

## 8

---

# Signaling Games

This above all: to thine own self be true, And it must follow,  
as the night the day, Thou canst not then be false to any man.

Shakespeare

### 8.1 Signaling as a Coevolutionary Process

A Thompson's gazelle who spots a cheetah, instead of fleeing, will often "stott," which involves an 18-inch vertical jump, with legs stiff and white rump patch fully displayed to the predator. The only plausible explanation for this behavior (Alcock 1993) is that the gazelle is signaling the cheetah that it would be a waste of both their times and energies for the cheetah to chase the gazelle, because the gazelle is obviously very fit. Of course, if the cheetah could not understand this signal, it would be a waste of time and energy for the gazelle to emit it. Also, if the signal could be easily falsified, and the ability to stott had nothing to do with the probability of being caught, cheetahs would never have evolved to heed the signal in the first place.<sup>1</sup>

A *signal* is a special sort of physical interaction between two agents. Like other physical interactions, a signal changes the physical constitution of the agents involved. But unlike interactions among nonliving objects, or between a nonliving object and a living agent, a signal is the product of a *strategic dynamic* between sender and receiver, each of whom is pursuing distinct but interrelated objectives. Moreover, a signal is a specific *type* of strategic physical interaction, one in which the content of the interaction is determined by the sender, and it changes the receiver's behavior by altering the way the receiver evaluates alternative actions.

<sup>1</sup>For a review of evidence for costly signaling in birds and fish in the form of colorful displays that indicate health and vigor, see Olson and Owens 1998. On the more general topic of costly signaling, see Zahavi and Zahavi 1997 and section 8.4.

The most important fact about a signal is that it is generally the result of a *coevolutionary process between senders and receivers* in which both benefit from its use. For if a signal is costly to emit (and if its use has been stable over time), then the signal is most likely both *beneficial to the sender* and *worthy of belief for the receiver*; a sender is better off sending that signal rather than none, or some other, and a receiver is better off acting on it the way receivers traditionally have, rather than ignoring it or acting otherwise. The reason is obvious: if the receiver were *not* better off acting this way, a mutant who ignored (or acted otherwise on) the signal would be more fit than the current population of receivers, and would therefore increase its frequency in the population. Ultimately, so many receivers would ignore (or act otherwise on) the signal that, being costly to the sender, it would not be worth sending unless, of course, the “otherwise” were also beneficial to the sender.

Signaling systems are not always in equilibrium, and potentially beneficial mutations need not occur. Moreover, human beings are especially adept both at dissimulating (emitting “false” signals) and detecting such dissimulation (Cosmides and Tooby 1992). However, human beings are disposed to taking the signals around them at face value unless there are good reasons for doing otherwise (Gilbert 1991). The treatment of signals as emerging from a coevolutionary process and persisting as a Nash equilibrium of the appropriate game, is the starting point for a theory of signaling.

## 8.2 A Generic Signaling Game

In signaling games, player 1 has a “type” that is revealed to player 2 via a special “signal,” to which player 2 responds by choosing an “action,” the payoffs to the two players being a function of player 1’s type and signal and player 2’s action. Thus, the stage game that played so prominent a role in the general Bayesian game framework collapses, in the case of signaling games, to a pair of payoff functions.

Specifically, there are three players Sender, Receiver, and Nature. Nature begins by choosing from a set  $T$  of possible *types* or *states of affairs*, choosing  $t \in T$  with probability  $\rho(t)$ . Sender observes  $t$  but Receiver does not. Sender then transmits a *signal*  $s \in S$  to Receiver, who uses this signal to choose an *action*  $a \in A$ . The payoffs to the two players are  $u(t, s, a)$  and  $v(t, s, a)$ , respectively. A pure strategy for Sender is thus a function  $f : T \rightarrow S$ , where  $s = f(t)$  is the signal sent when Nature reveals type  $t$ ,

and a pure strategy for Receiver is a function  $g : S \rightarrow A$ , where  $a = g(s)$  is the action taken when Receiver receives signal  $s$ . A mixed strategy for Sender is a probability distribution  $P_S(s; t)$  over  $S$  for each  $t \in T$ , and a mixed strategy for Receiver is a probability distribution  $p_R(a; s)$  over  $A$  for each signal  $s$  received. A Nash equilibrium for the game is thus a pair of probability distributions  $(P_S(\cdot; t), p_R(\cdot, s))$  for each pair  $\{(t, s) | t \in T, s \in S\}$  such that each agent uses a best response to the other, given the probability distribution  $\rho(t)$  used by Nature to choose the type of Sender.

