, where cost-free signals of strength were exchanged and the weaker animal always retreated immediately, the mean fitness would be $V/2 = 50$ (assume that equally matched animals allocate the resource randomly). In the experimental condition, such ideal situations clearly did not evolve—the negative mean implies that the $V/2$ expected payoff was balanced against greater energy and injury costs.[4] The

[4] It may seem strange that the mean energy payoff in the experimental condition could be negative. After all, a strategy of complete cowardice, in which an animal ran away from all contests as quickly as possible, would mean never winning the resource, but it would also guarantee not bearing any costs due to injury or aggressive display. A cowardly mutant could therefore expect an energy payoff of zero, and would do relatively well in the experimental

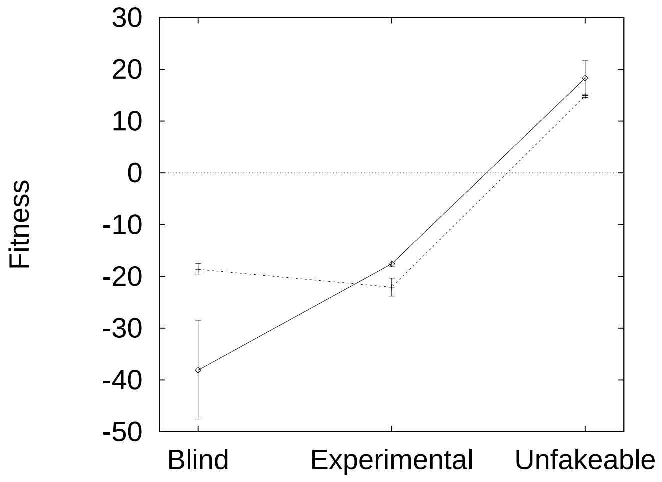

Figure 8.3: Mean fitness (energy payoff) ±1 s.e., for CFA (solid line) and DFA (dashed line), across the three conditions.

unfakeable condition provides an index of how efficiently the animals can allocate V when they have reliable, cost-free information about their opponent's strength, and in practice their efficiency approaches 40% of the ideal. The fact that mean fitness in the experimental condition was significantly lower than in the unfakeable condition (CFA: $t_{18} = 10.57$, $p < 0.001$; DFA: $t_{18} = 20.90$, $p < 0.001$) is strong evidence that the sort of reliable cost-free signalling system described by Hurd and Enquist did not evolve.

It is possible in principle that communication about RHP was occurring but that, because the signals were costly or unreliable, energy expenditure due to fighting was nevertheless greater than in the unfakeable control. If this was so, then we would expect fitness to be higher (i.e., energy expenditure to be *lower*) in the experimental group than in the blind control group—the blind control shows us what happens when communication and co-ordination of any kind is completely prevented because the animals cannot perceive each other's movements. Figure 8.3 suggests that, at least in the CFA case, the animals are indeed doing better in the experimental condition than in the blind control. Assuming an alpha-level of 0.05, the difference in means is of marginal statistical significance ($t_{18} = 2.13$, $p = 0.047$); in the DFA case there is no significant difference ($t_{18} = 1.68$, $p = 0.11$). What does this result mean? Clearly, being able to observe one's opponent's intention movements makes a difference in the amount of energy that is "wasted" on combat, although only when fighting ability is continuously distributed. However, this is not the same thing as establishing that signalling is occurring.

Figure 8.4 gives us another way of comparing the experimental and control conditions. The figure shows the percentage of contests that were resolved by all-out fights, i.e., by one animal overcoming the other (the remainder were almost all resolved by one animal fleeing). In the unfakeable condition, as expected, fights were relatively rare: if animals can perceive the strength of their opponent, and if the animal with higher fighting ability will always win an escalated fight, then fights should only occur when contestants are so closely matched that they cannot tell which

condition. However, such cowardice cannot be an ESS: a population of cowards represents a great opportunity for a mutant that simply waits around for its opponent to flee.

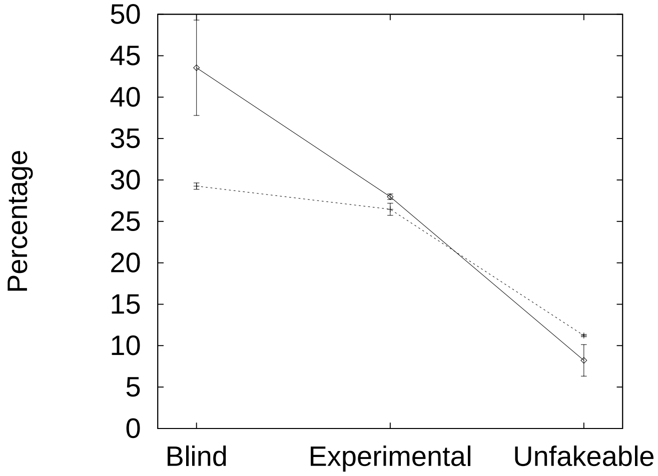

Figure 8.4: Mean percentage of contests resolved by all-out fights ±1 s.e., for CFA (solid line) and DFA (dashed line), across the three conditions.

one is stronger. In the other two conditions, escalated fights were more common but still occurred less than half the time. However, in the CFA version of the experimental condition, fights occurred 28.0% of the time, which was significantly less often ($t_{18} = 2.70$, $p = 0.015$) than in the CFA version of the blind control, where they occurred 43.5% of the time. Why were there fewer fights in the experimental condition? Again, a possible explanation is that a less-than-perfect signalling system is in place, allowing the animals to avoid costly fighting some of the time. However, this interpretation is mitigated against by the fact that when fights did occur in these two conditions, they were equally likely to be between closely matched opponents ($|f_a - f_b| < 2$). Specifically, in the CFA experimental condition, 37.0% of fights were between well-matched opponents, and in the CFA blind control the figure was 36.1% ($t_{18} = 1.80$, $p = 0.089$). If a signalling system were in place, it would presumably lead not only to fewer fights overall but to a greater proportion of well-matched opponents when fights did take place.

Overall, stronger animals were more aggressive than weaker ones; stronger animals were more likely to have higher Θ-values than weak animals at any particular time-step in the contest. For example, Figure 8.5 shows the mean Θ-positions of various categories of contestant throughout the contests that occurred in generation 5000 of the first run in the CFA version of the experimental condition. Note that the position axis is scaled such that $A = 100$; thus the average movement was toward aggression in all categories. Not only were stronger animals more aggressive than weaker ones, but there is a suggestion that strong animals were more aggressive when their *opponent* was strong too, and similarly that weaker animals were more aggressive against strong than against weak opponents. Figure 8.6 deals with the DFA version of the experimental condition (run 1), and shows the evolution of typical behaviour at time-step 5, an early point in the contest. The figure shows that the mean Θ-position of strong contestants was dramatically different from that of weak ones. The position axis is scaled in the same way as in Figure 8.5, so we can see that, after an initial era of neutral behaviour, strong animals evolve an extremely aggressive strategy, whereas weaker animals are likely to be neutral (on average) against weak opponents but to retreat somewhat from stronger opponents.

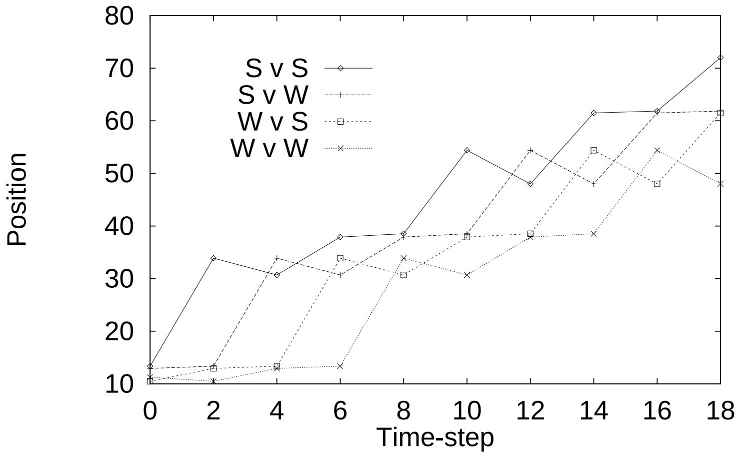

Figure 8.5: Typical behaviour through the course of a contest. Experimental condition with CFA; run 1 of 10 at generation 5000. Contestants have been divided at the mean into strong (S) and weak (W) categories of fighting ability, and the graph shows the mean Θ-positions of strong contestants against strong opponents, strong contestants against weak opponents, and so on.

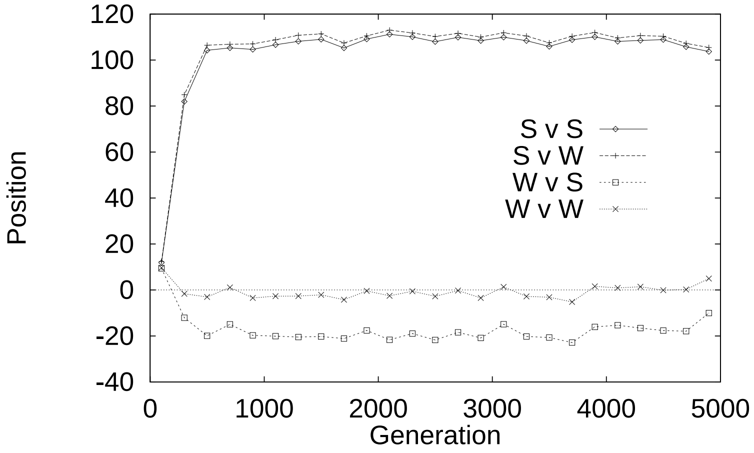

Figure 8.6: Evolution of typical behaviour at time-step 5 over generational time. Experimental condition with DFA; run 1 of 10. Contestants are either strong (S) or weak (W). The graph shows the mean Θ-positions of strong contestants against strong opponents, strong contestants against weak opponents, and so on. Data points on the time axis have been clumped into blocks of 200 generations and a mean value plotted.

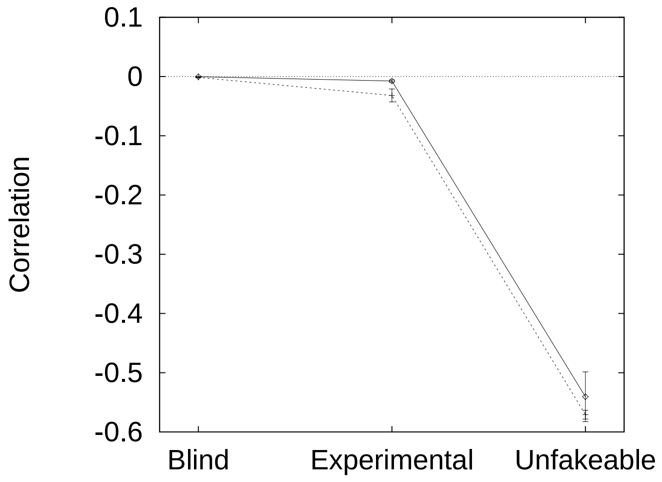

Figure 8.7: Mean correlation between Θ_{self} and $f_{opponent}$ ±1 s.e., for CFA (solid line) and DFA (dashed line), across the three conditions.

Whilst it is not surprising that stronger animals should be more aggressive than weaker ones, these findings present us with a conundrum. If fighting ability is linked to the mean expected position of an animal at a particular time-step in the contest, then should not the contestants themselves have been able to discover this relationship and exploit it? The data in Figures 8.5 and 8.6 suggest that to some extent they have done so; recall that in the experimental condition the animals cannot perceive the strength of their opponent. The fact that. for example, the weak animals in Figure 8.6 are more likely to retreat when their opponent is strong indicates that they are registering the behaviour of their opponents and responding accordingly. But does this constitute communication?

If a communication system existed, and the animals were able to assess the strength of their opponents, then we would expect a generally negative relationship between an animal's Θ-position and the strength of its opponent. That is, the animals would take in the opponent's signals of RHP and then respond aggressively towards weaker opponents and run away from stronger ones. Figure 8.7 looks at the the mean correlation coefficient between Θ_{self} and $f_{opponent}$ at time-step 4, when all contests were still in progress; the correlation was calculated separately for each generation and then averaged. In the unfakeable condition there is a negative relationship: unsurprisingly, animals that could reliably perceive their opponent's strength were likely to flee from stronger opponents. In the blind condition there is of course no relationship at all. In the experimental condition, there is the merest suggestion of a negative relationship, but it accounts for much less than 1% of the variance in Θ_{self}. It would appear that, whether or not we label it as communication, and despite the definite relationship between an animal's fighting ability and its expected Θ-position, there is not a great deal of information being transmitted about RHP.

A general link between f_{self} and Θ_{self} also suggests the possibility of bluffing, i.e., deception by moving to a Θ-value usually characteristic of higher f. Given the brevity of the contests and the likely importance of first impressions, bluffing has been investigated by tabulating the animals' opening moves (their $\Delta\Theta$ for time-step 1). Figures 8.8 and 8.9 show the frequencies of opening moves by fighting ability for CFA and DFA respectively. Substantial variation existed in these relationships across trials, and presenting mean values would obscure the situation—the two

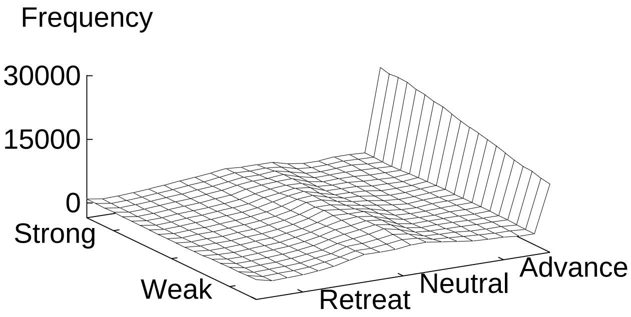

Figure 8.8: Total frequencies of opening moves by fighting ability, with each variable grouped into 20 bins for plotting. The plot shows data for the window period of run 5 in the experimental condition with CFA.

figures show data from typical runs.

In both cases there is evidence of stereotypical, extreme responses. In the CFA case (Figure 8.8) animals tend to make either an extremely aggressive first move or a neutral one. Contestants with higher fighting ability are more likely to do the former. Thus, the most frequent opening move for a contestant of maximum RHP ($f = 15$) is to move towards the attack line as rapidly as possible; this happens about 42% of the time. The same move is performed about 25% of the time by the weakest animals, in which case it seems reasonable to describe the advance as an attempted bluff: a weak animal is behaving in a way that usually characterizes a strong one. Figure 8.9 shows what happens when fighting ability is discretely distributed and animals are either weak or strong. The weak animals play a range of neutral first moves, whereas strong ones almost always behave with maximum aggression. In contrast to the CFA results, there is almost no evidence of bluffing; i.e., the weak animals almost never make highly aggressive first moves.

If bluffing in the experimental condition was occurring *successfully*, then this would presumably result in a relatively high proportion of contests being won by weaker animals: bluffing is impossible in the blind control and pointless in the unfakeable. Figure 8.10 shows the percentage of contests in which the weaker animal gains the resource (note that the DFA results are not directly comparable as half of the time there *was* no weaker animal). While the experimental condition, unsurprisingly, leads to more "upset wins" than the unfakeable condition (CFA case, $t_{18} = 9.63$, $p < 0.001$; DFA case, $t_{18} = 2.21$, $p = 0.041$), the difference between the experimental and the blind conditions is of marginal significance. This suggests that we do not need to invoke the hypothesis of successful bluffing in the context of a signalling system in order to account for the results in the experimental condition.

8.5 Variations on the model

Three variations on the basic model were devised, with the intention of finding out more about the conditions that might foster communication.

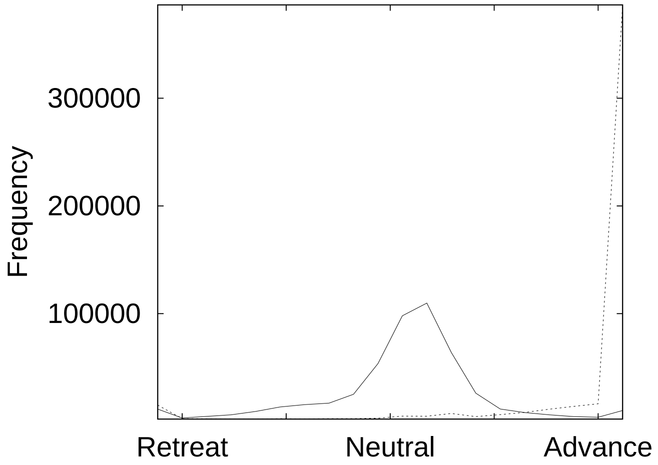

Figure 8.9: Total frequencies of opening moves by weak (solid line) and strong (dashed line) fighting ability, with the move data grouped into 20 bins for plotting. The plot shows data for the window period of run 1 in the experimental condition with DFA.

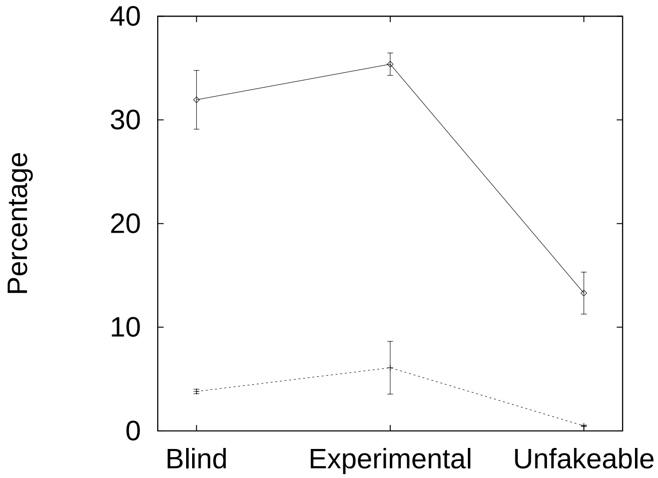

Figure 8.10: Mean percentage of contests won by the weaker animal ±1 s.e., for CFA (solid line) and DFA (dashed line), across the three conditions.

Firstly, the model was extended so that there were two ways in which the contestants might come to communicate. In addition to the ability of the animals to perceive each other's intention movements, they were now also given an arbitrary signalling channel. The results thus far indicate that cost-free, reliable signalling of RHP does not evolve in the way that Enquist and Hurd suggest; however, this two-channel condition allows us to meet the possible objection that whereas Enquist's and Hurd's results were to do with arbitrary signalling systems, the simulation described here forces the animals to use their one-dimensional behaviour both for conducting the contest (i.e., fighting or fleeing) and possibly for signalling as well. It seems reasonable to suggest that the evolution of a signalling system might be blocked because an animal's movements up or down on the Θ-continuum cannot do double duty in this way. The use of an arbitrary signalling channel can be illustrated by imagining that competing mantis shrimps (for example) can not only pay attention to significant movements like claw-spreading, advancing and retreating, but can also observe apparently irrelevant movements such as antenna-waving. Enquist and Hurd's logic suggests that in such a situation antenna-waving will be a good candidate for the transmission of conventional signals of fighting ability. The arbitrary signalling channel was implemented by giving the animals access to the activity value of a neuron in their opponent's network; specifically, the current activation level of neuron 1 in each contestant was used as one of the five inputs to the other animal's network (the other four inputs were a random value, Θ_{self}, $\Theta_{opponent}$ and f_{self}). The activity level of neuron 0 continued to function as the main behavioural output. Thus, if there is selective pressure to use the arbitrary signal to send honest information about RHP, it would be a simple matter for a positively-weighted connection to evolve between the Θ_{self} input and neuron 1.

The second variant allowed contestants the possibility of assessing each other's fighting ability in an indirect and costly fashion. Parker (1974) originally suggested that the function of displays during contests might be to facilitate the exchange of information about RHP; not through conventional signalling, but by allowing each contestant the chance to observe and assess unfakeable cues as to the fighting ability of their opponent. However, in the basic simulation model this is not possible. Although signalling may or may not occur, the animals cannot do anything to find out for themselves how strong their opponent is, short of engaging in an escalated contest and noting how long it takes for one or the other to be overcome. Therefore a variation was constructed in which animals were given informational access to their own energy level for the current contest. Recall that once an animal has lost $C = 200$ units of energy during one contest, it has been physically overcome. In this assessment variation the animals were aware of their energy level "counting down" to -200; the value was used as the fifth input to each animal's network. It was therefore possible for them to assess the strength of an attacking opponent by noting the rate at which their own energy level was being depleted through injury. A pair of contestants could in theory test the waters by engaging in a brief mutual attack, and then either might withdraw if it had discovered that its opponent was stronger. If this sort of behaviour in fact occurred, it would be debatable whether it should be classified as communication: the proper function of A's attack on B is presumably to overpower A and gain the resource; nevertheless B can exploit information arising from the attack in order to judge A's relative strength.