We say a signal  $s \in S$  is *along the path of play*, given the strategy profile  $(P_S(\cdot; t), p_R(\cdot; s))$ , if there is a strictly positive probability that Sender will transmit  $s$ , that is, if

$$\sum_{t \in T} \rho(t) P_S(s; t) > 0.$$

If a signal is not along the path of play, we say it is *off the path of play*. If  $s$  is along the path of play, then a best response for Receiver maximizes Receiver's expected return, with a probability distribution over  $T$  given by

$$P[t|s] = \frac{P_S(s; t)\rho(t)}{\sum_{t' \in T} P_S(s; t')\rho(t')}.$$

We thus require of  $P_S$  and  $p_R$  that

- a. For every state  $t \in T$ , and all signals  $s' \in S$  such that  $P_S(s'; t) > 0$ ,  $s'$  maximizes

$$\sum_{a \in A} u(t, s', a) p_R(a; s)$$

over all  $s \in S$ ; that is, Sender chooses a best response to Receiver's pattern of reacting to S's signals;

- b. For every signal  $s \in S$  along the path of play, and all actions  $a' \in A$  such that  $p_R(a'; s) > 0$ ,  $a'$  maximizes

$$\sum_{t \in T} v(t, s, a) P[t|s]$$

over all  $a \in A$ ; that is, Receiver chooses a best response to Sender's signal.

- c. If a signal  $s \in S$  is not along the path of play, we may choose  $P[t|s]$  arbitrarily such that (b) still holds. In other words, Receiver may respond arbitrarily to a signal that is never sent, provided this does not induce Sender to send the signal.

### 8.3 Sex and Piety: The Darwin-Fisher Model

In most species, females invest considerably more in raising their offspring than do males; for instance, they produce a few large eggs as opposed to the male's millions of small sperm. So, female fitness depends more on the *quality* of inseminations, whereas male fitness depends more on the *quantity* of inseminations (§6.26). Hence, in most species there is an *excess demand for copulations* on the part of males, for whom procreation is very cheap, and therefore there is a *nonclearing market for copulations*, with the males on the long side of the market (§9.13). In some species this imbalance leads to violent fights among males (dissipating the rent associated with achieving a copulation), with the winners securing the scarce copulations. But in many species, *female choice* plays a central role, and males succeed by being attractive rather than ferocious.

What criteria might females use to choose mates? We would expect females to seek mates whose appearance indicates they have genes that will enhance the survival value of their offspring. This is indeed broadly correct. But in many cases, with prominent examples among insects, fish, birds, and mammals, females appear to have *arbitrary prejudices* for dramatic, ornamental, and colorful displays even when such accoutrements clearly reduce male survival chances; for instance, the plumage of the bird of paradise, the elaborate structures and displays of the male bowerbird, and the stunning coloration of the male guppy. Darwin speculated that such characteristics improve the mating chances of males at the expense of the average fitness of the species. The great biologist R. A. Fisher (1915) offered the first genetic analysis of the process, suggesting that an arbitrary female preference for a trait would enhance the fitness of males with that trait, and hence the fitness of females who pass that trait to their male offspring, so the genetic predisposition for males to exhibit such a trait could become common in a species. Other analytical models of sexual selection, called *Fisher's runaway process* include Lande (1981), Kirkpatrick (1982), Pomiankowski (1987), and Bulmer (1989). We will follow Pomiankowski (1987), who showed that *as long as females incur no cost for being choosy, the Darwin-Fisher sexual selection process works, but even with a slight cost of being choosy, costly ornamentation cannot persist in equilibrium*.

We shall model runaway selection in a way that is not dependent on the genetics of the process, so it applies to cultural as well as genetic evolution. Consider a community in which there are an equal number of males and

females, and there is a cultural trait that we will call *pious fasting*. Although both men and women can have this trait, only men act on it, leading to their death prior to mating with probability  $u > 0$ . However, both men and women pass the trait to their children through family socialization. Suppose a fraction  $t$  of the population have the pious-fasting trait.

Suppose there is another cultural trait, a *religious preference for pious fasting*, which we call being “choosy” for short. Again, both men and women can carry the choosy trait and pass it on to their children, but only women can act on it, by choosing mates who are pious fasters at rate  $a > 1$  times that of otherwise equally desirable males. However, there may be a cost of exercising this preference, because with probability  $k \geq 0$  a choosy woman may fail to mate. Suppose a fraction  $p$  of community members bears the religious preference for pious fasters.

We assume parents transmit their values to their offspring in proportion to their own values; for instance, if one parent has the pious-fasting trait and the other does not, then half their children will have the trait. Males who are pious fasters then exercise their beliefs, after which females choose their mates, and a new generation of young adults is raised (the older generation moves to Florida to retire).

Suppose there are  $n$  young adult males and an equal number of young adult females. Let  $x_{tp}$  be the fraction of young adults who are “choosy fasters,”  $x_{-p}$  the fraction of “choosy nonfasters,”  $x_{t-}$  the fraction of “nonchoosy fasters,” and  $x_{--}$  the fraction of “nonchoosy nonfasters.” Note that  $t = x_{tp} + x_{t-}$  and  $p = x_{tp} + x_{-p}$ . If there is no correlation between the two traits, we would have  $x_{tp} = tp$ ,  $x_{t-} = t(1 - p)$ , and so on. But we cannot assume this, so we write  $x_{tp} = tp + d$ , where  $d$  (which biologists call *linkage disequilibrium*) can be either positive or negative. It is easy to check that we then have

$$\begin{aligned}x_{tp} &= tp + d \\x_{t-} &= t(1 - p) - d \\x_{-p} &= (1 - t)p - d \\x_{--} &= (1 - t)(1 - p) + d.\end{aligned}$$

Although male and female young adults have equal fractions of each trait because their parents pass on traits equally to both pious fasting and mate choosing can lead to unequal frequencies in the “breeding pool” of parents in the next generation. By assumption, a fraction  $k$  of choosy females do

not make it to the breeding pool, so if  $t^f$  is the fraction of pious-faster females in the breeding pool, then

$$t^f = \frac{t - kx_{tp}}{1 - kp},$$

where the denominator is the fraction of females in the breeding pool, and the numerator is the fraction of pious-faster females in the breeding pool. Similarly, if  $p^f$  is the fraction of choosy females in the breeding pool, then

$$p^f = \frac{p(1 - k)}{1 - kp},$$

where the numerator is the fraction of choosy females in the breeding pool.

We now do the corresponding calculations for males. Let  $t^m$  be the fraction of pious-faster males, and  $p^m$  the fraction of choosy males in the breeding pool, after the losses associated with pious fasting are taken into account. We have

$$t^m = \frac{t(1 - u)}{1 - ut},$$

where the denominator is the fraction of males, and the numerator is the fraction of pious-faster males in the breeding pool. Similarly,

$$p^m = \frac{p - ux_{tp}}{1 - ut},$$

where the numerator is the fraction of choosy males in the breeding pool.

By assumption, all  $n^f = n(1 - kp)$  females in the breeding pool are equally fit. We normalize this fitness to 1. The fitnesses of pious and nonpious males in the breeding pool are, however, unequal. Suppose each female in the breeding pool mates once. There are then  $n^f(1 - p^f)$  nonchoosy females, so they mate with  $n^f(1 - p^f)(1 - t^m)$  nonpious males and  $n^f(1 - p^f)t^m$  pious males. There are also  $n^f p^f$  choosy females, who mate with  $n^f p^f(1 - t^m)/(1 - t^m + at^m)$  nonpious males and  $n^f p^f at^m/(1 - t^m + at^m)$  pious males (the numerators account for the  $a : 1$  preference for pious males, and the denominator is chosen so that the two terms add to  $n^f p^f$ ). If we write

$$r_- = (1 - p^f) + \frac{p^f}{1 - t^m + at^m}$$

$$r_t = (1 - p^f) + \frac{ap^f}{1 - t^m + at^m},$$

then the total number of matings of nonpious males is  $n^f(1 - t^m)r_-$ , and the total number of matings of pious males is  $n^f t^m r_t$ . The probability that a mated male is pious is therefore  $t^m r_t$ . Because the probability that a mated female is pious is  $t^f$ , and both parents contribute equally to the traits of their offspring, the fraction of pious traits in the next generation is  $(t^m r_t + t^f)/2$ . If we write  $\beta_t = t^m r_t - t$  and  $\beta_p = p^f - p$ , then the change  $\Delta t$  in the frequency of the pious trait can be written as

$$\Delta t = \frac{t^m r_t + t^f}{2} - t = \frac{1}{2} \left( \beta_t + \frac{d\beta_p}{p(1-p)} \right). \quad (8.1)$$