Finally, the third variant was inspired by the data shown in Figures 8.8 and 8.9 which make it clear that the first move made by an animal carries some information about its fighting ability:

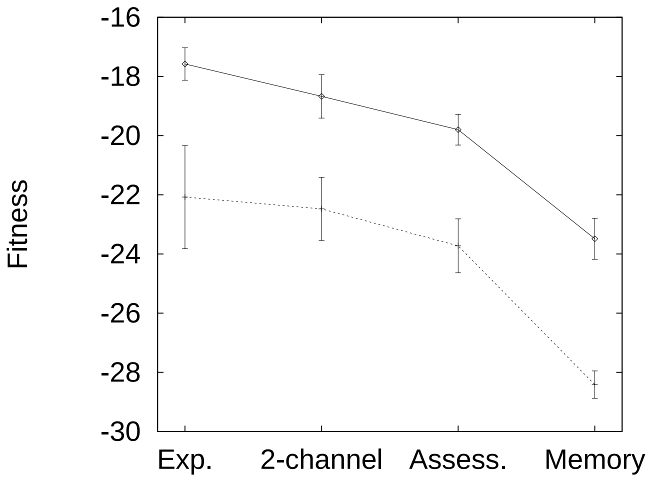

Figure 8.11: Mean fitness ±1 s.e., for CFA (solid line) and DFA (dashed line), for the experimental condition and across the three variations: the two-channel condition, the assessment-through-damage-perception condition, and the memory-for-opponent's-opening-move condition.

weak animals tend to start with relatively neutral moves, while strong ones often start very aggressively. Although the neural-net architecture of the simulated animals should have been complex enough that they could develop the ability to register and act on this information, it was decided to give them a helping hand in this variation. The animals were accordingly given a memory register which took a snapshot of the opponent's first move and then held that constant as a fifth input to the neural net (note that the memory register could not be overwritten as the contest proceeded). It was felt that this might make it easier for the animals to recognize signs of strength or weakness in their opponent and act accordingly.

Results for the three variations were disappointing. Figure 8.11 shows the mean fitness or energy payoff for the experimental condition and for the variants. Performance in terms of energy efficiency was either similar to or somewhat worse than the experimental condition. This means that such methods as providing a second signalling channel, making it possible to assess an opponent's strength, and providing animals with a snapshot memory are all ineffective in improving the animals' abilities to judge the strength of their opponent and thereby avoid some of the costs of fighting. The level of energy efficiency achieved in the unfakeable control condition remains a clear index of what can happen when animals *do* know the strength of their opponents, and this level is not approached in the experimental condition or in any variation thereof.

Results for the variations on other measures, such as the proportion of contests won by fighting, typical behaviour through the course of a contest, and the correlation between Θ_{self} and $f_{opponent}$ were generally very similar to the results for the experimental condition. The different sensory inputs available to the animals in the various conditions did not have very much influence on their behaviour.

Brief consideration of a fourth variation allows us to clear up an ambiguity raised earlier, however. In Figure 8.3 we saw that fitness in the experimental condition was significantly lower than fitness in the unfakeable condition, indicating that if any signalling was occurring in the experimental case then it was certainly not reliable and cost-free. However, fitness in the CFA experimental condition was higher than in the CFA blind control, which shows that animals must gain something from being able to perceive the intention movements of their opponent. The ambiguity is about whether, in the experimental case, they are gaining the benefit of proper signals that carry information about RHP, or whether there is a more basic benefit that comes merely from being able to see what one's opponent is doing. The variant simulation that should settle this question involves taking the blind control and adding an arbitrary signalling channel as described above. In such a situation the animals cannot perceive the movements of their opponents towards aggression or flight, but there is still the possibility of signals being exchanged through the arbitrary channel. If some sort of inefficient signalling system evolves in the experimental case, then communication should also evolve in the arbitrary-channel case, and mean fitness across the two conditions should be comparable. If, on the other hand, there is no proper signalling in the experimental condition, then there should be no signalling via the arbitrary channel either, and mean fitness in this final variation should be as low as it is in the blind control. It turns out that the latter is the case: mean fitness in the arbitrary-channel variation with CFA is -52.5 (s.e. $= 14.46$), which is significantly lower than in the experimental condition ($t_{18} = 2.42$, $p = 0.027$) and represents even worse performance than in the blind control.

8.6 Discussion

Several general points are worth making before we consider the fate of specific hypotheses. Overall, evolved behaviour in the experimental condition was more like the behaviour of "blind" animals than that of animals able to perceive the fighting ability of their opponent. Being able to observe an opponent's intention movements does not seem to be as useful as knowing their fighting ability. The unfakeable control condition seems to have functioned as intended; it is clear that contestants in this condition were able to use the information they were given about their opponent's RHP in order to conserve energy by fighting less often. Finally, across all conditions the evolved animals did not consistently escalate but often chose to retreat. This behaviour is as expected given that the costs of being seriously injured were greater than the benefits of winning control of the resource.

The predictions derived from the work of Enquist (1985) and Hurd (1997b)—that cost-free signalling of fighting ability would evolve—were certainly not supported. The fact that mean energy expenditure was significantly greater in the experimental condition than in the unfakeable control is enough to establish this. If cost-free signalling had evolved, then results in the two conditions should have been the same. This negative conclusion holds for both the continuous and the discrete fighting ability cases. The most likely reason for the failure of Enquist's prediction is that the condition established for the stability of honest signalling, namely that $D > \frac{V}{2} + C$, has not been met. The condition requires that the injury cost for a weak animal of meeting a strong one must be greater than the cost of an escalated fight between two weak animals. However, given that most contests were over quickly, that weak animals were less aggressive than strong ones, and that some animals were observed to "change their minds" and flee from highly aggressive opponents (see Figure 8.2), it is probable that weak animals were often able to flee from stronger opponents before suffering any injury. On the other hand, a protracted contest with another weak animal would cost close to 200 units of energy, even for the winner. In nature also it seems likely that a weaker contestant, realizing it is outmatched, will be able to flee from a strong opponent without suffering more than minor injuries on average, whereas an escalated contest with an animal of similar fighting ability is likely to be drawn-out and dangerous. For a detailed argument that Enquist's condition is inherently implausible, see Caryl (1987).

The possibility remains that less-than-ideal signalling was occurring in the experimental condition. However, this looks unlikely. As we have seen, escalated fights in the experimental condition were no more likely to be between well-matched opponents than in the blind control. If a signalling system had evolved, even an unreliable or a costly one, it should have caused some out-classed competitors to retreat, and thus increased the proportion of closely-matched fights. Furthermore, the results from the fourth, arbitrary-channel variation suggest that the advantage accrued in the experimental condition relative to the blind control was due merely to being able to observe one's opponents rather than being able to communicate with them.

The predictions derived from Maynard Smith's (1982) hawk-dove game provide a much better fit to the data. Communication of RHP does not appear to have evolved, and there is some evidence for a mixed strategy equilibrium: Figure 8.8 shows that contestants in the CFA version were likely to begin the contest with *either* a highly aggressive move or a neutral one. As animals became stronger they were more likely to be aggressive, but in the CFA case it paid animals across the entire range of fighting ability to be somewhat unpredictable. In this sense the results support the standard game-theoretic position on signalling during contests: when deception is possible, far from sending honest signals of RHP, competing animals will be selected to reveal as little as possible about their status.

There is a minor paradox in the results, in that there is clearly some information to be had in observing the intention movements of one's opponent. Figure 8.9 in particular shows that, in the DFA case when animals were either strong or weak, animals in the two categories behaved in different ways. It should have been possible for an animal to determine the strength of its opponent with some confidence, simply by observing the opponent's characteristic first move. Figure 8.6 shows that, at least on average and at least in the DFA case, this was occurring, because weak animals were behaving differently depending on the strength of their opponent. However, this is not enough to establish that communication had evolved. Given that our definition of proper signalling specifies that both the signal and the response must have been selected for *qua* signal and response, we would need to show that the stereotypical behaviour of strong and weak animals had been selected for partly because of its signalling value. There is no evidence that this is the case—animals in the blind control condition, in which communication was impossible, behaved in very similar ways.

Exploitation does not qualify as proper signalling; this concept may help us to understand what is going on in the simulation. If we assume for a moment that the "poker-faced competitors" picture from game theory is correct, then it is likely that in the experimental condition the animals are trying to be as uninformative as possible. However, they have other constraints on their behaviour: taking a random walk up and down the Θ-continuum would be an excellent way of being uninformative, but it would leave one vulnerable to opponents that attacked as rapidly as possible. In the DFA case, a weak animal faces an especially difficult dilemma: half of the time its opponent will be strong, and this represents a fight that it cannot win, so it should flee. But the other half of the time its opponent will be weak, in which case the animal does not want to be the first to flee because simply waiting around for the competitor to depart will mean winning the resource. The balance between these two impulses means that the evolved strategy of weak animals is to make a neutral first move, most of the time. However, this neutral move has not been selected for because of its value as a signal of weakness. Any information that the opponent can glean from this move constitutes exploitation in the sense outlined in section 3.3, just as a particularly perceptive mantis shrimp that noticed signs of recent moulting in its opponent would be exploiting that opponent and not receiving a signal from it.

A poker analogy may be useful here: the way a player bets may give you information about the strength of their hand, especially if that player is less than perfectly rational. But it is not the function of betting behaviour to give away this information; the function of betting is of course to win money. Betting is not communicating.

We should also remember that even though animals are giving away information by making movements that are indicative of their strength, observing one's opponent's first move does not give definite knowledge as to the opponent's fighting ability. Even in the DFA case, *sometimes* a strong animal would make a neutral or even a retreating first move, and *sometimes* a weak animal would start with an aggressive move. Therefore some uncertainty about the opponent's RHP would always remain. The only way to probe an opponent and find out for sure how strong they were would have been to escalate at all times, but it was clearly not worthwhile for even the strongest animals to challenge every possible bluff—they would have been engaged in a prohibitively expensive number of escalated fights. Moreover, even in the DFA case, the information gained by observing strength-typical moves cannot have been particularly reliable or fitness would have risen to levels comparable with the unfakeable condition.

The hypotheses about sustainable bluffing discussed at the beginning of the chapter assume that a signalling system exists and that a level of deception can be maintained within it. Because no proper signalling system evolved, these predictions are of debatable relevance. Nevertheless, in terms of Gardner and Morris's (1989) model the simulation results suggest that the cost of bluffing was low. Recall that Gardner and Morris argued that, given the high cost of injury, a low cost of bluffing would lead to a non-signalling equilibrium in which all competitors bluff, while a high cost of bluffing would lead to honest signalling. We can say that bluffing was probably cheap for the same reasons that we dismissed Enquist's condition $D > \frac{V}{2} + C$: weak animals had time to retreat if a bluff did not work, and could therefore avoid serious injury costs.

Adams and Mesterton-Gibbons's (1995) prediction, that the strongest animals would signal their strength and the weakest would attempt to bluff, was not borne out. If it had been, we would have expected to see in Figure 8.8 that both the strongest and the very weakest animals started the contest with an aggressive first move, whereas animals of average ability were more neutral. Indeed, the data on bluffing strongly support the conventional game-theoretic view (Maynard Smith, 1982; Krebs & Dawkins, 1984) that participants in aggressive interactions will eventually come to pay little attention to each other's manipulative "signals". For example, in the CFA case (Figure 8.8), the weakest animals start with an aggressive move about 25% of the time. These bluffs at least occasionally result in the animal gaining the resource (see Figure 8.10), but to explain this we do not need to propose that the animal's opponent is paying any attention, because in the blind condition similar results are observed.

As no proper signalling evolved, it is difficult to assess the implications of the work for the idea that intention movements can serve as raw material for signal evolution—although the data from Figure 8.6, for example, certainly suggest that the animals had no trouble in evolving strategies that took their opponent's intention movements into account. However, the point is that in nature animals are obviously not given dedicated, artificial communication channels *deus ex machina*. Plausible simulation models must incorporate phenomena like intention movements in order to investigate what happens when evolution co-opts an existing behaviour for the new purpose of signalling.

Finally, the findings serve as a reminder that accounting for the use of conventional signals of RHP in aggressive interactions, if indeed such signals exist, stands as an open problem for biological modelling.

Chapter 9

Honest signalling and sexual selection

Are sexual advertisements proper signals? That is the question confronted in this chapter. To answer it in the affirmative, we have to be able to show that male advertisements have been selected for because they map to some underlying quality that is of interest to females.[1] It is not obvious that this is the case. Recalling the distinctions established in section 3.3, we can safely assume that when a female chooses one male from amongst several on the basis of an advertisement trait, this constitutes an influence interaction. However, it is possible that the successful male is manipulating the female, relying on a tendency to respond that has been selected for in some other context. It is also possible in principle that the female is exploiting the males, gaining information about their quality as potential mates through their "advertisement" traits which have in fact been selected for utility in some other domain. In neither of these situations could we refer to the advertisement trait as a proper signal.

If we postulate that a male advertisement *does* function as a signal, and thus that it has been selected to convey information about males to females, we come up against the basic problem of honesty yet again. Why should low-quality males ever honestly signal their condition, when by doing so they will make themselves unlikely to be chosen as mates? Why wouldn't all males produce the maximum advertisement, regardless of their true quality—all claiming, in effect, to be the most desirable. Zahavi's handicap principle, which has been discussed at some length in chapter 2, provides a possible mechanism by which a sexual signalling system could be kept honest.

The plausibility of the handicap principle has been demonstrated by many models, both mathematical and simulation, in recent years. However, most of these models have made a major simplifying assumption: namely, that the underlying male quality of interest to females is environmentally determined. This could mean, for example, that males are advertising their level of nutrition, or the quality of the territory they possess. Clearly, in the case of sexual signalling this simplifying assumption misses the interesting subset of cases in which males are believed to be informing females of their *genetic* quality. For example, when sage grouse *Centrocercus*

[1] Throughout this chapter, sexual advertisement traits will be assumed to be expressed only by males and evaluated only be females. Whilst this is the general rule in nature, there are of course some exceptions. No sex bias is intended by the adoption of this convention.

urophasianus mate the males contribute only their sperm, leaving all other aspects of the project of raising offspring to females. Nevertheless, the females choose their mate carefully on the basis of his ornaments, display behaviour, and central position in the mating arena (Wiley, 1973; Bradbury, Gibson, & Tsai, 1986). If the males are advertising anything in this case, it must be their inherited genetic quality.

The primary goal of this chapter is therefore to present a simulation model that investigates the evolutionary stability of the honest signalling of genetically determined male quality. However, in order to place such a model in its theoretical context, we will also look at the signalling of environmentally determined quality; that this latter kind of signalling can be evolutionarily stable seems to have been well established.

9.1 Signalling of environmentally determined quality

Models that have demonstrated the stable signalling of environmentally determined male quality include those of Grafen (1990a, 1990b), Hurd (1995) and Bullock (1997b). The logical structure of these models is similar; Grafen's (1990b) will be used as an example. Grafen postulates a population of male and female organisms with the simplest genetic system possible: a single haploid locus. When in a male body this locus specifies an advertisement strategy, and when in a female body it specifies a response strategy. The organisms have a four-stage life cycle.

1. The males are randomly assigned a quality level. Thus a male's quality has nothing to do with the quality of his parents (indeed, his mother does not have such a property) but is down to the luck of the draw.

2. The males then refer to their genetically specified advertisement strategy in order to determine the magnitude of their advertisement trait; we can assume that the advertisement equates to tail length or something of the sort. The advertisement strategy is a function of quality, but the function may be either increasing, decreasing, or flat. So, for example, one male may have an honest advertisement strategy, and produce a tail that is proportional in length to his quality level, while another's strategy may be to grow a tail of a particular length regardless of his underlying quality.

3. Male survival is assessed. A high quality level makes a male more likely to survive to breeding age, but at the same time a high-valued advertisement trait (i.e., a long tail) makes him less likely to do so. A key feature of the model is put into effect at this stage: while quality is good for survival and advertising is bad for survival, the unit costs of advertisement are lower for higher quality males. This is the condition referred to as Grafen's proviso in chapter 2, and is the real "engine" behind the handicap principle. In simple terms it means that extending your tail by one centimetre is cheaper if you are a high quality male.

4. Breeding takes place. Females cannot perceive male quality directly but they gain fitness benefits (extra offspring) if they mate with high-quality males. The females randomly encounter the surviving males and choose them to mate with if the male's advertisement trait is above the female's current aspiration level; the aspiration level is specified by the female's inherited response strategy. Once a female has mated she is out of the mating pool thereafter. There is therefore selective pressure on females not to be too choosy and end up unmated, and not to be too eager and mate with the first male they meet. After all, mating with the first male encountered means mating with a male whose expected quality is only

> average, but if males are using honest advertisement strategies, then it is worth waiting for a long-tailed suitor as he is likely to be of high quality.[2]

Grafen develops a general population-genetic model that captures analytically the way such a system would be expected to evolve. He demonstrates that there are two equilibria: the first is a non-signalling strategy in which none of the males advertise and the females can do no better than to choose their mates at random. The second is a handicap signalling strategy, in which males produce costly advertisements that are informative as to their quality level. At the signalling equilibrium females are prepared to bear a cost associated with their preference; i.e., accessing the information about male quality that is inherent in the advertisement trait is worth something to them. In this way Grafen demonstrates that honest signalling of environmentally determined male quality can be evolutionarily stable.

9.2 Signalling of genetically determined quality

If we suppose that males might be advertising their genetic quality to females then the situation becomes more difficult to model mathematically. It is not clear that an honest signalling equilibrium will exist in the same way as described above. As noted in section 2.7, a central problem is that on first consideration we would not expect there to be any residual variation in male quality and there would therefore be nothing to signal about. Suppose that males vary on a heritable trait we shall call viability: high viability males are more resistant to disease, as are their offspring, and the trait is therefore fitness-related. If the males were honestly advertising their viability level, and females were choosing to mate with high-valued males, then after a few generations the males will all have trait values clustered around the optimum. As Williams (1975) and Maynard Smith (1978) have argued, there should be no heritable variation remaining in fitness-related traits at equilibrium.

And yet we find female sage grouse paying the costs of choice (e.g., time costs and predation risk) in order to choose the best male, when the male will contribute only his genes. This is known as the paradox of the lek: why aren't modern sage grouse males all maximally viable, and thus equally attractive to females? The most likely answer is that mutation on fitness-related traits (e.g., our viability example) is negatively biased. That is, a single mutation event affecting the genes controlling a fitness-related trait is more likely than not to decrease the value of that trait. There is some empirical evidence for this idea: Partridge (1980) found that at least one component of fitness was mildly heritable in an experiment with *Drosophila melanogaster*; Partridge reasoned that this could only occur if mutational load kept fitness-related traits below their optimum value. Pomiankowski and Møller (1995), reviewing evidence in many avian species, reached a similar conclusion. Whereas mutation is usually thought of as being equally likely to move a one-dimensional trait in either direction, it seems reasonable to assume that in reality it is more likely to move certain traits downward.

Iwasa et al. (1991) constructed a population-genetic model of the evolution of costly male advertisements and female preferences; they incorporated just such a negative mutation bias on the viability trait. Iwasa et al.'s model purports to show that honest signalling of genetically

[2] Grafen included in his model the idea that females are mating over the course of a breeding season and would prefer to mate in the middle of the season when conditions are optimal. Female aspiration levels were therefore functions that tended to go down over the course of the season. We need not be concerned with this added complexity.

determined male quality can be evolutionarily stable. It is one of the very few models to deal with genetically determined quality and will provide a starting point for the simulation described in this chapter.

Iwasa et al. ask us to consider a sexual population of organisms with genes coding for three traits: male advertisement, female preference, and a general viability trait. The expression of the first two traits is sex-limited, i.e., females carry advertisement genes and males carry preference genes but they do not express them. The viability trait is expressed by both sexes: a higher value on this trait means that an animal is more likely to survive to breeding age.[3] As in Grafen's (1990b) model, high values of the male advertisement trait are detrimental to male survival. The female preference trait can be either positive or negative, indicating the strength of a preference for males with larger-than-average or smaller-than-average ornaments respectively. (In fact Iwasa et al.'s analysis concentrates on situations where the females prefer larger ornaments—we can do the same.) Females pay a fitness cost that increases exponentially as the absolute value of their preference trait increases; a zero preference value indicates random mating, which incurs no costs. Male mating success increases exponentially with the magnitude of their phenotypic advertisement, with the rate being proportional to the population mean female preference.

The model is an additive quantitative genetic one: it is assumed that there are a number of loci contributing in an *additive* fashion to the overall value for a particular trait, and that therefore each trait can be safely modelled as a real number. The effect of sexual reproduction is that an offspring's value for any trait will be the mean of the two parental values, and the effect of mutation is to perturb the value for a particular trait. Additive quantitative genetic modelling is a common convention in population genetics.

Iwasa et al. assume a population with discrete, non-overlapping generations. They then construct an equation to describe the rate of change in the population mean values of the male advertisement (t), female preference (p), and viability (v) traits. It is reproduced here.

$$\begin{pmatrix} \Delta\bar{t} \\ \Delta\bar{p} \\ \Delta\bar{v} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} G_t & B_{tp} & B_{tv} \\ B_{tp} & G_p & B_{pv} \\ B_{tv} & B_{pv} & G_v \end{pmatrix} \times \begin{pmatrix} \partial \ln W_m / \partial t \\ \partial \ln W_f / \partial p \\ \partial \{\ln W_m + \ln W_f\} / \partial v \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -w \end{pmatrix}$$

The first matrix term on the right-hand side is the additive genetic variance-covariance matrix; the terms on the main diagonal refer to the genetic variance in each trait, while the other terms describe the co-variance (i.e., degree of linkage) between traits. The second term is the selection vector, which specifies the effect that small changes in t, p and v would have on an individual's fitness; W_m and W_f refer to male and female fitness respectively. The final matrix term is the implementation of the negative mutation bias on viability: note that whatever else happens to the mean value of v in each generation it will decrease by an amount w, whereas there is no bias on t or p.

Iwasa et al. are interested in whether there is an equilibrium at which females show costly preferences for extreme values of the male advertisement trait. They reason that if females are

[3] Up to a point: Iwasa et al. actually set an optimal value for the viability trait; animals with viability scores higher than this arbitrary optimum would be less likely to survive. However, the negative mutation bias on viability means that the optimum is not reached in practice.

prepared to pay a cost for their preference, then there must be information worth having in the expressed values of the advertisement trait, and it is therefore an honest indicator of quality. For the population to be at an equilibrium would mean that the values of $\bar{t}$, $\overline{p}$ and $\overline{v}$ were no longer changing; the authors therefore set the three values in the matrix on the left-hand side of the equation equal to zero, and re-arranged terms to get an expression for the mean female preference value at equilibrium. They asked what it would take for this value to be positive, i.e., for there to be an equilibrium at which females preferentially mated with long-tailed males and were prepared to pay costs in order to do so. After making certain assumptions which we will consider below, Iwasa et al. came up with two conditions for the existence of such an equilibrium.

The first condition was that w had to be positive—there needed to be negatively-biased mutation on the viability trait. Otherwise, values of v in the population would be clustered around the optimum, and the females would then be in a position where random mating was just as likely to result in a high-viability partner as was a costly preference.

The second condition was that $\rho_{pv} > \rho_{tp}\rho_{tv}$. That is, the genetic correlation between preference and viability had to be greater than the product of the correlations between advertisement and preference and between advertisement and viability. Another way of putting this is that there must be a link between preference and viability that does not come about solely because of their joint relationship with the male advertisement trait: if ρ_{ab} and ρ_{bc} are both positive, then ρ_{ac} will be equal to their product simply because a is linked to c through b; ρ_{ac} will be greater than the product of ρ_{ab} and ρ_{bc} only if there is some additional causal link between a and c.

Recalling for a moment the three variations on the handicap principle spelt out in section 2.4.2, we can see how this second equilibrium condition implies that whereas the conditional and revealing handicaps will work, the pure epistasis handicap will not. In all cases, genetic linkage between the preference and viability traits (i.e., high ρ_{pv}) will come about if high-preference females tend to mate with high-viability males. Females cannot, of course, perceive viability directly, but they can perceive the phenotypic value of the male advertisement trait. If there is a correlation between a male's viability and his expressed advertisement trait, then it is possible that a genetic correlation will develop between p and v. In the conditional and revealing handicaps, the viability trait v directly affects the expression of the male's advertisement—general viability modifies the expression of the genes for growing an ornament of a particular size, or, in the case of the revealing handicap, low viability means that a large ornament cannot be successfully maintained as an adult. Valuable information for females concerning male viability has thus been built into the expressed male trait. It is therefore possible for a direct correlation between female preference and viability to evolve. However, in the pure epistasis handicap, the realized size of the male advertisement is only linked to viability indirectly. The size of the male ornament depends on t, the underlying genetic value that codes for it. There is in turn a correlation between t and v that comes about through epistatic fitness interactions, i.e., through the fact that high-t, low-v males tend to die before reproducing. Thus, any correlation between p and v comes about only because of their joint linkage to t, and ρ_{pv} will be equal to the product of ρ_{tp} and ρ_{tv} but not greater than it.

One difference between Grafen's (1990b) model and that of Iwasa et al. (1991) is that the former assumes the expressed male advertisement trait is related to underlying quality because of a genetically specified strategy, whereas in the latter model male advertisement is a straightforward heritable trait. Of course, in Grafen's model male quality is randomly determined and it would therefore make no sense for the male advertisement to be a simple genetic trait: there needs to be some way in which the expressed advertisement can at least potentially serve as an indicator of quality. In Iwasa et al.'s model, on the other hand, there is no a priori reason why the size of the male advertisement could not have been due to an inherited function of quality rather than being the direct expression of a real-valued gene; in the simulation described below both possibilities will be investigated.

On first examination, it seems that Grafen's proviso—the stipulation that the unit costs of advertisement must be lower for higher-quality males—has not been explicitly implemented in Iwasa et al.'s (1991) model. We are told that the male fitness function, W_m, depends on the values of t, v, $\bar{t}$, $\overline{p}$ and $\overline{v}$, but Iwasa et al. have treated this function in a very general fashion, specifying only that male fitness should decrease as t and v deviate from their optimal values. This seems to imply that the cost for males of the advertisement trait could be independent of their viability. However, later on in their paper Iwasa et al. establish that Grafen's proviso is indeed a pre-condition for the evolutionary stability of honest-advertisement equilibria. They argue that the conditional handicap can lead to an equilibrium with honest signalling and costly female preference, because $\partial s/\partial v > 0$ (s is the size of the advertisement trait that is actually expressed, as opposed to the genetic specification, and v is viability). The females therefore gain information about viability from male ornament size. However, it turns out that for this equilibrium to exist, the second derivative of the cost function for s with respect to v must be negative; this is in addition to s depending on both t and v. In other words, the cost per unit of s must get lower as v increases. This is simply another way of stating Grafen's proviso.

Iwasa et al.'s (1991) result is potentially a very general and powerful one. However, the authors have had to make some assumptions in order to render the mathematics tractable. They have been forced to neglect the higher-order terms in the Taylor expansions of the various fitness functions, although this probably does not upset the validity of the model (see Gomulkiewicz, 1998). More seriously, they assume that the genetic co-variances between male advertisement, female preference, and the viability trait will all be positive and constant. At first glance, this appears to be begging the question: one hallmark of an honest signalling equilibrium would be a correlation between the viability trait, which female observers cannot detect, and the advertisement trait, which they can; assuming that such a correlation exists looks like cheating. However, the logic of Iwasa et al.'s work asks: *if* these conditions hold, is there a signalling equilibrium? It might have turned out, for example, that despite such generous assumptions, no plausible conditions for the existence of a costly-preference equilibrium could be found.

Still, this critical assumption must detract from the generality of the work. As Andersson (1994) has pointed out in his review of the sexual selection literature, Iwasa et al.'s (1991) model is important because it attempts, as few others have done, to deal with inherited male quality, but it is not clear whether the conclusions would hold without assuming constant positive co-variances between t, p and v. In the real world, genetic co-variances are of course not constant but change as the population evolves over time; even if we suppose that the co-variances might start out positive, it is not clear that they would remain so. And it is more likely that the co-variances in a plausible initial population would be close to zero. The problem highlights a weakness of the population-genetic approach: despite the name, there is in fact no population, which means that such important variables as genetic variances and co-variances must be input into the model as parameters, rather than being measurements that are made with respect to an evolving lineage.

On the other hand, the mathematics become intractable if the co-variance assumptions are not made. This presents an excellent opportunity for an evolutionary simulation. In the work described in this chapter, we will try to find out what should be expected if the co-variances were not held artificially constant but allowed to vary in the natural way.

9.3 Zahavi vs. Fisher: Measuring sexual signalling

How can we tell, in an evolutionary simulation or in the real world, when males are producing costly advertisements that serve to provide information about their underlying viability? In other words, how can we tell when proper sexual signalling is occurring?

Werner (1996), describing an evolutionary simulation model, suggests that a general criterion for detecting the effects of sexual *selection* is to note the ratio at which males and females survive to reproductive age. If an equal proportion of males and females survive, then this is evidence that the males are not producing any ornaments that are particularly costly to their survival. However, if substantially fewer males survive then they must be producing costly advertisements. Of course, we need to assume that the higher male mortality rate is not due to inter-male combat or some other factor, but this assumption can be safely made in a simple three-trait model. The males *could* have evolved such that they produced no ornaments at all, and therefore survived to adulthood just as often as females. We must invoke sexual selection in order to explain the fact that they are bearing significant advertisement costs. However, this manifestation of sexual selection might not involve any honest indicators of underlying viability. The Fisherian runaway process could be at work, and high male mortality might be the result of genetic linkage between the male ornament trait and the female preference (see section 2.7). A female-biased survival ratio is certainly evidence that the males are bearing costs, but it does not prove that sexual signalling is occurring.

Similarly, directly observing the production of costly traits by males establishes that sexual selection is at work, but no more than that. Costly male displays are a necessary but not sufficient condition for handicap signalling of underlying quality; such displays are predicted by both Fisher and Zahavi.

Iwasa et al. (1991) suggest that the best way to measure the existence of a handicap signalling equilibrium is to determine whether females are prepared to pay a cost for their preferences regarding the male advertisement trait. If females bear the costs of preference, then it must be worthwhile for them to select particular males; females with a preference must be fitter than females that mate randomly. Fitness benefits for female preference could come about if preferring a particular class of male ornaments resulted in mating with higher-quality males. This, in turn, would occur if there was a correlation between the visible male trait and underlying male quality, i.e., if the advertisement trait bore information about viability. The existence of costly female preference is thus a reasonable pointer to the existence of a signalling system. However, the same authors (Pomiankowski, Iwasa, & Nee, 1991) have elsewhere established that the Fisher process can also lead to costly female preference under certain circumstances (although models of the Fisher process have more often assumed that female preference was not costly). If runaway sexual selection can sometimes lead to costly female preferences then this measure also fails to qualify as a sufficient condition for the existence of a sexual signalling system.

Finally, Grafen (1990b) has suggested the "Fisher index" as a way of determining the extent to which costly male advertisements are the result of runaway selection on one hand and handicap signalling on the other. Grafen argues that in the former case male ornaments are exaggerated but only because of a self-reinforcing link with female preference. Variation in male ornament size has nothing to do with male viability; it must represent natural genetic variation in the advertisement trait, or, in the case of a strategic model, variation in the heritable strategies for mapping quality to advertisement. In the case of handicap signalling, in contrast, variation in the expressed male trait values will be at least partially due to variation in the underlying viability level, assuming that a signalling equilibrium has been reached. Grafen therefore suggests that the proportion of variance in the male advertisement trait that can be explained by underlying differences in quality or viability is a good measure of the extent to which honest signalling is occurring; his proposed Fisher index is actually equal to one minus this value, and represents the degree to which the situation can be explained in terms of runaway selection. In the simulation presented here we will assess this by looking at the correlation between the expressed male advertisement trait and the underlying male quality at the time mating takes place—note that only those males that survive to breed can contribute to this statistic. This measure will be considered along with those discussed above, in order to classify the results from different simulation runs. If the male-female survival ratio is less than one, if males are exhibiting costly ornaments, if females are paying the costs of preference, *and* there is information in the size of expressed advertisement traits concerning underlying quality, then it seems safe to say that proper signalling is occurring.

9.4 Description of the model

The work reported here is an attempt to translate the population-genetic model of Iwasa et al. (1991) into an individual-based evolutionary simulation. However, in addition to providing a test-bed for the possibility of honest signalling of genetically determined male quality, variations on the simulation are intended to exhibit (for comparison within the same general framework) the signalling of environmentally determined quality, and Fisherian runaway sexual selection.

The population consists of sexual individuals breeding in discrete, non-overlapping generations. Individual organisms have both a genotype and a phenotype; the genotype consists of real-valued genetic parameters.[4] In the standard conditions each organism carries a gene for the male advertisement or ornament trait (t_{gen}), the female preference trait (p_{gen}), and the general viability trait (v_{gen}). An individual's phenotype is derived from its genotype but is not necessarily identical to it: firstly, the phenotype is of course sex-limited, in that only males express the advertisement trait, and only females express preference. Secondly, the phenotype is derived in different ways in the different conditions investigated. Nevertheless, across all conditions, the phenotype always

[4] Simulations were also performed with a more realistic implementation of the genotype: each organism had a long binary chromosome, and the value for a particular trait was the sum of the bit-values along a section of the chromosome. This method allows the use of the standard genetic-algorithm operators of crossover and mutation. It also means that an organism's value on a trait will only be equal to the mean of the parental values on average; there is room for variation such as resembling one's mother more than one's father. However, the binary-genotype implementation was relatively expensive in computational terms, and pilot runs indicated that the results did not differ significantly from the method presented here.

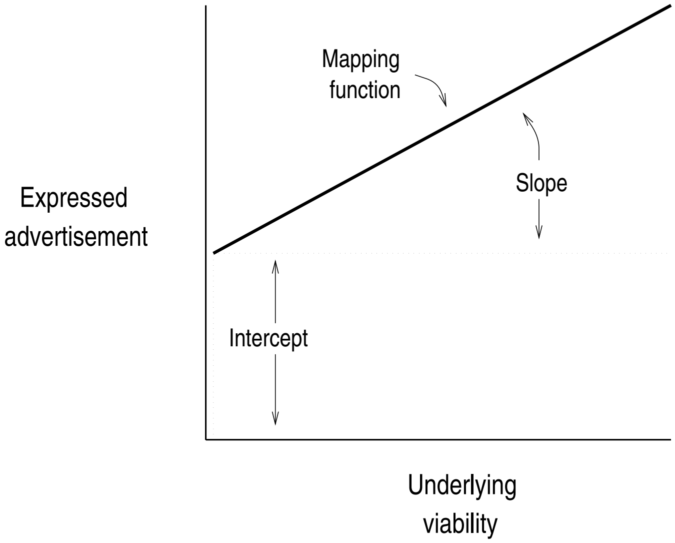

Figure 9.1: Male advertisement as a linear function of viability. The two genetic parameters are the y-intercept, which can vary between -1 and 1, and the slope of the line, which is expressed as an angle between $-\frac{\pi}{2}$ and $+\frac{\pi}{2}$ radians. The scheme makes positively-correlated, negatively-correlated and flat advertising functions possible. If the specified advertisement value would be greater than 1 or less than 0, it is truncated accordingly. This makes discontinuous advertising strategies possible: e.g., not advertising at all for $v < 0.5$, but producing an advertisement that is positively correlated with viability when $v \geq 0.5$.

consists of two real values: either t_{phen} or p_{phen}, depending on sex, and v_{phen}. Genotypic and phenotypic parameters are always real numbers between zero and one inclusive. The organisms go through a life cycle similar to the one in Grafen's (1990b) model.

Development stage

Each individual's sex is chosen at random, and its phenotypic trait values are determined. Normally, each trait is read off the genome, then a random gaussian error term is added ($\mu = 0$, $\sigma = 0.005$), and the resulting value stands as the expressed trait. However, several variations are possible. To investigate environmentally determined quality, the genetically specified value v_{gen} is ignored, and phenotypic viability is instead determined according to a uniform random distribution. Similarly, when investigating runaway sexual selection, v_{phen} is again disregarded and all individuals are given a phenotypic viability of 1—this ensures that there is no variation in viability and thus nothing for males to honestly advertise.

The phenotypic male advertisement is normally read off the genotypic value of t_{gen}. However, in the conditional and revealing handicaps the male's viability also influences the expressed ornament size—the phenotypic viability is therefore calculated first. For the conditional handicap, the advertisement size that would otherwise be expressed is reduced by an amount proportional to v_{phen}. In other words, $t_{phen} = t_{gen} \times v_{phen}$. Only males with the maximum possible viability actually produce an advertisement that is as big as their genotype specifies.

Sometimes male advertisement is not treated as a simple trait but as an inherited strategy relating ornament size to underlying viability. This is obviously necessary in the case of environ-

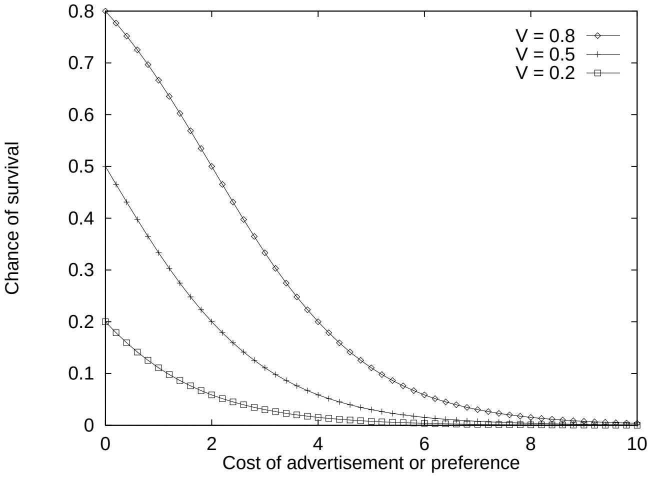

Figure 9.2: Function for determining the survival cost of the male advertisement and female preference traits. We assume that an individual produces an advertisement or preference of medium size, i.e., that t_{phen} or $p_{phen} = 0.5$. The figure then shows the probability of survival for individuals of three different viability levels, as the cost of advertising (C_{adv}) or the cost of preference (C_{pref}) increases.

mentally determined (i.e., random) viability, but it is also investigated for genetically determined viability. In these cases, instead of a gene t_{gen} for advertisement size, organisms carry two genes: one for the intercept and one for the slope of a linear function relating viability to size of advertisement. These two values range between zero and one like all other parameters, but during the development stage they are scaled to the range ± 1 for the intercept, and $\pm\frac{\pi}{2}$ for the slope, which represents not a gradient but an angle. Figure 9.1 shows how these two scaled parameters define an advertising strategy; the method is based on the implementation of male advertisement strategies described by Bullock (1997b).

Survival stage

The question of which individuals survive to adult reproductive age, and which ones die young, is then settled. An individual's basic probability of survival is simply equal to its phenotypic viability: less viable animals are less likely to survive. However, both male advertisements and female preferences are supposed to be costly, and the cost of these characteristics is manifested as a reduction in an individual's probability of survival, according to the degree of the trait's phenotypic expression.

Grafen's proviso, in which the unit costs of advertisement are lower for higher-quality signallers, is enforced at this stage. The basic probability of survival (v_{phen}) is first converted to an odds ratio, and then scaled by $(1 - t_{phen})^{C_{adv}}$, where C_{adv} represents the cost of advertising. If $C_{adv} = 0$ then there is no cost at all to males for growing ornaments; if $t_{phen} = 0$ then a male will pay no costs regardless of how high C_{adv} might be. The scaled odds ratio is then converted back to a probability value. The result of all this manipulation is the following expression for the probability of survival:

$$p_{survival} = \frac{v_{phen}(1-t_{phen})^{C_{adv}}}{v_{phen}(1-t_{phen})^{C_{adv}}-v_{phen}+1}$$

The scaling factor implements Grafen's proviso, because individuals with high phenotypic viability will be best able to bear the costs of advertisement. Figure 9.2 illustrates the operation of the function.

The survival costs of female preference are assessed in exactly the same way as the costs of the male trait: p_{phen} is simply substituted for t_{phen}, and C_{pref} for C_{adv}, in the expressions above. Theories of handicap signalling generally do not require that female preference should involve anything other than a simple cost that is independent of viability; however, calculating female costs in the same way as male ones allows the costs borne by each sex to be directly compared.

Once the probability of an individual's survival has been calculated, the brute fact as to whether it survives or not is determined using a pseudo-random number generator. In the rare event that no males (or no females) survive to adulthood, it is necessary to randomly select one male (or one female) to be resurrected; otherwise the population would suffer extinction.

Mating stage

Once the individual phenotypes have been fleshed out and the issue of survival to adulthood determined, the surviving males and females are able to breed. This is the point at which females get to exercise their preferences, and males may experience the benefits of their costly ornaments.

There are many ways in which female preference and choice could have been implemented, and indeed several different methods were experimented with in pilot studies. The method chosen has the virtue of simplicity: a surviving female is randomly chosen, and she is then presented with a "lek" of eight males, also selected at random. With a probability equal to her preference value (p_{phen}), she selects the male with the largest expressed advertisement trait to mate with. If she does not choose this male, she chooses randomly from among the eight males. The effects of this procedure are in keeping with the way preference is described by Iwasa et al.: high-preference females are likely to end up mating with the male with the biggest ornament, while zero-preference females will mate with anyone. The results of Werner's (1996) simulation suggests that this method can be effective in producing sexual-selection effects, and that eight is a reasonable lek size. Note that the way female preference manifests itself is in contrast to Iwasa et al.'s model, in which preferences were expressed relative to the male population mean advertisement. Having a model in which individuals really exist allows us to avoid the rather dubious assumption that females could know what the population mean advertisement was; instead, females choose a mate from among those males they happen to come into contact with.