What about the change in  $p$  across generations? The fraction of mated, choosy females is simply  $p^f$ , because all females in the breeding pool mate. The number  $n^m$  of males in the breeding pool is  $n^m = n(1 - ut)$ , of which  $n x_{-p}$  are nonpious and choosy, whereas  $n(1 - u)x_{tp}$  are pious and choosy. Each nonpious male has  $n^f r_-/n^m$  offspring, and each pious male has  $n^f r_t/n^m$  offspring, so the total number of choosy male offspring per breeding female is just

$$p^{m'} = n x_{-p} r_- / n^m + n(1 - u)x_{tp} r_t / n^m.$$

A little algebraic manipulation shows that this can be written more simply as

$$p^{m'} = p + \frac{d\beta_t}{t(1-t)}.$$

Then the change  $\Delta p$  in the frequency of the choosy trait can be written as

$$\Delta p = \frac{p^{m'} + p^f}{2} - p = \frac{1}{2} \left( \beta_p + \frac{d\beta_t}{t(1-t)} \right). \quad (8.2)$$

Let us first investigate (8.1) and (8.2) when choosy females are not less fit, so  $k = 0$ . In this case,  $p^f = p$ , so  $\beta_p = 0$ . Therefore,  $\Delta t = \Delta p = 0$  exactly when  $\beta_t = 0$ . Solving this equation for  $t$ , we get

$$t = \frac{(a-1)p(1-u) - u}{u(a(1-u) - 1)}. \quad (8.3)$$

This shows that there is a range of values of  $p$  for which an equilibrium frequency of  $t$  exists. Checking the Jacobian of the right-hand sides of (8.1) and (8.2) (see section 11.7 if you do not know what this means) we

find that stability requires that the denominator of (8.3) be positive (do it as an exercise). Thus, the line of equilibria is upward sloping, and  $t$  goes from zero to one as  $p$  goes from  $u/(a - 1)(1 - u)$  to  $au/(a - 1)$  (you can check that this defines an interval contained in  $(0, 1)$  for  $0 < u < 1$  and  $a(1 - u) > 1$ ). This set of equilibria is shown in figure 8.1. This shows that the Darwin-Fisher sexual selection process is plausible, even though it lowers the average fitness of males in the community. In essence, the condition  $a(1 - u) > 1$  ensures that the benefit of sexual selection more than offsets the cost of the ornamental handicap.

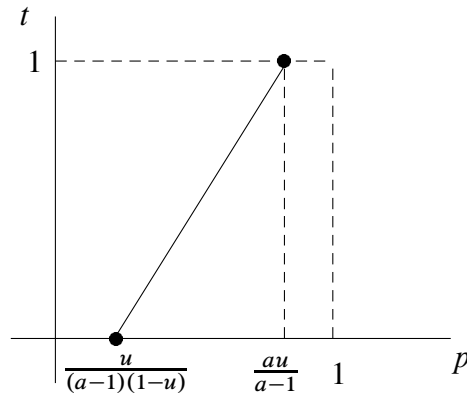


Figure 8.1. Equilibria in Darwin-Fisher sexual selection model when there is no selection against choosy females

Suppose, however,  $k > 0$ . If we then solve for  $\Delta t = \Delta p = 0$  in (8.1) and (8.2), we easily derive the equation

$$d^2 = t(1 - t)p(1 - p).$$

But  $t(1 - t)p(1 - p) = (x_{t-} + d)(x_{-p} + d)$ , which implies  $x_{t-} = x_{-p} = 0$ . But then, nonchoosy females must mate only with nonpious males, which is impossible so long as there is a positive fraction of pious males. We conclude that *when choosiness is costly to females, sexual selection cannot exist*. Because in most cases we can expect some positive search cost to be involved in favoring one type of male over another, we conclude that sexual selection probably does not occur in equilibrium in nature. Of course, random mutations could lead to a disequilibrium situation in which females prefer certain male traits, leading to increased fitness of males with those traits. But when the fitness costs of such choices kick in, choosy females will decline until equilibrium is restored.