When male viability is environmentally determined, the females presumably need some incentive for choosing high-viability males. In the case of inherited viability, there is the obvious benefit of passing on good genes to one's offspring, but when male viability has been randomly determined there is no reason why a female should prefer a male with $v_{phen} = 0.95$, who survived easily, over a male with $v_{phen} = 0.05$, who is lucky to be alive. Therefore a viability mating bonus was devised for this condition: when a male and female copulate and produce one offspring, there is a probability equal to the male's v_{phen} that they will immediately produce a second offspring. This gives females a reason to be interested in high-viability males.

Some artificial manipulation proved necessary with respect to the female preference scores: phenotypic preference values of less than 0.1 are set equal to zero in practice. That is to say, females with sufficiently low preference values mate randomly. This is to avoid a situation in which there is selection pressure for zero preference in females (i.e., selection favours random mating) but the mean value of p never quite reaches zero due to recurrent mutation. This would lead in turn to a small female preference being manifested, which might well be enough to push males towards advertising when they otherwise would not have invested in ornaments. In other words, if this adjustment is not made then we risk artificially preventing the organisms from reaching a *non*-signalling equilibrium, by never allowing the females to be truly random in their mate selection.

The mate selection process continues as described until sufficient offspring have been produced to stock the next generation. Crossover and mutation are extremely simple: newborn individuals inherit the mean of their two parents' values for each real-valued genetic parameter. The mutation operator consists of adding a random gaussian ($\mu = 0$, $\sigma = 0.03$) to each genetic parameter. The all-important negative mutation bias on viability is implemented by subtracting 0.003 from whatever value a newborn individual's genetic viability would otherwise have been. If the mutated value of any trait would be less than zero or greater than one, it is truncated accordingly.

9.5 Results

The population consisted of 100 individuals, and evolution proceeded in each run for 5000 generations. Unless otherwise stated, the results summarize a window period over the last 500 generations, and are averaged across 10 repeated runs in each case. The repeated runs in the various conditions were each performed with a different seed for the pseudo-random number generator.

The simulations have been conducted over a range of values for the advertising and preference costs C_{adv} and C_{pref}. Werner's (1996) work suggests that males will be prepared to bear much higher costs in advertising than females will tolerate in expressing a preference, and the range of cost levels investigated reflects this. It is not that we have strong hypotheses about the sorts of behaviour to expect when the advertising and preference costs are at particular values, but simply that we need to investigate a reasonable range of costs in order to find out whether the phenomena we are interested in—such as non-zero preference in females, and informative advertising by males—are going to evolve at all. If we ran the simulation using only one (albeit plausible) cost level for each trait, then we may well conclude that a certain behaviour does not occur, when in fact it would have evolved in a nearby region of "cost space". In Iwasa et al.'s (1991) analytic model, costs are dealt with as unspecified functions, or left as variables in algebraic expressions, and this allows very general conclusions to be drawn. In contrast, in a simulation such as the one presented here, costs must be set to particular values for any one run. The only way that we can approach the generality of mathematical methods is to observe what happens as critical variables such as cost are allowed to vary across runs.

The investigation of different initial conditions would have been a valuable extension of the simulation, but regrettably time and space constraints mean that only random initial conditions have been employed. Looking at what happens when the population starts with genes set to an honest advertisement strategy, or to a non-signalling strategy in which males do not advertise and females mate randomly, must wait for future work. In all conditions, all genetic parameters for the individuals in the first generation were set to uniformly distributed random values between zero and one inclusive. The one case where an alternative genetic starting point has been briefly

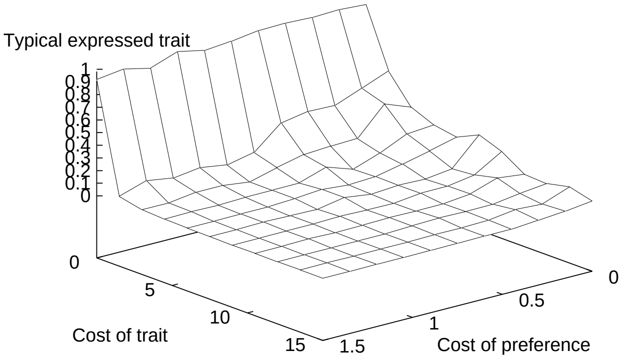

Figure 9.3: Typical expressed advertisement trait values in the environmentally determined quality condition, by C_{adv} and C_{pref}. The typical trait value is calculated by substituting 0.5, the mean random viability value, into the genetically specified advertisement strategy of each member of the population.

investigated is noted.

9.5.1 Environmentally determined quality

In this condition, viability was determined randomly, and the expression of the male advertisement trait was determined as a heritable function of viability according to the scheme shown in Figure 9.1. Females were motivated to mate with high viability males because there was a chance equal to the male's v_{phen} that the pair would have two offspring rather than one. This situation is very similar to those modelled by Grafen (1990b) and Bullock (1997b). Generalizing from these earlier models, we would expect sexual signalling to evolve in this case.

Figure 9.3 shows the average male advertisement values. We can see that when $C_{adv} = 0$ and advertising was not costly, most males grew ornaments of the maximum size. (This makes sense: if large ornaments are free, then a male might as well have one in order to improve his chances of being the most attractive on the lek.) As the costs of advertising and the costs of preference went up, males tended to produce no ornaments at all. However, there is a range of values for which male advertisement and female preference are not too expensive, in which males produce modest ornaments. Figure 9.4 shows the mean female preference values: these are high when preference is cost-free, but fall off as the cost of preference goes up.

So, we have established that when the cost values are right, males will produce ornaments, and females will exhibit preferences. Does this constitute sexual selection, or, more rigourously, does it constitute sexual signalling? As noted in section 9.3, the male-female survival ratio was proposed by Werner (1996) as a way of measuring whether female choice was shaping the selective landscape for males, i.e., as a way of showing whether or not sexual selection was occurring. If we discount the unrealistic cases when either advertising or preference are completely cost-free, then the most extreme male-female survival ratio observed was 0.552. This occurs when $C_{adv} = 3.0$ and $C_{pref} = 0.15$. At this point the typical expressed advertisement was 0.365, the

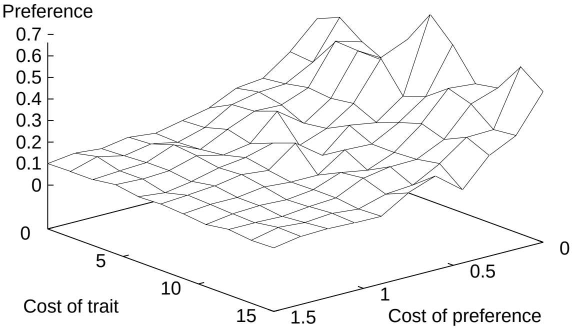

Figure 9.4: Mean preference values in the environmentally determined quality condition, by C_{adv} and C_{pref}.

mean female preference was 0.456, and the correlation between the expressed advertisements of males and their underlying viability was 0.389. In lieu of statistical tests to establish that these values are significantly different from zero, we can look at what happens when the cost levels are very different. For example, when advertisement and preferences were very costly ($C_{adv} = 15$ and $C_{pref} = 1.5$), the male-female survival ratio was 1.084 (i.e. more males survived than females), the typical advertisement was a tiny 0.001, the mean preference was 0.091 (which is less than 0.1 and would therefore result in random mating; see section 9.4), and the correlation between expressed advertisement and viability was only 0.023.

Thus we can conclude that, at least in one region of the cost landscape, males are producing signals that are costly enough to affect their survival, females are bearing a cost for preferring ornamented males, and male ornaments are providing information about underlying viability. Considered together, these characteristics mean that proper signalling is occurring. It is critical to this conclusion that there is a correlation between a male's viability and his expressed advertisement—Figure 9.5 shows this correlation over the full range of cost values, and we can see that there is informational value in the male advertisement trait over much of the landscape. Males are being informative about their viability even though, in the case of low-viability individuals, this will mean they are less likely to be chosen for mating opportunities. The results support the generally accepted thesis that Grafen's proviso can lead to the honest signalling of (environmentally determined) male quality despite the inherent conflict of interests between males and females.

Note also that the scales for the advertisement-cost and preference-cost axes in Figures 9.3, 9.4 and 9.5 are very different; as Werner (1996) found, males are prepared to bear substantially higher costs in order to advertise than females are prepared to bear in order to exhibit a preference.

9.5.2 Genetically determined quality

We now turn to the main purpose of the simulation model, and consider the possibility of honest advertisement of heritable viability. Each of the variations on the handicap principle discussed by Iwasa et al. (1991)—the pure epistasis, the conditional, and the revealing handicaps—will be

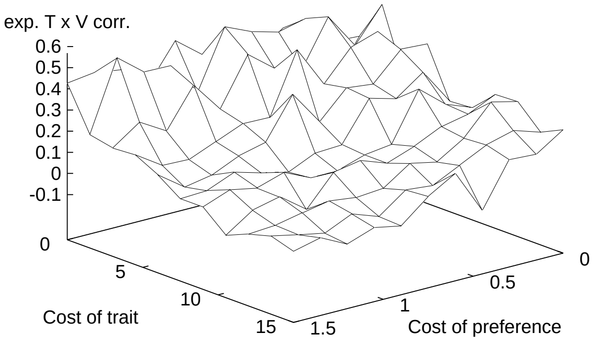

Figure 9.5: Correlation between expressed advertisement and underlying viability in the environmentally determined quality condition, by C_{adv} and C_{pref}.

examined in turn. We will also consider the strategic signalling of inherited quality.

A control condition of sorts was devised for these experimental runs, in which females paid the survival costs for their genetically specified preference value, but in practice it was ignored and consistently random mating was enforced. In other respects the control condition was identical to the pure epistasis handicap condition described below. If females are prevented from realizing any preference for male ornaments, then we would expect that the values of p_{gen} and t_{gen} will go to zero. There is no point in costly male advertisements that cannot influence female choice. The results for the control condition were approximately as expected. Discounting cases where C_{adv} or $C_{pref} = 0$, the average male-female survival ratio was 0.928, the mean advertisement trait was 0.054, the mean female preference was 0.137, and the average correlation between male advertisement and viability was 0.024. The fact that these values are close to zero (or close to one in the case of the male-female survival ratio) should help to increase our confidence in the validity of the simulation. However, the fact that they are not actually equal to zero means they can be regarded as baseline values: results in an experimental condition should exceed these levels before being taken seriously. In particular, it is worth noting that the mean female preference value can reach as high as 0.137 even when it is futile for females to attempt to exhibit a preference. This tells us that the rather low values of C_{pref} (in comparison to the high costs of advertisement) are only causing modest selection pressure with respect to the preference genes. The manipulation of preference described in section 9.4, in which values of less than 0.1 are treated as zero in practice, may also be having an effect.

Pure epistasis handicap

In this condition viability is inherited with a negative mutation bias. The term “pure epistasis” refers to the fact that there is no direct connection between the phenotypic expression of a male’s advertisement trait and his underlying viability. Any relationship between the two must be based on epistatic fitness interactions, i.e., on the fact that low-viability males will almost certainly die if they produce a large advertisement. Iwasa et al. (1991) found that the pure epistasis handicap was not evolutionarily stable; based on their model, we would expect no sexual signalling to evolve in

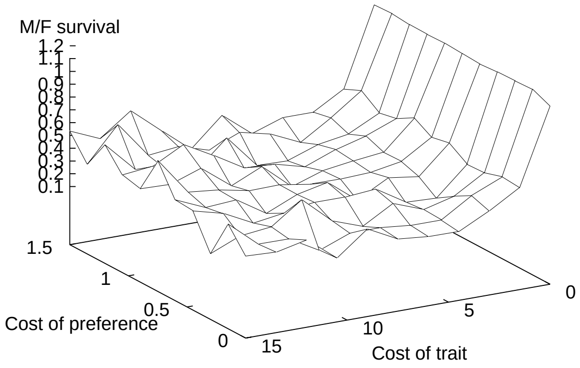

Figure 9.6: Male-female survival ratio in the pure epistasis handicap condition, by C_{adv} and C_{pref}. Note that the graph has been rotated for clarity.

this condition.

Figure 9.6 shows the rate at which males survived to reproductive age, compared to female survival. We note that as soon as the cost of advertising is greater than zero, many males are dying young because of their advertisement traits. Mean advertisement trait values were very high when $C_{adv} = 0$, falling off gradually to around 0.35 as the cost of advertising increased. Mean preference values were more uneven, with an overall average of 0.327. Discounting cases where C_{adv} or $C_{pref} = 0$, the most extreme male-female survival ratio was 0.132, when $C_{adv} = 10.5$ and $C_{pref} = 0.15$. At this point the mean value for the advertisement gene was 0.684 and the mean value for the preference gene was 0.400. The correlation between expressed male advertisement and viability was only 0.020, which suggests that despite costly advertisement and preference values, the size of male ornaments does not carry much information about underlying male quality. However, Figure 9.7 shows the value of this correlation across the cost landscape. It is clear that when the cost of advertising is greater than zero but less than about 5, male ornament size is modestly correlated with viability, and therefore *does* carry information. As in the environmentally determined quality condition, signalling of viability does not occur across the full range of cost values, but it certainly appears to be occurring in one region. This contradicts Iwasa et al.'s (1991) claim that the honest advertisement of viability cannot be evolutionarily stable under the terms of the pure epistasis handicap.

Conditional handicap

In the conditional handicap, viability is still inherited with a negative mutation bias. The difference between this condition and the pure epistasis handicap is that the expressed value of the male advertisement trait depends not only on t_{gen} but on v_{phen} as well. That is, the expression of the advertisement is condition-dependent. Iwasa et al. found that this version of the handicap principle could lead to evolutionarily stable costly-preference equilibria.

The mean male-female survival ratios are shown in Figure 9.8. As with the pure epistasis handicap, we see that as soon as the cost of advertisement becomes non-zero, males are dying much more often than females, because of their costly ornaments. Mean values for the advertise-

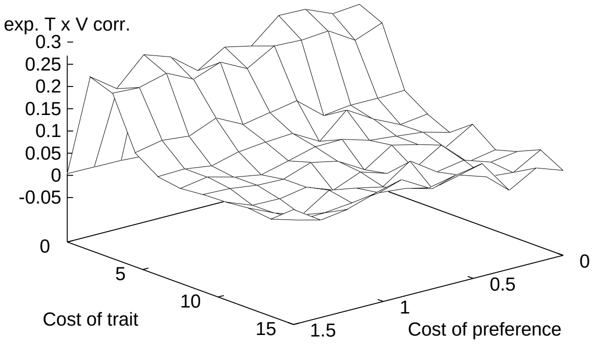

Figure 9.7: Correlation between expressed advertisement and underlying viability in the pure epistasis handicap condition, by C_{adv} and C_{pref}.

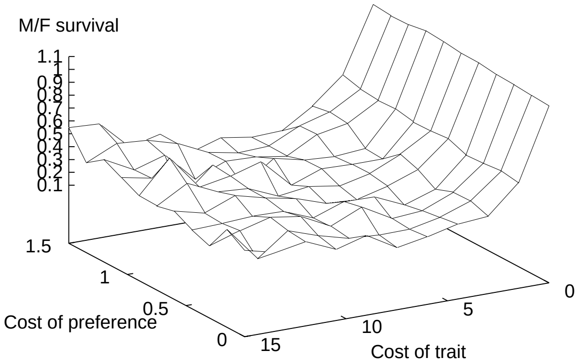

Figure 9.8: Male-female survival ratio in the conditional handicap condition, by C_{adv} and C_{pref}.

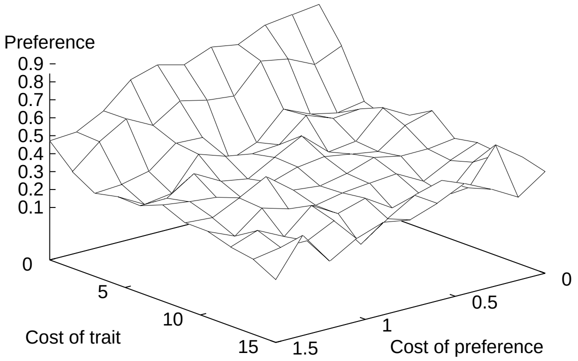

Figure 9.9: Mean preference values in the conditional handicap condition, by C_{adv} and C_{pref}.

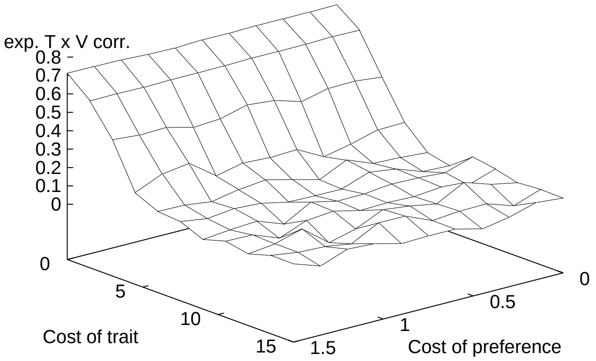

Figure 9.10: Correlation between expressed advertisement and underlying viability in the conditional handicap condition, by C_{adv} and C_{pref}.

ment trait were high when advertising was cheap and gradually became lower as C_{adv} increased, falling off to around 0.4 when $C_{adv} = 15$. This was also similar to the pattern found for the pure epistasis handicap. Figure 9.9 shows the mean values for female preference. It is interesting to observe that, although the mean value falls off somewhat as the cost of preference increases, high preference values are most likely to evolve when the cost of *advertising* is low.

Figure 9.10 shows the correlation between expressed male advertisement and underlying viability. This correlation was as high as 0.712 at one point. Discounting cases where C_{adv} or $C_{pref} = 0$, the maximum correlation was 0.618, at which time the male-female survival ratio was 0.488, the mean advertisement trait value was 0.961 and the mean preference was 0.596. At this point we have costly advertisements, costly preference, and the transmission of information about male quality. The results therefore support Iwasa et al.'s conclusion that an honest-signalling equilibrium could be evolutionarily stable under the terms of the conditional handicap.

Comparison of Figure 9.7 with Figure 9.10 shows us that the correlation between expressed

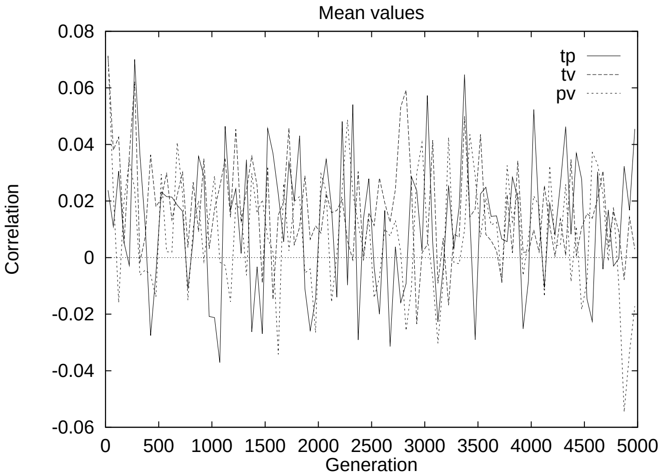

Figure 9.11: Correlations between the three genetic traits t, p and v over generational time. Data points on the time axis have been clumped into blocks of 50 generations and a mean value plotted.

advertisement and viability was generally lower in the pure epistasis handicap condition than in the conditional handicap case. For example, the maximum correlation in the pure epistasis condition was 0.269, which occurred when $C_{adv} = 1.5$ and $C_{pref} = 0.45$. At the same point in the conditional handicap case, the observed correlation was 0.613. In crude terms, this means that females gain more information about male quality when the conditional handicap is in place. Furthermore, it is interesting to note that in the pure epistasis condition, when advertising was free ($C_{adv} = 0$) the advertisements of males could not be trusted, i.e., they had almost zero correlation with viability. In the conditional handicap the exact opposite occurs: it is the cost-free advertisements that provide the most information about quality.

All of the simulations presented here have implemented a negative mutation bias on viability, which was Iwasa et al.'s (1991) first condition for stable signalling equilibria. However, their critical assumption that the genetic co-variances B_{pv}, B_{tp} and B_{tv} will remain constant and positive, and their second condition that ρ_{pv} should be greater than $\rho_{tp} \times \rho_{tv}$, have not been built into the simulation. We can now look at some additional results for the conditional handicap condition and try to determine—for a case in which honest signalling has evolved—whether the critical assumption has been satisfied and whether the second condition has been met.

Under the terms of the conditional handicap, and discounting cases when either advertisement or preference was cost-free, the maximum observed correlation between expressed advertisement and viability was 0.618. This occurred when $C_{adv} = 1.5$ and $C_{pref} = 0.15$, i.e., at the minimal positive cost values. In what follows, this point in the cost landscape will serve as an exemplar for the evolution of sexual signalling. Figure 9.11 shows, for the first of the ten simulation runs that were performed with these cost values, the correlations over generational time between each of the three genetic traits. The graph may appear complicated and difficult to interpret, but the basic message should be clear: the genetic correlations between the three traits were not constant and were not consistently positive. This implies that nor would the co-variances B_{pv}, B_{tp} and B_{tv} have been positive constants; strictly speaking, Iwasa et al.'s (1991) assumption is not supported. The correlations were also very low, never moving very far from zero, but signalling nevertheless evolved.

The same run was performed with perfect positive correlations between t, p and v imposed in the initial generation. The graph is not shown as the results were for all practical purposes identical to those of Figure 9.11. The strong genetic linkages between each trait disappeared after only a handful of generations.

Nevertheless, inspection of Figure 9.11 shows that the mean value for each of the three correlations is going to be weakly positive. Returning to data that has been averaged across the ten simulation runs conducted with $C_{adv} = 1.5$ and $C_{pref} = 0.15$, we find that the mean overall values were as follows: $\rho_{tp} = 0.012$, $\rho_{tv} = 0.029$, and $\rho_{pv} = 0.010$. These are extremely low correlations, and they imply that only about one-hundredth of one percent of the variance in one trait could be explained by the variance in another. Although it is always dangerous to overlook the cumulative effect of small factors in evolution, it is difficult to believe that these correlations are actually responsible for the maintenance of an honest signalling equilibrium. It cannot be the tiny correlation of 0.029 between the advertisement and viability genes that makes it worthwhile for females to bear costly preferences. There is another more important factor at work to produce the vastly higher correlation of 0.618 between the expressed advertisements of males and their underlying viability: it is of course the direct link between viability and advertisement imposed by the condition-dependent expression of the ornament trait. The implications of this will be discussed in section 9.6 below. Still, we can observe for the record that, given the observed correlations, the condition that ρ_{pv} must be greater than $\rho_{tp} \times \rho_{tv}$ has been satisfied.

Revealing handicap

The revealing handicap is similar to the conditional handicap, in that the male advertisement that females actually get to see has been influenced not only by the male's genetic tendency to grow a large ornament (t_{gen}) but by his viability (v_{phen}) as well. In the conditional handicap, viability exerts its moderating effect on the advertisement trait at the development stage, before the survival costs of advertisement have been determined. By contrast, in the revealing handicap males are assumed to produce an ornament as specified by their genome, and to bear the associated costs, but then before mating takes place their expressed advertisement is scaled according to their viability, i.e., $t_{new} = t_{old} \times v_{phen}$. This is supposed to reflect the idea that less viable males find it more difficult to maintain their ornament at its peak. We can imagine, for example, that less viable peacocks cannot avoid succumbing to parasite infestations that reduce the impact of their ornamental tails (see Hamilton & Zuk, 1982).

The results for the revealing handicap simulations were very similar to those for the conditional handicap, and therefore no additional graphs will be presented here. Neglecting cases where C_{adv} or $C_{pref} = 0$, male-female survival averaged 0.300, the mean value for the advertisement gene was 0.682, the mean value for preference was 0.354, and the average correlation between expressed advertisement and underlying viability was 0.183. The pattern of the data on correlation between advertisement and viability was the same as in the conditional handicap, with the highest values (up to 0.607) being observed when the cost of advertising was low. Overall, the findings from the simulation support Iwasa et al.'s (1991) claim that the revealing handicap, like the conditional one, can lead to costly-preference, honest-advertisement equilibria.

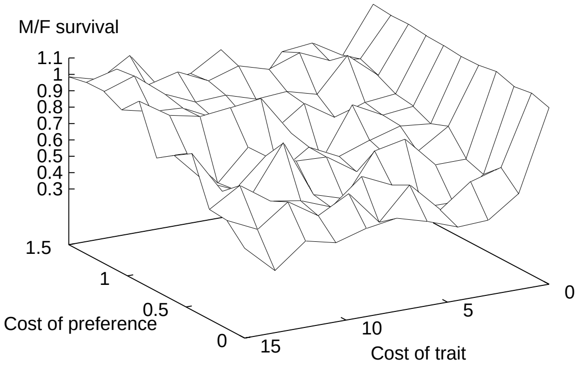

Figure 9.12: Male-female survival ratio in the strategic advertisement condition, by C_{adv} and C_{pref}.

Strategic advertisement-trait expression

Iwasa et al.'s (1991) model is all about the possibility that the expression of a male advertisement trait could come to provide worthwhile information about male viability to females. However, the advertisement trait itself is assumed to have a straightforward genetic basis. The model does not look at what should be expected if the expression of the male ornament depended not on a genetic trait, but on an inherited strategy for mapping viability to advertisement level. Expressing advertisement according to a strategy is what happens in Grafen's (1990b) model and in the simulation of the signalling of environmentally determined quality described in section 9.5.1. As noted earlier, there is no reason why heritable male quality could not also be the basis for a strategic advertisement scheme. Rather than males inheriting both viability and advertisement trait values from their parents, they could inherit a viability level plus a strategy or function for translating that viability level into a visible advertisement.

Simulation runs were performed to investigate the strategic signalling of heritable male quality. Viability and preference genes were dealt with as usual. As in the earlier model dealing with randomly determined viability, the strategy for mapping viability level into an advertisement was specified using two real-valued genes according to the scheme detailed in Figure 9.1. Note that—in contrast to the pure epistasis, conditional, and revealing handicap models—the males are now able to "choose" whether or not the expression of their advertisement will be condition-dependent. If the function mapping viability to advertisement has a positive slope, then expressed advertisements will be more or less honest, and will give females information about underlying male quality. But this is not enforced: the process of selection might also lead to uninformative advertisement strategies, such as producing an ornament of maximum size no matter what one's viability is.

Figure 9.12 shows the male-female survival ratio across the cost landscape. We can see that compared to the pure epistasis and conditional handicap conditions (Figures 9.6 and 9.8 respectively) males are not suffering such high mortality costs due to their ornaments. However, their survival is still significantly lower than female survival, especially when it is cheap for females to express a preference. Figure 9.13 shows the typical size of the expressed male advertisement. For medium to low values of advertisement and preference costs, males are clearly using strategies

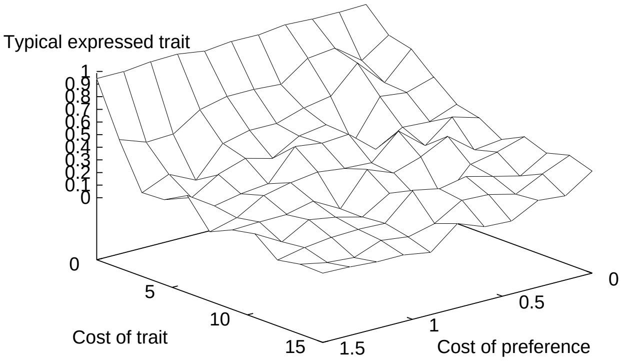

Figure 9.13: Typical expressed advertisement trait values in the strategic advertisement condition, by C_{adv} and C_{pref}. The typical trait value is calculated by substituting 0.5 into the genetically specified advertisement strategy of each member of the population.

that lead to the production of ornaments (although we need to be slightly cautious in interpreting this graph, as the typical advertisement is calculated by working out what advertisement each individual would produce if it were male and had a viability of 0.5). Mean preference values were somewhat uneven, but there was a general tendency for lower values as the cost of preference increased. The overall mean value (discounting cases where C_{adv} or $C_{pref} = 0$) was 0.299. Finally, the correlations between expressed advertisement and underlying viability are shown in Figure 9.14. For the lower values of C_{pref}, these correlations reached moderate values, up to a maximum of 0.417. We can therefore conclude that, in some parts of the cost landscape, proper signalling was occurring. The fact that moderate positive correlations were observed between advertisement and viability tells us that honest advertisement strategies were adopted when, in principle, males could easily have chosen to be uninformative.

9.5.3 Fisherian runaway sexual selection

In the sexual selection literature, the idea that males signal their underlying quality is only one of several competing explanations for the evolution of costly male ornaments (see section 2.7). The most important alternative theory is that male ornaments are the result of Fisher's runaway process, in which genetic linkage between the male advertisement and the female preference leads to a cycle of exaggeration which continues until checked by the mortality costs of the over-sized advertisements.

The results from the simulations reported thus far have established that, when the costs of advertisement and preference are right, male advertisements that function as honest indicators of quality can evolve. This has been the case for environmentally determined viability and for all four of the simulation variants involving genetically determined viability. According to the logic of Grafen's (1990b) proposed "Fisher index", the fact that we have consistently observed significant positive correlations between expressed advertisement and underlying viability rules out the possibility that we are looking at the results of a runaway process alone. On the other hand,

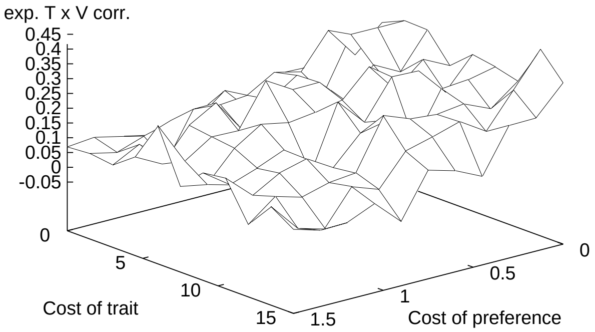

Figure 9.14: Correlation between expressed advertisement and underlying viability in the strategic advertisement condition, by C_{adv} and C_{pref}.

the fact that these correlations are not perfect suggests that runaway processes may be playing a part. In Grafen's terms, we have established that the Fisher index is not equal to one, but that does not mean that it is equal to zero.

It therefore seems prudent to investigate what happens in a simulation variant in which all individuals are always given the same constant value as a viability level. With no variation in viability, there is nothing for males to signal about. If costly advertisements and preferences nevertheless evolve in this condition, they must presumably be due to Fisher's runaway process. This would in turn suggest that runaway processes are partially responsible for the observed advertisement and preference levels in other conditions.

In this variation, the genetic viability trait technically still existed, but it was ignored and the phenotypic viability of a newborn individual was always set to 1.0. It follows that males with no ornament and females with no preference were guaranteed to survive to reproductive age. It is interesting to note that when this condition was first run, the small amount of variance in viability resulting from the random error term applied during the development stage was enough to get a signalling system going. It proved necessary to remove this phenotypic error term in order to ensure that every member of the population had *exactly* the same viability value of 1.0.

Figure 9.15 shows the male-female survival ratios. As the cost of advertisement increased, fewer males were surviving: this is evidence that sexual selection was occurring. The mean values for the advertisement gene were as high as about 0.95 when C_{adv} was low, and fell off smoothly to approximately 0.55 when $C_{adv} = 15$. Means for preference were more uneven, but the overall average was 0.44 and the highest values were observed when the cost of *advertising* was low. There was, of course, zero correlation between the expressed advertisement trait and underlying viability, which equates to a Fisher index of one. The evidence is unequivocal: costly male ornaments and costly female preference evolve within the model even when they cannot have a communicative function. It would be unwise to compare specific trait and preference means with results in the other conditions, because automatically assigning every individual a viability of 1.0 means that survival was easier in this condition, but generally speaking the male traits and the female

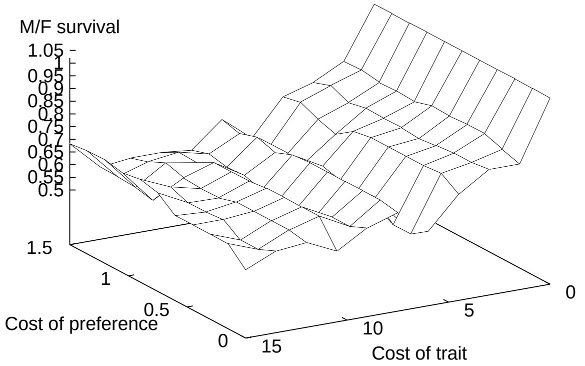

Figure 9.15: Male-female survival ratio in the runaway sexual selection condition, by C_{adv} and C_{pref}.

preferences were equal to or even greater than those observed in signalling scenarios.

The Fisher process is supposed to be driven by a genetic linkage between trait and preference. The observed genetic correlation between trait and preference averaged 0.021, discounting cases where C_{adv} or $C_{pref} = 0$. This value seems quite low, but we must assume that it was sufficient to start the runaway process because no other explanation for the observed costly ornaments and preferences is available. For comparison, the equivalent genetic correlation from the pure epistasis handicap condition was 0.023, and for the conditional handicap case it was 0.021. The close match between these values suggests that the runaway process was also playing a part in the earlier conditions: the honest advertisement of viability can co-exist with the pointless exaggeration of trait and preference due to genetic linkage.

9.6 Discussion

The results of the various simulations suggest that sexual advertisements can be proper signals of male quality, whether that quality is environmentally or genetically determined. The Fisher process of runaway sexual selection also appears to play a part in the evolution of costly male advertisements and female preferences.

Honest signalling of viability occurs in the pure epistasis handicap, despite the fact that Iwasa et al. (1991) claimed it could not. In the conditional and revealing handicaps, on the other hand, the results presented here are in accordance with Iwasa et al.'s prediction that honest signalling could be evolutionarily stable. Iwasa et al. intend their paper to clarify some of the controversies around the handicap principle. They argue that their findings explain why some earlier papers have concluded that the handicap principle can work while others have concluded that it cannot: different authors have tried to model different versions of the idea. Iwasa et al.'s intended clarification is an admirable goal; however, the results of the simulations presented here suggest that their conclusions must be taken with a grain of salt. In particular, their central assumption that genetic co-variances between advertisement, preference and viability could be treated as positive constants does not appear to be a reasonable one.

The conditional and revealing handicaps deserve closer scrutiny. Consider Figure 9.10, which shows the correlation between expressed advertisement and viability in the conditional handicap case; results for the revealing handicap were similar. The graph shows that the highest correlations were achieved when advertising was cost-free. In the pure epistasis handicap, by contrast, we find that "talk is cheap" in these cases: male advertisement was never an indicator of quality when the cost of advertising was zero. Why then, in the conditional and revealing handicap conditions, can females trust the advertisement levels of males who, in theory, can choose any advertisement level they like because there is no cost involved? The answer is that the males *cannot* choose any advertisement level that they might like. The stipulation that the expression of the ornament trait is condition-dependent (i.e., modified by viability) builds in an informational link between advertisement and viability in a rather uninteresting way. It seems disingenuous of Iwasa et al. to hold up the existence of costly female preference and honest advertisements as a deep result when the way in which the male trait is expressed itself enforces honesty. Thus we find that the genetic correlation between the advertisement trait and viability remains low in the conditional handicap case, but the correlation between the actual expressed advertisements and the underlying viability of adult males is very much higher: the condition-dependent expression of the ornament means that females automatically get useful information about viability. In addition, it is a little odd to claim that "handicap" signalling is occurring when the cost-free signals are the most reliable. Considered closely, Iwasa et al.'s claims for the conditional and revealing handicaps amount to little more than the uncontroversial observation that females will attend to unfakeable information about male quality.

Some caveats are necessary. Firstly, the results have shown, in various conditions, the simultaneous existence of costly advertisements, costly preferences, and a correlation between advertisement and viability. This constellation of symptoms has been interpreted as evidence for the evolution of communication, but in some cases the information transfer that was occurring may not quite meet the strict definition of proper signalling that was outlined in section 3.3. If a correlation develops between t and v, two separate genetic traits, and some observer infers something about v by observing t, then that is not proper signalling. In Millikan's terminology, the producer of the signal has not been selected to generate that signal in accordance with any kind of mapping rule. The male advertisements are not produced in accordance with a mechanism that relates them to viability; the phenotypic value of the advertisement trait is simply read off the genes. If female observers come to exploit an underlying genetic correlation, then it is a case of exploitation rather than one of proper signalling.

However, when the advertisement trait is determined as a strategic function of viability (either in the environmental or genetic viability cases) then the same logic does not apply. An advertisement trait that has been produced in accordance with a strategy, and thus may or may not be informative about quality, really can be regarded as a signal. The inherited strategy is the mapping rule that allows the expressed advertisement to qualify as an intentional icon and a proper signal. The fact that the advertisement strategies in the heritable viability case evolved toward honesty (at least under certain cost regimes) has additional implications. It suggests that in real-world cases, in which long-term evolution could presumably result in male advertisement being *either* the expression of a simple genetic trait or the expression of a strategy, selection will favour the latter. After all, the strategic advertisement condition is a general case that subsumes the pure epistasis and conditional handicaps: if the slope of the strategy evolves to zero, then the intercept can be thought of as a simple genetic trait; if the slope of the strategy is positive, then the expression of the advertisement is condition-dependent. Although the conditional handicap condition has been criticised above for enforcing honesty through its assumptions, it is surely of interest to find that selection will push for condition-dependent advertisement expression when other, uninformative strategies are also possible.

Finally, it should be noted that all of the simulation results depend on Grafen's proviso that the unit costs of advertisement (and in our case preference as well) should be lower for higher-quality individuals. One's faith in the simulation results must depend on one's faith in Grafen's proviso as a real-world condition. Some pilot studies, not reported here, indicated that for very narrow cost windows modest levels of honest sexual signalling might be possible without the proviso, i.e., when the costs of advertisement were independent of viability. An obvious topic for future simulation work would be to look at the effects of partial and complete failure to comply with Grafen's proviso.

Chapter 10

Conclusions

The thesis began with questions about the function of animal signalling systems. What are the selective advantages of the behaviours we call signals? Under what conditions will animals evolve to communicate with each other? The time has come to assess the extent to which we have been able to find answers.

Our first port of call was the biological literature on signalling. The earliest writers on the subject believed that the function of a signal was simply to transmit information about an animal's internal state, which was in its turn supposed to be an inherently good idea. The game-theoretic revolution in theoretical biology put an end to that notion, and in the more recent literature we find various explanations for the evolution of signalling. These range from Krebs and Dawkins's (1984) claim that signals *per se* do not really exist—that there are only attempts by one animal to manipulate another—through to Zahavi's (1975) handicap principle, which states that the function of a signal is to be costly and thus guarantee honesty. Although the authors of these theories sometimes claim that they are universally applicable (see e.g., Zahavi & Zahavi, 1997), it appears that some of them are better suited to particular ecological contexts than others. For example, the handicap principle seems most at home in the domain of sexual signalling, in which it was originally formulated. The argument from game theory that animals will maximize ambiguity about their intentions is most applicable to contests over finite resources, in which two animals' interests are maximally opposed. This observation partly inspired the construction of simulation models, described in chapters 7, 8 and 9, that were focussed on different kinds of possible communication behaviour.

In chapter 3 some philosophical foundations were laid. Most importantly, proper signalling was defined as a special kind of influence interaction; one in which a history of selection has favoured both the production of the signal and the performance of the response. Proper signalling was distinguished from cases of accidental influence, manipulation, and exploitation. This was not meant as a means of defining away the more problematic forms of animal communication, but as an argument for the importance of clarity and precision in the use of words like "signal" and "communicate". It may not be the case that all of the phenomena we carelessly lump together under the label "animal signalling" will admit of the same kinds of explanation. For instance, the manipulation and mind-reading described so vividly by Krebs and Dawkins no doubt exists in the animal world, but Millikan's notion of an intentional icon (upon which our definition of proper signalling was based) suggests that there may be behaviors out there that have a much closer fit to our everyday ideas about communication.

Many of the theories on signalling in the biological literature, as discussed in chapter 2, relied either on loose verbal arguments or on simple game-theoretic models with highly restrictive assumptions. This was one of the keystones of the argument built up in chapter 4 that evolutionary simulation models could help improve our understanding of the selective stories behind animal signalling systems. A critical review of work on this topic within the new field of artificial life (see chapter 5) showed that although the potential exists for artificial-life models to function as scientific tools, they have generally not done so to date.

With this groundwork out of the way, the main original contributions of the thesis were then presented in chapters 7, 8 and 9. In chapter 7 we used the evolution of food and alarm calls as a test-bed for examining Krebs and Dawkins's (1984) idea about there being two kinds of signal evolution: costly signalling when the interests of the participants conflict, and cheap conspiratorial whispers when the participants have common interests. Despite the popularity of this idea, Krebs and Dawkins's predictions were not in fact borne out in the context of a simple signalling game. Signalling generally only evolved when the participants had common interests; we cannot assume that the "costly signalling arms races" described by Krebs and Dawkins will occur in all contexts. In terms of answering a question about why animals communicate, the results from this chapter support the (admittedly somewhat banal) conclusion that signalling may serve to inform others of a certain state of affairs that will be mutually beneficial for the signaller and the receiver. However, there was also a novel finding that signals would be more costly when the positive payoff to the receiver was marginal.