#### 8.4 Biological Signals as Handicaps

Zahavi (1975), after close observation of avian behavior, proposed an alternative to the Darwin-Fisher sexual selection mechanism, a notion of costly signaling that he called the *handicap principle*. According to the handicap principle, a male who mounts an elaborate display is in fact signaling his good health and/or good genes, because an unhealthy or genetically unfit male lacks the resources to mount such a display. The idea was treated with skepticism for many years, because it proved difficult to model or empirically validate the process. This situation changed when Grafen (1990b) developed a simple analytical model of the handicap principle. Moreover, empirical evidence has grown in favor of the costly signaling approach to sexual selection, leading many to favor it over the Darwin-Fisher sexual selection model, especially in cases where female mate selection is costly.

Grafen's model is a special case of the generic signaling model presented in section 8.2. Suppose a male's type  $t \in [t_{\min}, \infty)$  is a measure of male vigor (e.g., resistance to parasites). Females do best by accurately determining  $t$ , because an overestimate of  $t$  might lead a female to mate when she should not, and an underestimate might lead her to pass up a suitable mate. If a male of type  $t$  signals his type as  $s = f(t)$ , and a female uses this signal to estimate the male's fitness as  $a = g(s)$ , then in an equilibrium with truthful signaling we will have  $a = t$ . We suppose that the male's fitness is  $u(t, s, a)$ , with  $u_t > 0$  (a male with higher  $t$  is more fit),  $u_s < 0$  (it is costly to signal a higher level of fitness), and  $u_a > 0$  (a male does better if a female thinks he's more fit). We assume the male's fitness function  $u(t, s, g(s))$  is such that a more vigorous male will signal a higher fitness; that is,  $ds/dt > 0$ . Given  $g(s)$ , a male of type  $t$  will then choose  $s$  to maximize  $U(s) = u(t, s, g(s))$ , which has first-order condition

$$U_s(s) = u_s(t, s, g(s)) + u_a(t, s, g(s)) \frac{dg}{ds} = 0. \quad (8.4)$$

If there is indeed truthful signaling, then this equation must hold for  $t = g(s)$ , giving us the differential equation

$$\frac{dg}{ds} = -\frac{u_s(g(s), s, g(s))}{u_a(g(s), s, g(s))}, \quad (8.5)$$

which, together with  $g(s_{\min}) = t_{\min}$ , uniquely determines  $g(s)$ . Because  $u_s < 0$  and  $u_a > 0$ , we have  $dg/ds > 0$ , as expected.

Differentiating the first-order condition (8.4) totally with respect to  $t$ , we find

$$U_{ss} \frac{ds}{dt} + U_{st} = 0.$$

Because  $U_{ss} < 0$  by the second-order condition for a maximum, and because  $ds/dt > 0$ , we must have  $U_{st} > 0$ . But we can write

$$\begin{aligned} U_{st} &= u_{st} + u_{at} g'(s) \\ &= \frac{u_{st} u_a(g(s), s, g(s)) - u_{at} u_s(g(s), s, g(s))}{u_a} > 0. \end{aligned}$$

Therefore,

$$\frac{d}{dt} \left[ \frac{u_s(t, s, g(s))}{u_a(t, s, g(s))} \right] = \frac{U_{st}}{u_a} > 0. \quad (8.6)$$

We can now rewrite (8.4) as

$$u_a(t, s, g(s)) \left[ \frac{u_s(t, s, g(s))}{u_a(t, s, g(s))} + g'(s) \right] = 0. \quad (8.7)$$

Because the fraction in this expression is increasing in  $t$ , and the expression is zero when  $t = g(s)$ , this shows  $s = g^{-1}(t)$  is a local maximum, so the male maximizes fitness by truthfully reporting  $s = g^{-1}(t)$ , at least locally.

For an example of the handicap principle, suppose  $u(t, s, a) = a^r t^s$ ,  $0 < t < 1$ , so (8.5) becomes  $g'/g = -(1/r) \ln g$ , which has solution  $\ln g = ce^{-s/r}$ . If we use  $g(s_{\min}) = t_{\min}$  this gives

$$g(s) = t_{\min} e^{-\frac{s-s_{\min}}{r}}$$

and

$$f(t) = s_{\min} - r \ln \frac{\ln t}{\ln t_{\min}}.$$

The reader will note an important element of unrealism in this model: it assumes that the cost of female signal processing and detection is zero, and hence signaling is perfectly truthful and reliable. If we allow for costly female choice, we would expect that signal detection would be imperfect, and there would be a positive level of dishonest signaling in equilibrium,