In chapter 8, a model of animal contests over an indivisible resource was constructed. Enquist (1985) and Hurd (1997b) had claimed that cost-free, reliable signals of strength could evolve in such a case. More traditional game-theoretic views (Maynard Smith, 1982) suggested that a strength-signalling system could not be evolutionarily stable as it would always be open to exaggeration and bluff. On a variety of measures, there was no evidence that a communication had evolved in the simulation—this favours the standard game-theoretic view that predicts poker faces in contests. This tells us that "signalling one's strength to one's opponent" is not a plausible function of any signal, at least not in the kind of contest that was described.

Finally, a simulation of sexual signalling was developed in chapter 9. Iwasa et al.'s (1991) model of handicap signalling of heritable quality was tested with some of its restrictive assumptions concerning genetic co-variances relaxed. In keeping with Iwasa et al.'s findings, the simulation results suggested that male advertisement traits can indeed function as honest signals of underlying genetic quality, if we assume either the conditional or revealing versions of the handicap principle. There was a further limitation that the cost of the male advertisement trait should not be too high. However, in a finding that contradicted Iwasa et al., the simulation showed that the signalling of male quality could also be stable under the terms of the pure epistasis handicap. This latter version of the handicap principle is closer to Zahavi's original formulation, and the simulation result suggests that it can sometimes be the function of male sexual advertisement signals to significantly decrease the survival chances of the signaller, and to thereby serve as an honest index of quality.

Due to the deliberate investigation of the possibility of signalling in different ecological contexts, it is difficult to capture the results in a single generalization. However, if a one-line summary was absolutely necessary, it would have to be that communication is very hard to get started when any kind of conflict of interests exists. In chapter 7, with the exception of certain variant conditions, communication only evolved when there were positive payoffs for signalling and for receiving that outweighed the costs. In chapter 8 communication did not evolve. In chapter 9, on the other hand, we did find communication under many contexts, as long as the balance of costs was right, but all of these simulations enforced Grafen's proviso—the idea that the unit costs of display are lower for higher quality signallers—and it has already been established (Grafen, 1990a) that this condition is a force for honesty.

10.1 Biological implications

The simulation results may be of interest to those with some personal stake in the theoretical biology literature, but a field biologist might well complain that claims like "the honest signalling of quality will evolve as long as it is not too expensive" are not very informative. "What predictions do your simulations suggest", she might ask, "that I could go out and test in the real world?"

This is a valid question. While there are indeed some specific real-world predictions that arise from the simulations presented here, it should first be pointed out that the simulations could have been of some value even if no such predictions arose. The simulations were never intended to be detailed models of behaviour in a particular species, but as demonstrations of general principles. Their short-term point was—in keeping with the Quinean view of the nature of science sketched out in chapter 4—to settle particular theoretical questions. For example, in chapter 8 Enquist and Hurd were pitted against Maynard Smith. In chapter 9, we wanted to see whether Iwasa et al.'s predictions would hold up without their dubious assumptions. Clearly, if the point of theoretical biology is more than just to provide employment for theoretical biologists, then all of these issues must ultimately relate back to the real world, and be translated into testable, empirical predictions. However, that does not mean that simulations are themselves theories, and must make immediate testable claims. In biology as in most sciences there are many more theories in print than can possibly be true, and there is much to be said for a method that allows us to favour some over others. This is particularly so in cases like the evolution of behaviour, where field data is very hard to come by.

The idea is, then, that the biggest biological implications of the thesis cast their shadow on *theoretical* biology. Krebs and Dawkins should acknowledge the elusiveness of their costly signalling arms races. Enquist and Hurd should not go on claiming with impunity that cost-free signals of fighting ability can be evolutionarily stable, at least not without responding to the simulation presented here. Iwasa et al. should not imagine that the pure epistasis handicap never works, and that only the conditional and revealing handicaps are worth bothering with. Of course, the proponents of a particular mathematical ESS model—faced with a contradictory simulation—could always claim that their logic had not been faulted, and that they had described an ESS within the context of their model. This is so; the authors whose work is contradicted here are not being accused of mere calculation errors. However, if a simple mathematical model is supposed to apply to real-world cases at least in a limited way, then surely it should work with reference to simulated cases of intermediate complexity.

Having said all that, there are in fact some aspects of the three simulations that do suggest testable hypotheses. The novel finding from chapter 7, that signals will be more costly when the return to the receiver is marginal, could be cashed out in many ways. For example, in a situation in which nestlings are begging to their parents for food, the parental donation of food items certainly benefits the chick, and presumably benefits the parent in terms of inclusive fitness. However, to the extent that there are extra-pair copulations in this species, there is a possibility that the apparent father is not the genetic father. This means that across the two "games" that the chicks are playing with the two parents, the positive inclusive-fitness returns will be more marginal for the father than for the mother. We could therefore predict that chicks will signal in a louder and costlier fashion to their fathers than to their mothers. The complete lack of altruistic communication in the simulation described in chapter 7 also suggests that no altruistic food or alarm calls will evolve for use between non-relatives if there is no potential for reciprocity in the species. The logic behind this claim runs as follows: the simulation established that variations in the cost of the signal cannot get altruistic signalling started, and the only other plausible mechanisms that remain are kin selection and reciprocal altruism. Therefore altruistic calls will not evolve between unrelated animals that cannot reciprocate. It follows that we should expect not to find altruistic calls in simple animals that are not capable of recognizing their conspecifics, or that do not interact repeatedly—except when these simple animals are interacting predominantly with their kin.

The findings from chapter 8 imply that animals engaging in contests over food, mates or territory will not signal their strength or fighting ability honestly if they have a choice in the matter. That is, if there are no unfakeable cues about strength available—let us suppose that the competitors are all roughly the same size, and differ only in their muscular efficiency—then selection will not favour any behaviours that serve to indicate an animal's strength. (Of course, this relies on the fairly reasonable assumption that the costs of serious injury are greater than the reward for winning a contest.) The extreme behaviours observed in the simulation, in which animals tended to either be very aggressive or to flee immediately, are strongly reminiscent of the mixed-strategy equilibrium in Maynard Smith's original hawk-dove game; this suggests that some of the real-world behaviour observed in animal contests and often interpreted as "threats" or "bluffs" might in fact be the result of stochastically selected extreme strategies. It was an interesting feature of the simulation that the mean fitness in the experimental condition was negative, while fitness in the unfakeable control was positive; this means that if the animals had had an option by which they could avoid fighting altogether, they might have been expected to take it in the experimental case but not in the unfakeable control. In the real world, non-aggression is certainly an option exercised by many species. Although this may be reading too much into the details of a particular model, we can speculate that those animal species that *do* engage in aggressive confrontations may be precisely those that have access to unfakeable cues about fighting ability.

Finally, the results from chapter 9 provide good evidence that, as long as Grafen's proviso can be shown to be true in a particular case, heritable male quality can be honestly advertised. Therefore the observation that a male advertisement trait is correlated with some measure of genetic quality does not mean that the expression of the trait is condition-dependent—the pure epistasis handicap can work too. The data on genetic correlations from the simulation suggest that very low correlations between trait, preference and viability, as measured in newborn organisms, may nevertheless support the evolution of an honest signalling equilibrium. This, as well as Grafen's (1990b) work on the Fisher index, implies that the important correlation to measure is the one between the magnitude of the expressed advertisement and underlying genetic quality in males that survive to breeding age.

10.2 Limitations of the thesis

Each of the simulations have important limitations that have been described in the respective chapters. To reiterate: in chapter 7 the simulation only captures those situations in which the signaller is ambivalent about the receiver's response in the low state; the model is also founded on a certain view of what constitutes a conflict of interests. It is perhaps a deeper limitation that the thesis assumes the origin of such things as signal perception and turn-taking to be relatively unproblematic. If these were treated as part of the phenomenon of interest, rather than being assumed as prerequisites, it is not clear how far our views on communication would have to shift (although see Di Paolo, 1997b).

The generality of the results from chapter 8 is greatly constrained by the fact that the simulation models the unusual situation in which animals cannot in any way detect each other's strength and have nothing like recognition or memory for the results of previous contests. It is also the case that after each contest the animal starts afresh, with a new randomly determined fighting ability and no advantage or disadvantage in the current contest based on the result of the previous one. Furthermore, the animals cannot influence the frequency with which they get involved in contests, aggression is assumed to be captured by a single dimension, the animals cannot (except in a variant condition) detect the level of damage that they have suffered, and there is an artificial cap on the amount of damage that can be sustained in any one contest.

In chapter 9 there is a basic assumption that Grafen's proviso is true for the advertisement trait in question. There are several other more prosaic limitations: for instance, genes are implemented as real values, rather than any attempt being made to simulate a chromosome. Females, if they exercise a preference, choose the best male from a lek of a constant size; other methods of female choice were experimented with in pilot studies but none have been reported here.

More generally, the limits on the sensible use of *any* simulation have already been discussed in great detail in chapters 4 and 5: simulations do not constitute empirical data; one cannot *prove* anything of empirical interest with a simulation, only demonstrate sufficiency or the lack of it, etc. It is hoped that by now the reader will agree that simulations are nevertheless useful tools in determining which theories are worth exploring further in particular domains.

Another general limitation is that the tight links to biological theory present throughout the thesis mean that we have only looked at issues that have previously been treated in the biological literature. There may be aspects of communication behaviour that have not yet been considered by biologists, and that we have therefore needlessly neglected. In one sense this is certainly true. However, the implicit objection here can be answered in two ways. Firstly, there is nothing to stop evolutionary simulation methods from being applied to ideas that come from further afield than biology. Secondly, the limitation of having to take small steps from what has gone before is inevitable in what Kuhn (1962) somewhat disparagingly referred to as normal science.

A final limitation of the thesis is that any simulation result of the form “X did not evolve”, as is asserted in chapters 7 and 8, makes the assumption that the phenomenon X was an accessible point in the evolutionary space being explored. It is always possible, however, that some artefact in the simulation has prevented the evolution of the phenomenon, despite the fact that under more realistic conditions it would be selected for. This is of particular concern for the simulation described in chapter 8: if the CTRNNs used as a control architecture turned out to be inadequate to embody an evolved signalling system, then the reported result would be a false negative. There is no quick and easy solution to this problem. Given that no fixed guidelines yet exist for the use of such tools as genetic algorithms and artificial neural networks, it would appear that evolutionary-simulation builders must keep one eye on the research literature for the phenomena they are modelling, and another on the literature associated with the tools of their trade.

10.3 Future work

In chapter 4 and in other parts of the thesis, an argument has been developed for the use of evolutionary simulation models as a way of extending the reach of theoretical biology on complex topics such as animal communication. The three simulations presented could only hope to be examples plucked from a vast field of possible modelling projects. Thus the avenues for future work are extremely broad, and no attempt will be made here to list all of the possible domains in which evolutionary simulation modelling might be fruitfully applied.

However, a problem with this kind of work, as discussed in section 10.1 above, is that the empirical implications may not always be clear. Evolutionary simulations can be abstract testing grounds for comparing the plausibility of different theories—as has been their role here—but they can also implement detailed models of particular behaviours in particular species. A sensible direction for the future seems to lie in progressively extending models like those in chapters 7, 8 and 9 until they are capable of making concrete predictions. A good example of the idea is Davis and Todd’s (1998) study of parental feeding strategies in the Western bluebird *Sialia mexicana*; this is not just an abstract model of a foraging problem, but incorporates real data on bluebird metabolism, typical foraging flight times, number of days before a chick leaves the nest, etc. One can imagine an extension of the model presented in chapter 8, for instance, that was specifically targeted at the mantis shrimp *Gonodactylus bredini* and permitted such clearly testable predictions as “Shrimps will refuse to fight within 4 hours either side of moulting”, or “Three claw strikes is the optimal threat display.”

Such detailed simulations have their own problems of data-gathering and computational complexity, however. In the more immediate future, minor extensions of the work presented here could investigate several promising factors. Firstly, the spatial arrangement of a population is one of the most obvious ways in which simulation models can improve upon mathematical ones, and yet the effects of space and locality have only been investigated to a very limited extent in chapter 7. What would happen in a fully spatial simulation of animal contests, for example, in which confrontations only occurred between animals that approached each other in a larger environment? Might there be important spatial effects in lek-based mating, such that the superior males gain central positions on the lek and this in itself serves as an advertisement of quality? The effects of learning and memory could also be of interest: if alarm- or food-calling animals could remember whether their partner co-operated with them last time, or if a pair of competing animals both knew who had won the previous encounter between the two, then the results would probably be very different from those presented here. Similarly, social effects could be implemented in a richer way: if lekking females could observe the mate choices of other females, what they saw would be likely to influence their own choices. If food- or alarm-calling was going on both within a tight kin group and amongst a broader herd-like aggregation, then there might be room for two kinds of signal co-evolution after all.

Bibliography

Ackley, D. H., & Littman, M. L. (1994). Altruism in the evolution of communication. In Brooks, R. A., & Maes, P. (Eds.), *Artificial Life IV*, pp. 40–48. MIT Press, Cambridge, MA.

Adams, E. S., & Caldwell, R. L. (1990). Deceptive communication in asymmetric fights of the stomatopod crustacean *Gonodactylus bredini*. *Animal Behaviour*, *39*, 706–716.

Adams, E. S., & Mesterton-Gibbons, M. (1995). The cost of threat displays and the stability of deceptive communication. *Journal of Theoretical Biology*, *175*, 405–421.

Allen, C., & Bekoff, M. (1997). *Species of Mind: The Philosophy and Biology of Cognitive Ethology*. MIT Press, Cambridge, MA.

Andersson, M. (1976). Social behaviour and communication in the Great Skua. *Behaviour*, *58*, 40–77.

Andersson, M. (1986). Evolution of condition-dependent sex ornaments and mating preferences: Sexual selection based on viability differences. *Evolution*, *40*, 804–816.

Andersson, M. (1994). *Sexual Selection*. Princeton University Press, Princeton, NJ.

Arak, A., & Enquist, M. (1993). Hidden preferences and the evolution of signals. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *340*, 207–213.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books, New York.

Batali, J. (1994). Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax. In Brooks, R. A., & Maes, P. (Eds.), *Artificial Life IV*, pp. 160–171. MIT Press, Cambridge, MA.

Bekoff, M., & Allen, C. (1992). Intentional icons: Towards an evolutionary cognitive ethology. *Ethology*, *91*(1), 1–16.

Bell, G. (1978). The “handicap” principle of sexual selection. *Evolution*, *32*, 872–885.

Bennett, J. (1976). *Linguistic Behaviour*. Cambridge University Press, Cambridge.

Bhaskar, R. (1978). *A Realist Theory of Science* (Second edition). Harvester Wheatsheaf, Hemel Hempstead, UK.

Bickerton, D. (1994). Origin and evolution of language. In Asher, R. E. (Ed.), *The Encyclopedia of Language and Linguistics*, pp. 2881–2883. Pergamon Press, Oxford.

Binmore, K. (1992). *Fun and Games: A Text on Game Theory*. Heath, Lexington, MA.

Boorman, S. A., & Levitt, P. R. (1972). Group selection on the boundary of a stable population. *Proceedings of the National Academy of Sciences, USA*, *69*(9), 2711–2713.

Boorman, S. A., & Levitt, P. R. (1973). Group selection on the boundary of a stable population. *Theoretical Population Biology*, *4*(1), 85–128.

Bradbury, J. W., Gibson, R. M., & Tsai, I. M. (1986). Hotspots and the evolution of leks. *Animal Behaviour*, *34*, 1694–1709.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA.

Briskie, J. V., Naugler, C. T., & Leech, S. M. (1994). Begging intensity of nestling birds varies with sibling relatedness. *Proceedings of the Royal Society of London: Biological Sciences*, *258*, 73–78.

Bullock, S. (1997a). *Evolutionary Simulation Models: On their Character, and Application to Problems Concerning the Evolution of Natural Signalling Systems*. Ph.D. thesis, School of Cognitive and Computing Sciences, University of Sussex, Brighton, UK.

Bullock, S. (1997b). An exploration of signalling behaviour by both analytic and simulation means for both discrete and continuous models. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 454–463. MIT Press / Bradford Books, Cambridge, MA.

Bullock, S. (1998). 1-800 CALL ME: The myth of the passive prospect. Paper presented to the Artificial Life Reading Group, School of Cognitive and Computing Sciences, University of Sussex, 13 February.

Burghardt, G. M. (1970). Defining 'communication'. In Johnston, Jr., J. W., Moulton, D. G., & Turk, A. (Eds.), *Communication by Chemical Signals*. Appleton-Century-Crofts, New York.

Caro, T. M., & Hauser, M. D. (1992). Is there teaching in nonhuman animals? *Quarterly Review of Biology*, *67*, 151–174.

Caryl, P. G. (1979). Communication by agonistic displays: What can games theory contribute to ethology? *Behaviour*, *67*, 136–169.

Caryl, P. G. (1982). Animal signals: A reply to Hinde. *Animal Behaviour*, *30*, 240–244.

Caryl, P. G. (1987). Acquisition of information in contests: The gulf between theory and biology. Paper presented at the ESS Workshop on Animal Conflicts, Sheffield, UK, July.

Cheney, D. L., & Seyfarth, R. M. (1982). How vervet monkeys perceive their grunts: Field playback experiments. *Animal Behaviour*, *30*, 739–751.

Cheney, D. L., & Seyfarth, R. M. (1990). *How Monkeys See the World*. University of Chicago Press, Chicago.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. (1968). *Language and Mind*. Harcourt, Brace and World, New York.

Chomsky, N. (1975). *Reflections on Language*. Pantheon Books, New York.

Clark, A. (1996). Happy couplings: Emergence and explanatory interlock. In Boden, M. A. (Ed.), *The Philosophy of Artificial Life*, pp. 262–281. Oxford University Press, Oxford.

Clutton-Brock, T., Albon, S. D., Gibson, R. M., & Guinness, F. E. (1979). The logical stag: adaptive aspects of fighting in red deer *(Cervus elaphus L.)*. *Animal Behaviour*, *27*, 211–225.

Collins, R. J., & Jefferson, D. R. (1991). Antfarm: Towards simulated evolution. In Langton, C. G., Taylor, C., Farmer, J. D., & Rasmussen, S. (Eds.), *Artificial Life II*, pp. 579–601. Addison-Wesley, Redwood City, CA.

Cullen, J. M. (1972). Some principles of animal communication. In Hinde, R. A. (Ed.), *Non Verbal Communication*, pp. 101–122. Cambridge University Press, Cambridge.

Cummins, R. (1994). Functional analysis. In Sober, E. (Ed.), *Conceptual Issues in Evolutionary Biology*, pp. 49–70. MIT Press / Bradford Books, Cambridge, MA.

Dabelsteen, T., & Pedersen, S. B. (1990). Song and information about aggressive responses of blackbirds, *Turdus merula:* Evidence from interactive playback experiments with territory owners. *Animal Behaviour*, *40*, 1158–1168.

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection*. John Murray, London.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*. John Murray, London.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. John Murray, London.

Dautenhahn, K. (1995). Getting to know each other—artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, *16*, 333–356.

Davidson, D. (1970). Mental events. In Foster, L., & Swanson, J. W. (Eds.), *Experience and Theory*. University of Massachusetts Press, Amherst, MA.

Davis, J. W. F., & O'Donald, P. (1976). Sexual selection for a handicap: A criticial analysis of Zahavi's model. *Journal of Theoretical Biology*, *57*, 345–354.

Davis, J. N., & Todd, P. M. (1998). Simple decision rules for parental investment. In Gigerenzer, G., & Todd, P. M. (Eds.), *Simple Heuristics That Make Us Smart*. Oxford University Press, New York.

Dawkins, M. S. (1993). Are there general principles of signal design? *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *340*, 251–255.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press, Oxford.

Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation? In Krebs, J. R., & Davies, N. B. (Eds.), *Behavioural Ecology: An Evolutionary Approach*, pp. 282–309. Blackwell, Oxford.

de Boer, B. (1997). Generating vowel systems in a population of agents. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 503–510. MIT Press / Bradford Books, Cambridge, MA.

de Bourcier, P., & Wheeler, M. (1994). Signalling and territorial aggression: An investigation by means of synthetic behavioural ecology. In Cliff, D., Husbands, P., Meyer, J.-A., & Wilson, S. W. (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pp. 463–472. MIT Press / Bradford Books, Cambridge, MA.

de Bourcier, P., & Wheeler, M. (1995). Aggressive signaling meets adaptive receiving: Further experiments in synthetic behavioural ecology. Cognitive science research paper 364, School of Cognitive and Computing Sciences, University of Sussex, Brighton, UK.

de Bourcier, P., & Wheeler, M. (1997). The truth is out there: The evolution of reliability in aggressive communication systems. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 444–453. MIT Press / Bradford Books, Cambridge, MA.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press / Bradford Books, Cambridge, MA.

Dennett, D. C. (1991a). *Consciousness Explained*. Allen Lane, London.

Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy*, *88*, 27–51.

Di Paolo, E. A. (1996). Some false starts in the construction of a research methodology for artificial life. In Noble, J., & Parsowith, S. R. (Eds.), *The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex*. Cognitive science research paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Di Paolo, E. A. (1997a). An investigation into the evolution of communication. *Adaptive Behavior*, *6*(2), 285–324.

Di Paolo, E. A. (1997b). Social coordination and spatial organization: Steps towards the evolution of communication. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 464–473. MIT Press / Bradford Books, Cambridge, MA.

Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press / Bradford Books, Cambridge, MA.

Dugatkin, L. A., & Reeve, H. K. (1994). Behavioral ecology and the "levels of selection": Dissolving the group selection controversy. *Advances in the Study of Behavior*, *23*, 101–133.

Dunham, D. W. (1966). Agonistic behaviour in captive Rose-breasted Grosbeaks, *Pheucticus ludovicianus* (L.). *Behaviour*, *27*, 160–173.

Emlen, J. M. (1973). *Ecology: An Evolutionary Approach*. Addison-Wesley, Reading, MA.

Endler, J. A. (1993). Some general comments on the evolution and design of animal communication systems. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *340*, 215–225.

Enquist, M. (1985). Communication during aggressive interactions with particular reference to variation in choice of behaviour. *Animal Behaviour*, *33*, 1152–1161.

Evans, C. S., & Marler, P. (1994). Food-calling and audience effects in male chickens (*Gallus gallus*): Their relationships to food availability, courtship and social facilitation. *Animal Behaviour*, *47*, 1159–1170.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press, London.

Fodor, J. (1968). *Psychological Explanation: An Introduction to the Philosophy of Psychology*. Random House, New York.

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press, Cambridge, MA.

Fodor, J. (1990). *A Theory of Content*. MIT Press, Cambridge, MA.

Fontana, W., Wagner, G., & Buss, L. W. (1994). Beyond digital naturalism. *Artificial Life*, *1*(1/2), 211–227.

Franceschini, N., Pichon, J. M., & Blanes, C. (1992). From insect vision to robot vision. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *337*, 283–294.

Gardner, M. (1970). Mathematical games. *Scientific American*, *October*, 120–123.

Gardner, R., & Morris, M. R. (1989). The evolution of bluffing in animal contests: An ESS approach. *Journal of Theoretical Biology*, *137*, 235–243.

Godfray, H. C. J. (1991). Signalling of need by offspring to their parents. *Nature*, *352*, 328–330.

Gomulkiewicz, R. (1998). Game theory, optimization, and quantitative genetics. In Dugatkin, L. A., & Reeve, H. K. (Eds.), *Game Theory and Animal Behavior*, pp. 283–303. Oxford University Press, New York.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London: Biological Sciences*, *205*, 581–598.

Grafen, A. (1990a). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*, 517–546.

Grafen, A. (1990b). Sexual selection unhandicapped by the Fisher process. *Journal of Theoretical Biology*, *144*, 473–516.

Grafen, A. (1991). Modelling in behavioural ecology. In Krebs, J. R., & Davies, N. B. (Eds.), *Behavioural Ecology: An Evolutionary Approach* (Third edition), pp. 5–31. Blackwell, Oxford.

Grice, H. P. (1969). Utterer's meaning and intention. *Philosophical Review*, *68*, 147–177.

Guilford, T., & Dawkins, M. S. (1991). Receiver psychology and the evolution of animal signals. *Animal Behaviour*, *42*, 1–14.

Haldane, J. B. S. (1932). *The Causes of Evolution*. Longmans, London.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*, 1–52.

Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature*, *228*, 1218–1220.

Hamilton, W. D., & Zuk, M. (1982). Heritable true fitness and bright birds: A role for parasites? *Science*, *218*, 384–387.

Hammerstein, P. (1998). What is evolutionary game theory? In Dugatkin, L. A., & Reeve, H. K. (Eds.), *Game Theory and Animal Behavior*, pp. 3–15. Oxford University Press, New York.

Hansen, A. J. (1986). Fighting behaviour in bald eagles: A test of game theory. *Ecology*, *67*, 787–797.

Harper, D. G. C. (1991). Communication. In Krebs, J. R., & Davies, N. B. (Eds.), *Behavioural Ecology: An Evolutionary Approach* (Third edition), pp. 374–397. Blackwell, Oxford.

Harvey, I. (1996). Untimed and misrepresented: connectionism and the computer metaphor. *AISB Quarterly*, *96*, 20–27.

Hashimoto, T. (1997). Usage-based structuralization of relationships between words. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 483–492. MIT Press / Bradford Books, Cambridge, MA.

Hasson, O. (1994). Cheating signals. *Journal of Theoretical Biology*, *167*, 223–238.

Hauser, M. D. (1996). *The Evolution of Communication*. MIT Press / Bradford Books, Cambridge, MA.

Helmreich, S. (1995). The word for world is computer: An anthropological expedition into artificial worlds, second natures, and artificial life. Tech. rep., Santa Fe Institute, Santa Fe, NM. Later withdrawn from the SFI's *Working Papers* series, may be hard to find.

Hinde, R. A. (1981). Animal signals: Ethological and games-theory approaches are not incompatible. *Animal Behaviour*, *29*, 535–542.

Hinton, G. E., & Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems*, *1*, 495–502.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.

Huberman, B. A., & Glance, N. S. (1993). Evolutionary games and computer simulations. *Proceedings of the National Academy of Sciences, USA*, *90*(16), 7715–7718.

Hume, D. (1748 / 1955). *An Inquiry Concerning Human Understanding*. Bobbs-Merrill, New York.

Hurd, P. L. (1995). Communication in discrete action-response games. *Journal of Theoretical Biology*, *174*, 217–222.

Hurd, P. L. (1997a). *Game Theoretical Perspectives on Conflict and Biological Communication*. Ph.D. thesis, Department of Zoology, Stockholm University, Sweden.

Hurd, P. L. (1997b). Is signalling of fighting ability costlier for weaker individuals? *Journal of Theoretical Biology*, *184*, 83–88.

Huxley, J. S. (1923). Courtship activities of the red-throated diver (*Colymbus stellatus* Pontopp.); together with a discussion on the evolution of courtship in birds. *Journal of the Linnaean Society*, *35*, 253–293.

Huxley, J. S. (1966). Ritualization of behaviour in animals and men. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *251*, 249–271.

Iwasa, Y., Pomiankowski, A., & Nee, S. (1991). The evolution of costly mate preferences II. The "handicap" principle. *Evolution*, *45*(6), 1431–1442.

Johnstone, R. A. (1994). Honest signalling, perceptual error and the evolution of 'all-or-nothing' displays. *Proceedings of the Royal Society of London: Biological Sciences*, *256*, 169–175.

Johnstone, R. A. (1997). The evolution of animal signals. In Krebs, J. R., & Davies, N. B. (Eds.), *Behavioural Ecology: An Evolutionary Approach* (Fourth edition), pp. 155–178. Blackwell, Oxford.

Johnstone, R. A. (1998). Game theory and communication. In Dugatkin, L. A., & Reeve, H. K. (Eds.), *Game Theory and Animal Behavior*, pp. 94–117. Oxford University Press, New York.

Kauffman, S. A. (1993). *The Origins of Order*. Oxford University Press, New York.

Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research*, *52*, 1–26.

Kirby, S., & Hurford, J. (1997). Learning, culture and evolution in the origin of linguistic constraints. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 493–502. MIT Press / Bradford Books, Cambridge, MA.

Kirkpatrick, M. (1986). The handicap mechanism of sexual selection does not function. *American Naturalist*, *127*, 222–240.

Kitano, H., Hamahashi, S., Kitazawa, J., Takao, K., & Imai, S.-i. (1997). The virtual biology laboratories: A new approach to computational biology. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 274–283. MIT Press / Bradford Books, Cambridge, MA.

Klump, G. M., & Shalter, M. D. (1984). Acoustic behaviour of birds and mammals in the predator context. i. factors affecting the structure of alarm signals. ii. the functional significance and evolution of alarm signals. *Zeitschrift für Tierpsychologie*, *66*, 189–226.

Krakauer, D. C., & Johnstone, R. A. (1995). The evolution of exploitation and honesty in animal communication: A model using artificial neural networks. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *348*, 355–361.

Krebs, J. R., & Davies, N. B. (1981). *An Introduction to Behavioural Ecology*. Blackwell, Oxford.

Krebs, J. R., & Davies, N. B. (Eds.). (1997). *Behavioural Ecology: An Evolutionary Approach* (Fourth edition). Blackwell, Oxford.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind reading and manipulation. In Krebs, J. R., & Davies, N. B. (Eds.), *Behavioural Ecology: An Evolutionary Approach* (Second edition), pp. 380–402. Blackwell, Oxford.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press, Chicago.

Langton, C. G. (1989). Artificial life. In Langton, C. G. (Ed.), *Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems (ALIFE '87)*, pp. 1–48. Addison-Wesley, Redwood City, CA.

Levins, R. (1970). Extinction. In Gerstenhaber, M. (Ed.), *Some Mathematical Questions in Biology*, Vol. 2 of *Lectures on Mathematics in the Life Sciences*, pp. 77–107. American Mathematical Society, Providence, RI.

Lewis, D. B., & Gower, D. M. (1980). *Biology of Communication*. Blackie, Glasgow.

Lorenz, K. (1935). Der Kumpan in der Umwelt des Vogels. *Journal of Ornithology*, *83*, 137–215; 289–413.

Lorenz, K. (1967). *On Aggression*. Methuen, London. Translated by M. Latzke.

MacLennan, B. (1991). Synthetic ethology: An approach to the study of communication. In Langton, C. G., Taylor, C., Farmer, J. D., & Rasmussen, S. (Eds.), *Artificial Life II*. Addison-Wesley, Redwood City, CA.

MacLennan, B. J., & Burghardt, G. M. (1994). Synthetic ethology and the evolution of cooperative communication. *Adaptive Behavior*, *2*(2), 161–188.

Marchetti, K. (1993). Dark habitats and bright birds illustrate the role of the environment in species divergence. *Nature*, *362*, 149–152.

Marler, P. (1957). Specific distinctness in the communication signals of birds. *Behaviour*, *11*, 13–40.

Mataric, M. (1994). Learning to behave socially. In Cliff, D., Husbands, P., Meyer, J.-A., & Wilson, S. W. (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pp. 453–462. MIT Press / Bradford Books, Cambridge, MA.

Maturana, H. R., & Varela, F. J. (1980). Autopoiesis: The organization of the living. In Maturana, H. R., & Varela, F. J. (Eds.), *Autopoiesis and Cognition: The Realization of the Living*, pp. 59–138. Reidel, Dordrecht, Holland.

May, R. M., Bonhoeffer, S., & Nowak, M. A. (1995). Spatial games and evolution of cooperation. In Morán, F., Moreno, A., Merelo, J. J., & Chacón, P. (Eds.), *Advances in Artificial Life: Third European Conference on Artificial Life (ECAL '95)*, Vol. 929 of *Lecture Notes in Artificial Intelligence*, pp. 749–759. Springer, Berlin.

Maynard Smith, J. (1974a). *Models in Ecology*. Cambridge University Press, Cambridge.

Maynard Smith, J. (1974b). The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*, *47*, 209–221.

Maynard Smith, J. (1976). Sexual selection and the handicap principle. *Journal of Theoretical Biology*, *57*, 239–242.

Maynard Smith, J. (1978). *The Evolution of Sex*. Cambridge University Press, Cambridge.

Maynard Smith, J. (1979). Game theory and the evolution of behaviour. *Proceedings of the Royal Society of London: Biological Sciences*, *205*, 475–488.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Maynard Smith, J. (1985). Sexual selection, handicaps and true fitness. *Journal of Theoretical Biology*, *115*, 1–8.

Maynard Smith, J. (1989). *Evolutionary Genetics*. Oxford University Press, Oxford.

Maynard Smith, J. (1991). Honest signalling: The Philip Sydney game. *Animal Behaviour*, *42*, 1034–1035.

Maynard Smith, J. (1993). *The Theory of Evolution* (Canto edition). Cambridge University Press, Cambridge.

Maynard Smith, J., & Harper, D. G. C. (1995). Animal signals: Models and terminology. *Journal of Theoretical Biology*, *177*, 305–311.

Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*, 15–18.

Meyer, J.-A. (1994). The animat approach to cognitive science. In Roitblat, H., & Meyer, J.-A. (Eds.), *Comparative Approaches to Cognitive Science*. MIT Press, Cambridge, MA.

Miller, G. F., & Todd, P. M. (1998). Mate choice turns cognitive. *Trends in Cognitive Sciences*, *2*(5), 161–201.

Miller, G. F. (1995). Artificial life as theoretical biology: How to do real science with computer simulation. Cognitive science research paper 378, School of Cognitive and Computing Sciences, University of Sussex, Brighton, UK.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. MIT Press / Bradford Books, Cambridge, MA.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice*. MIT Press / Bradford Books, Cambridge, MA.

Mitchell, M., & Forrest, S. (1994). Genetic algorithms and artificial life. *Artificial Life*, *1*(3), 267–289.

Møller, A. P. (1988). Female choice selects for male sexual tail ornaments in the monogamous swallow. *Nature*, *332*, 640–642.

Møller, A. P. (1989). Viability costs of male tail ornaments in a swallow. *Nature*, *339*, 132–135.

Møller, A. P. (1991). Parasite load reduces song output in a passerine bird. *Animal Behaviour*, *41*, 723–730.

Morris, D. (1957). 'Typical intensity' and its relation to the problem of ritualization. *Behaviour*, *11*, 1–12.

Morton, E. S. (1975). Ecological sources of selection on avian sounds. *American Naturalist*, *109*, 17–34.

Moukas, A., & Hayes, G. (1996). Synthetic robotic language acquisition by observation. In Maes, P., Matarić, M., Meyer, J.-A., Pollack, J., & Wilson, S. W. (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pp. 568–579. MIT Press / Bradford Books, Cambridge, MA.

Nelson, D. A. (1984). Communication of intentions in agonistic contexts by the pigeon guillemot, *Cepphus columba*. *Behaviour*, *88*, 145–189.

Noble, J. (1997). The scientific status of artificial life. Poster presented at the Fourth European Conference on Artificial Life (ECAL'97), Brighton, UK, 28–31 July. Available WWW: http://www.cogs.susx.ac.uk/ecal97/present.html [1997, 21 August].

Noble, J. (1998a). The evolution of communication with and without conflicts of interest: An argument for modelling proto-proto-language. Paper presented at the Second International Conference on the Evolution of Language, London, UK, 6–9 April.

Noble, J. (1998b). Evolved signals: Expensive hype vs. conspiratorial whispers. In Adami, C., Belew, R., Kitano, H., & Taylor, C. (Eds.), *Artificial Life VI*, pp. 358–367. MIT Press, Cambridge, MA.

Noble, J. (1998c). Tough guys don't dance: Intention movements and the evolution of signalling in animal contests. In Pfeifer, R., Blumberg, B., Meyer, J.-A., & Wilson, S. W. (Eds.), *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, pp. 471–476. MIT Press / Bradford Books, Cambridge, MA.

Noble, J., & Cliff, D. (1996). On simulating the evolution of communication. In Maes, P., Matarić, M., Meyer, J.-A., Pollack, J., & Wilson, S. W. (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pp. 608–617. MIT Press / Bradford Books, Cambridge, MA.

Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*, 826–829.

Nur, N., & Hasson, O. (1983). Phenotypic plasticity and the handicap principle. *Theoretical Biology*, *110*, 375–398.

Oliphant, M. (1996). The dilemma of Saussurean communication. *BioSystems*, *37*, 31–38.

Oliphant, M. (1997). *Formal Approaches to Innate and Learned Communication: Laying the Foundations for Language*. Ph.D. thesis, Department of Cognitive Science, University of California, San Diego.

Paley, W. (1802). *Natural Theology, or Evidence of the Existence and Attributes of the Deity Collected from the Apperances of Nature*. Faulder, London.

Parker, G. A. (1974). Assessment strategy and the evolution of animal conflicts. *Journal of Theoretical Biology*, *47*, 223–243.

Partridge, L. (1980). Mate choice increases a component of offspring fitness in fruit flies. *Nature*, *283*, 290–291.

Pattee, H. H. (1995). Artificial life needs a real epistemology. In Morán, F., Moreno, A., Merelo, J. J., & Chacón, P. (Eds.), *Advances in Artificial Life: Third European Conference on Artificial Life (ECAL '95)*, Vol. 929 of *Lecture Notes in Artificial Intelligence*, pp. 23–38. Springer, Berlin.

Pomiankowski, A., Iwasa, Y., & Nee, S. (1991). The evolution of costly mate preferences I. Fisher and biased mutation. *Evolution*, *45*, 1422–1430.

Pomiankowski, A., & Møller, A. P. (1995). A resolution of the lek paradox. *Proceedings of the Royal Society of London: Biological Sciences*, *260*, 21–29.

Popper, K. R. (1968). *The Logic of Scientific Discovery* (Revised edition). Hutchinson, London.

Poundstone, W. (1985). *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge*. William Morrow, New York.

Prem, E. (1997). Epistemic autonomy in models of living systems. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 2–9. MIT Press / Bradford Books, Cambridge, MA.

Proctor, H. C. (1991). Courtship in the water mite *Neumania papillator*: Males capitalize on female adaptations for predation. *Animal Behaviour*, *42*, 589–598.

Proctor, H. C. (1993). Sensory exploitation and the evolution of male mating behaviour: A cladistic test using water mites (Acari: *Parasitengona*). *Animal Behaviour*, *44*, 745–752.

Prusinkiewicz, P. (1994). Visual models of morphogenesis. *Artificial Life*, *1*(1/2), 61–74.

Putnam, H. (1960). Minds and machines. In *Mind, Language, and Reality. Philosophical Papers, Volume Two*, pp. 362–385. Cambridge University Press, Cambridge. Collection published 1975.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, *60*(1), 20–43.

Ray, T. S. (1994). An evolutionary approach to synthetic biology: Zen and the art of creating life. *Artificial Life*, *1*(1/2), 179–209.

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In Ortony, A. (Ed.), *Metaphor and Thought*. Cambridge University Press, Cambridge.

Reynolds, C. W. (1987). Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics*, *21*(4), 25–34. SIGGRAPH '87 Conference Proceedings.

Reynolds, V., & Reynolds, F. (1965). Chimpanzees of the Budongo Forest. In DeVore, B. I. (Ed.), *Primate Behavior: Field Studies of Monkeys and Apes*, pp. 368–424. Holt, Rinehart and Winston, New York.

Riechert, S. E. (1982). Spider interaction strategies: Communication vs. coercion. In Witt, P. N., & Rovner, J. (Eds.), *Spider Communication: Mechanisms and Ecological Significance*, pp. 281–315. Princeton University Press, Princeton, NJ.

Riechert, S. E. (1998). Game theory and animal contests. In Dugatkin, L. A., & Reeve, H. K. (Eds.), *Game Theory and Animal Behavior*, pp. 64–93. Oxford University Press, New York.

Robbins, P. (1994). The effect of parasitism on the evolution of a communication protocol in an artificial life simulation. In Cliff, D., Husbands, P., Meyer, J.-A., & Wilson, S. W. (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pp. 431–437. MIT Press / Bradford Books, Cambridge, MA.

Ryan, M. J. (1985). *The Túngara frog: A study in sexual selection and communication*. University of Chicago Press, Chicago.

Ryan, M. J. (1988). Constraints and patterns in the evolution of anuran acoustic communication. In Fritzsch, B., Ryan, M. J., Wilczynski, W., Hetherington, T. E., & Walkowiak, W. (Eds.), *The Evolution of the Amphibian Auditory System*, pp. 637–678. John Wiley & Sons, New York.

Ryan, M. J. (1990). Sexual selection, sensory systems and sensory exploitation. *Oxford Surveys in Evolutionary Biology*, *7*, 157–195.

Ryan, M. J., & Rand, A. S. (1993). Sexual selection and signal evolution: the ghost of biases past. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *340*, 187–195.

Saunders, G. M., & Pollack, J. B. (1996). The evolution of communication schemes over continuous channels. In Maes, P., Matarić, M., Meyer, J.-A., Pollack, J., & Wilson, S. W. (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pp. 580–589. MIT Press / Bradford Books, Cambridge, MA.

Selous, E. (1901). *Bird Watching*. Constable, London.

Selous, E. (1933). *Evolution of Habit in Birds*. Constable, London.

Seyfarth, R., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science*, *210*, 801–803.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Sherman, P. W. (1977). Nepotism and the evolution of alarm calls. *Science*, *197*, 1246–1253.

Simon, H. A. (1981). *The Sciences of the Artificial* (Second edition). MIT Press, Cambridge, MA.

Smithers, T. (1994). On why better robots make it harder. In Cliff, D., Husbands, P., Meyer, J.-A., & Wilson, S. W. (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pp. 64–72. MIT Press / Bradford Books, Cambridge, MA.

Sober, E. (1993). *Philosophy of Biology*. Oxford University Press, Oxford.

Sober, E. (1996). Learning from functionalism—prospects for strong artificial life. In Boden, M. A. (Ed.), *The Philosophy of Artificial Life*, pp. 361–378. Oxford University Press, Oxford.

Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, *2*(3), 319–332.

Steels, L. (1996a). Emergent adaptive lexicons. In Maes, P., Matarić, M., Meyer, J.-A., Pollack, J., & Wilson, S. W. (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pp. 562–567. MIT Press / Bradford Books, Cambridge, MA.

Steels, L. (1996b). Self-organising vocabularies. In Langton, C. G., & Shimohara, K. (Eds.), *Artificial Life V*. MIT Press, Cambridge, MA.

Steels, L., & Vogt, P. (1997). Grounding adaptive language games in robotic agents. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 474–482. MIT Press / Bradford Books, Cambridge, MA.

Stokes, A. W. (1962a). Agonistic behaviour among Blue Tits at a winter feeding station. *Behaviour*, *19*, 118–138.

Stokes, A. W. (1962b). The comparative ethology of Great, Blue, Marsh and Coal Tits at a winter feeding station. *Behaviour*, *19*, 208–218.

Sugiyama, Y. (1969). Social behavior of chimpanzees in the Budongo Forest, Uganda. *Primates*, *10*(3, 4), 197–225.

Taylor, C., & Jefferson, D. (1994). Artificial life as a tool for biological inquiry. *Artificial Life*, *1*(1/2), 1–13.

Tesfatsion, L. (1997). A trade network game with endogenous partner selection. In Amman, H., et al. (Eds.), *Computational Approaches to Economic Problems*, pp. 249–269. Kluwer.

Tinbergen, N. (1952). "Derived" activities; their causation, biological significance, origin, and emancipation during evolution. *Quarterly Review of Biology*, *27*(1), 1–32.

Tinbergen, N. (1953). *The Herring Gull's World*. Collins, London.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*, 410–433.

Tinbergen, N. (1964). The evolution of signalling devices. In Etkin, W. (Ed.), *Social Behavior and Organization Among Vertebrates*, pp. 206–230. University of Chicago Press, Chicago.

Todd, P. M., & Miller, G. F. (1995). The role of mate choice in biocomputation: Sexual selection as a process of search, optimization, and diversification. In Banzof, W., & Eeckman, F. H. (Eds.), *Evolution and biocomputation: Computational models of evolution*, Vol. 899 of *Lecture Notes in Computer Science*, pp. 169–204. Springer-Verlag, New York.

Toquenaga, Y., Kajitani, I., & Hoshino, T. (1994). Egrets of a feather flock together. *Artificial Life*, *1*(4), 391–411.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.

Trivers, R. L. (1974). Parent-offspring conflict. *American Zoologist*, *14*, 249–264.

van Rhijn, J. G., & Vodegel, R. (1980). Being honest about one's intentions: An evolutionary stable strategy for animal conflicts. *Journal of Theoretical Biology*, *85*, 623–641.

Veblen, T. (1899). The theory of the leisure class. In Lerner, M. (Ed.), *The Portable Veblen*, pp. 53–214. Viking, New York. Collection published 1948.

von Frisch, K. (1967). *The Dance Language and Orientation of Bees*. Belknap Press / Harvard University Press, Cambridge, MA.

von Neumann, J., & Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

von Uexküll, J. (1928). *Theoretische Biologie*. Suhrkamp, Frankfurt.

Vriend, N. J. (1995). Self-organization of markets: An example of a computational approach. *Computational Economics*, *8*, 205–231.

Waas, J. R. (1991). Do little blue penguins signal their intentions during aggressive interactions with strangers? *Animal Behaviour*, *41*, 375–382.

Wade, M. J. (1978). A critical review of the models of group selection. *Quarterly Review of Biology*, *53*, 101–114.

Webb, B. (1994). Robotic experiments in cricket phonotaxis. In Cliff, D., Husbands, P., Meyer, J.-A., & Wilson, S. W. (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pp. 45–54. MIT Press / Bradford Books, Cambridge, MA.

Wedell, N. (1994). Variation in nuptial gift quality in bush crickets (Orthoptera: Tettigoniidea). *Behavioural Ecology*, *5*, 418–425.

Werner, G. M. (1996). Why the peacock's tail is so short: Limits to sexual selection. In Langton, C. G., & Shimohara, K. (Eds.), *Artificial Life V*, pp. 85–91. MIT Press, Cambridge, MA.

Werner, G. M., & Dyer, M. G. (1991). Evolution of communication in artificial organisms. In Langton, C. G., Taylor, C., Farmer, J. D., & Rasmussen, S. (Eds.), *Artificial Life II*. Addison-Wesley, Redwood City, CA.

Werner, G. M., & Dyer, M. G. (1993). Evolution of herding behavior in artificial animals. In Meyer, J.-A., Roitblat, H. L., & Wilson, S. W. (Eds.), *From Animals to Animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*. MIT Press / Bradford Books, Cambridge, MA.

Werner, G. M., & Todd, P. M. (1997). Too many love songs: Sexual selection and the evolution of communication. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97)*, pp. 434–443. MIT Press / Bradford Books, Cambridge, MA.

Wheeler, M., & de Bourcier, P. (1995). How not to murder your neighbor: Using synthetic behavioral ecology to study aggressive signaling. *Adaptive Behavior*, *3*(3), 273–309.

Wiley, R. H. (1973). Territoriality and non-random mating in sage grouse (*Centrocercus urophasianus*). *Animal Behaviour Monographs*, *6*, 85–169.

Wiley, R. H. (1983). The evolution of communication: Information and manipulation. In Halliday, T. R., & Slater, P. J. B. (Eds.), *Communication*, Vol. 2 of *Animal Behaviour*, pp. 156–189. Blackwell, Oxford.

Williams, G. C. (1966). *Adaptation and Natural Selection*. Princeton University Press, Princeton, NJ.

Williams, G. C. (1975). *Sex and Evolution*. Princeton University Press, Princeton, NJ.

Wilson, D. S. (1975). A general theory of group selection. *Proceedings of the National Academy of Sciences, USA*, *72*, 143–146.

Wilson, D. S. (1980). *The Natural Selection of Populations and Communities*. Benajmin-Cummings, Menlo Park, CA.

Wilson, E. O. (1975). *Sociobiology: The New Synthesis*. Belknap Press / Harvard University Press, Cambridge, MA.

Wilson, S. W. (1995). Classifier fitness based on accuracy. *Evolutionary Computation*, *3*(2), 149–175.

Wright, L. (1994). Functions. In Sober, E. (Ed.), *Conceptual Issues in Evolutionary Biology*, pp. 27–48. MIT Press / Bradford Books, Cambridge, MA.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*(2), 97–158.

Wynne-Edwards, V. C. (1962). *Animal Dispersion in Relation to Social Behaviour*. Oliver and Boyd, Edinburgh.

Yamauchi, B. M., & Beer, R. D. (1994). Sequential behavior and learning in evolved dynamical neural networks. *Adaptive Behavior*, *2*(3), 219–246.

Yanco, H., & Stein, L. A. (1993). An adaptive communication protocol for cooperating mobile robots. In Meyer, J.-A., Roitblat, H. L., & Wilson, S. W. (Eds.), *From Animals to Animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, pp. 478–485. MIT Press / Bradford Books, Cambridge, MA.

Zaera, N., Cliff, D., & Bruten, J. (1996). (Not) evolving collective behaviors in synthetic fish. In Maes, P., Matarić, M., Meyer, J.-A., Pollack, J., & Wilson, S. W. (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pp. 635–644. MIT Press / Bradford Books, Cambridge, MA.

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, *53*, 205–214.

Zahavi, A. (1977). The cost of honesty (further remarks on the handicap principle). *Journal of Theoretical Biology*, *67*, 603–605.

Zahavi, A. (1987). The theory of signal selection and some of its implications. In Delfino, V. P. (Ed.), *International Symposium on Biological Evolution*, pp. 305–327.

Zahavi, A. (1991). On the definition of sexual selection, Fisher's model, and the evolution of waste and of signals in general. *Animal Behaviour*, *42*, 501–503.

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford University Press, Oxford.

Appendix A

Cooperative and competitive communication: derivations of game-theoretic results

A.1 The simple signalling game

A.1.1 Expected payoffs in an honest and trusting population

In looking at the simple signalling game presented in chapter 7, we are firstly concerned with the honest and trusting strategy (NS/Sig, Neg/Pos), which specifies that signallers will signal only in the high state, and receivers will respond positively only to a signal. In a population of players using this strategy, we can ask what the expected payoff per game would be. There are four situations that need to be taken into account: a player is equally likely to play the signalling or the receiving role in any one game, and this is crossed with the fact that the hidden environmental state is equally likely to be high or low. Table A.1 shows the payoff for a player in each of these cases. The average expected payoff for a player in an honest and trusting population is therefore the average of the expressions in the four cells of table A.1. This is how the value of $\frac{P_S - C_S + P_R - C_R}{4}$ given in section 7.2.1 was derived.

A.1.2 Conditions for evolutionary stability of the honest and trusting strategy

Evolutionary stability depends on a strategy being uninvadable because it is the best response to itself. Table A.2 shows the returns expected for each of the sixteen possible strategies in the simple game, assuming that the background population is playing the honest and trusting strategy. For simplicity, the necessary division by four has been factored out of each expression, and so the table is really showing the expected return per four games. The strategies are labelled with binary digits according to the scheme described in table 7.2; note that the code for the honest and trusting strategy itself is 0101.

For the honest and trusting strategy to be an ESS requires that the entry for 0101 in the table be greater than or equal to the payoff for any other strategy (and if it is equal to some other payoff value then that strategy must not do as well against itself as does the honest and trusting strategy). However, we do not need to solve fifteen inequalities in parallel in order to derive the conditions for evolutionary stability. By inspecting table A.2 we can see that the payoff P_S never occurs without

	Signaller	Receiver
Low state	0	0
High state	$P_S - C_S$	$P_R - C_R$

Table A.1: Payoffs for players in various cases, assuming a population that has fixated on the honest and trusting strategy.

	Expected payoff per 4 games
Strategy: 0000	0
0001	$P_R - C_R$
0010	$-C_R$
0011	$P_R - 2C_R$
0100	$P_S - C_S$
0101	$P_S - C_S + P_R - C_R$
0110	$P_S - C_S - C_R$
0111	$P_S - C_S + P_R - 2C_R$
1000	$-C_S$
1001	$-C_S + P_R - C_R$
1010	$-C_S - C_R$
1011	$-C_S + P_R - 2C_R$
1100	$P_S - 2C_S$
1101	$P_S - 2C_S + P_R - C_R$
1110	$P_S - 2C_S - C_R$
1111	$P_S - 2C_S + P_R - 2C_R$

Table A.2: Expected payoffs for all sixteen possible strategies against a background population that plays the honest and trusting strategy 0101.

	Expected payoff per 4 games
Strategy: 0000	P_S
0001	P_S
0010	$P_S + P_R - 2C_R$
0011	$P_S + P_R - 2C_R$
0100	$P_S - C_S$
0101	$P_S - C_S$
0110	$P_S - C_S + P_R - 2C_R$
0111	$P_S - C_S + P_R - 2C_R$
1000	$P_S - C_S$
1001	$P_S - C_S$
1010	$P_S - C_S + P_R - 2C_R$
1011	$P_S - C_S + P_R - 2C_R$
1100	$P_S - 2C_S$
1101	$P_S - 2C_S$
1110	$P_S - 2C_S + P_R - 2C_R$
1111	$P_S - 2C_S + P_R - 2C_R$

Table A.3: Expected payoffs for all sixteen possible strategies against a background population that plays the blind optimism strategy 0011.

the cost C_S, and similarly that the payoff P_R never occurs without the cost C_R. Furthermore, we know that while the payoffs P_S and P_R might sometimes be negative, the costs C_S and C_R are meant to represent energy expenditure or something similar and so cannot sensibly be negative. Finally, the zero payoff to the 0000 strategy sets up a minimum payoff level that honest-and-trusting must beat if it is to be an ESS. Thus, to guarantee that the entry for 0101 will be the highest in the table, we must stipulate that the payoffs P_S and P_R are both positive, and in each case larger than the respective cost values. These values would mean that a population of honest and trusting players could not be invaded by any mutant strategy. This is how the conditions $P_S > C_S > 0$ and $P_R > C_R > 0$ were derived in section 7.2.1.

A.1.3 Other possible ESSs

Showing that the honest and trusting strategy can be an ESS under certain conditions in no way implies that it is the only possible ESS in the game, or even that it is the only possible ESS under those conditions. It turns out that the strategy involving never signalling and always responding positively, i.e., 0011, can be an ESS of sorts (this strategy is referred to as "blind optimism" in chapter 7). It can always be invaded by the similar strategy 0010—which specifies never signalling and responding positively if no signal is given—and vice versa. However, the two strategies lead to the same behaviour when all players are using one or the other or a mixture of the two, and so we can consider their mutual uninvadability to constitute a joint ESS.

Table A.3 shows the expected payoffs for all sixteen possible strategies competing against a background population of blind optimists, i.e., players using the 0011 strategy. To show that blind optimism can be an ESS, we have to find the conditions under which the entries for 0011 and 0010 (i.e., $P_S + P_R - 2C_R$ in each case) will be the highest in the table. We can do this by considering which other payoff might be competitive. Firstly, we note that of the payoffs P_S, $P_S - C_S$, and $P_S - 2C_S$, the first of these must always be higher than the others, because C_S is an energy cost and therefore always positive. The payoff for blind optimism will be greater than P_S if $P_R - 2C_R > 0$, i.e., if $P_R > 2C_R$. Similarly, of the other two payoff values present in the table—$P_S - C_S + P_R - 2C_R$ and $P_S - 2C_S + P_R - 2C_R$—the payoff for blind optimists will always be higher as long as C_R is positive. Thus the important condition for the evolutionary stability of blind optimism is that $P_R > 2C_R$.

A minor complication is introduced if we consider what happens for a population dominated by the similar strategy 0010: it may be important, in such a context, not to give a signal in the high state—otherwise any 0010 receiver will fail to respond and the player will miss out on the P_S payoff. The equivalent of table A.3 for the 0010 case thus looks slightly different. This introduces a further ESS condition that $P_S > -C_S$, because if P_S happens to be negative then it might be worth paying the cost of signalling precisely in order to miss out on P_S.

The other two conditions for the stability of blind optimism given in section 7.2.1 follow from the discussion above. The requirement that $C_S > 0$ is simply making explicit the assumption that signalling has a positive cost. It is necessary that $P_R > 2C_R > 0$ because the payoff to receivers must be high enough to compensate for the costs of constant positive responses.

A.2 The variable-signal-cost game

There are 72 possible strategies in the more complex signalling game described in section 7.3.1. Showing that various strategies qualify (or fail to qualify) as ESSs by using complete strategy tables will therefore be avoided.

A.2.1 Stability of soft-signalling honest-and-trusting strategies

The first claim made in section 7.3.1 is that the strategies involving not signalling in the low state, signalling softly in the high state, and responding to soft signals—i.e., (NS/Soft, Neg/Pos/Pos) and (NS/Soft, Neg/Pos/Neg)—constitute a joint ESS. We can refer to these strategies as soft-signalling honest-and-trusting strategies. Either one played against itself or the other will result in an expected return of $P_S - C_S + P_R - C_R$ per four games—the logic of table A.1 still applies. If we consider a population dominated by some mixture of these two strategies, we can ask whether any mutant strategies would be able to invade.

Table A.4 shows the *change* in the expected payoff for various mutant strategies that might arise in a population of players using either of the soft-signalling honest-and-trusting strategies. Progressing through the table line by line, we can see that there is nothing to encourage a mutant that starts giving soft or loud signals (as opposed to no signal at all) in the low state. They will simply lower their expected payoff by $-C_S$ or $-2C_S$ respectively, and, given the assumption that C_S is a positive cost value, any such mutants will pose no threat of invasion.

A mutant that starts giving no signal in the high state, in contrast to the soft signals given by the rest of the population, will sacrifice $-P_S$, i.e., it will miss out on the payoff to signallers, but it will gain $+C_S$, i.e., it will not have to pay the cost of signalling. This mutation would be

	Change in expected payoff per 4 games
Signal in low state: soft	$-C_S$
loud	$-2C_S$
Signal in high state: none	$-P_S + C_S$
loud	$-P_S - C_S$ or $-C_S$
Response to no signal: positive	$-C_R$
Response to soft signal: negative	$-P_R + C_R$
Response to loud signal: negative	0
positive	0

Table A.4: Change in expected payoffs for various one-point mutant strategies playing against a background population of soft-signalling honest-and-trusting players in the variable-signal-cost game.

an improvement on the background population if C_S were greater than P_S. Thus we come to our first condition for the evolutionary stability of the soft-signalling honest-and-trusting strategy: that $P_S > C_S > 0$. (The second part of this condition, namely that $C_S > 0$, is simply making explicit the assumption that C_S is a positive cost value.)

A mutant that makes a loud signal in the high state may or may not miss out on the payoff to signallers (P_S) depending on whether the background population is playing (NS/Soft, Neg/Pos/Pos) or (NS/Soft, Neg/Pos/Neg). However, such a mutant will always be worse off by $-C_S$. Although in principle it might invade the population if P_S was strongly negative, in practice the possibility of this mutation does not impose any additional conditions on the evolutionary stability of the honest and trusting strategy.

Mutants that respond positively when no signal is given will make themselves worse off by $-C_R$. They therefore offer no threat of invasion.

Mutants that respond negatively to the soft signal, on the other hand, will improve their lot if $C_R > P_R$. This makes sense: when the cost of responding is greater than the payoff, one can do better by not responding at all. Thus we come to our second condition for the stability of the honest-and-trusting strategy: that $P_R > C_R > 0$, otherwise negative-response mutants will be able to invade.

Finally, we can see from the last two lines of table A.4 that a mutation which changes a player's response to the loud signal will have no effect on its expected payoff. This is because in a soft-signalling population, loud signals will never be heard. Selectively neutral drift between (NS/Soft, Neg/Pos/Pos) and (NS/Soft, Neg/Pos/Neg) is possible because of this fact.

The argument given above shows that no single-point mutation will threaten to invade a soft-signalling honest-and-trusting population as long as the payoffs to signallers and receivers are larger than their respective costs. The possibility of multiple-point mutations need not be dealt

	Change in expected payoff per 4 games
Signal in low state: soft	$-C_S$
loud	$-2C_S$
Signal in high state: none	$-P_S+2C_S$
soft	$+C_S$
Response to no signal: positive	$-C_R$
Response to soft signal: negative	0
Response to loud signal: negative	$-P_R+C_R$

Table A.5: Change in expected payoffs for various one-point mutant strategies playing against a background population of loud-signalling players in the variable-signal-cost game.

with explicitly, as mutations that change more than one aspect of a player's strategy will be additive in their effects.

A.2.2 Costly signalling not an ESS

The second claim made in section 7.3.1 is that none of the strategies involving costly signalling—i.e., the use of the loud signal to denote the high state—can be an ESS. We will first consider the strategy (NS/Loud, Neg/Pos/Pos) as an example. Players in a population dominated by this strategy can expect an average payoff of $P_S - 2C_S + P_R - C_R$ per four games. Table A.5 shows the change in the expected payoff for all possible single-point mutations that might arise in a population playing this strategy.

The most important aspect of the table, for our current purposes, is that the expected payoff to a mutant that began signalling softly in the high state (instead of loudly) would in fact improve by C_S. Such a mutant would therefore invade; because C_S is always positive, there are no circumstances under which the loud-signalling strategy (NS/Loud, Neg/Pos/Pos) can be an ESS.

The reason that shifting to a soft signalling strategy results in an increased expected payoff is because the background population happen to be ready to respond positively to soft signals. Of course, with no soft signals in general use, there is no selective pressure on the locus specifying what response to make to a soft signal; note that the change in expected payoff for a mutant that starts responding *negatively* to soft signals is zero. This point is the basis for the general argument that none of the costly signalling strategies in the variable-cost game can be an ESS: a uniform population must always settle on one signal for the high state and another for the low state. This takes selective pressure off the loci for responses to any other available signals, because such signals are never heard. But if the genetically specified responses to unheard signals can drift freely, then sooner or later a mutant that uses a cheaper signalling strategy will be able to invade. In the example given, (NS/Loud, Neg/Pos/Pos) can be invaded by (NS/Soft, Neg/Pos/Pos), a cheaper signalling strategy. The strategy (NS/Loud, Neg/Neg/Pos) specifies responding negatively to soft signals, and so it cannot be invaded in the same way. However, because soft signals are not used, there is nothing to stop the population drifting to the invadable (NS/Loud, Neg/Pos/Pos). Eventually the cheapest pair of signals—in this case No Signal and Soft—will come to be used for the low and high states respectively. Note that this argument stands even if any number of additional signals of varying cost levels are introduced to the game.