13

When It Pays to Be Truthful: Signaling in Games with Friends, Adversaries, and Kin

> Each discipline of the social sciences rules comfortably within its own chosen domain . . . so long as it stays largely oblivious of the others.
>
> *Edward O. Wilson (1998):191*

13.1 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, since 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.[1]

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.

[1]For a recent 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 §13.6.

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 to) 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 1992a). 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.

13.2 A Generic Signaling Game

Signaling games are special cases of Bayesian games, presented in chapter 12. In Bayesian games, players have "types" which may be partially or wholly revealed in the course of play. In signaling games, only player 1 has a "type," and this 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 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 $f : T \to S$, where $s = f(t)$ is the signal sent when Nature reveals type t, and a pure strategy for Receiver is a function $g : S \to A$, where $a = g(s)$ is the action taken when Receiver receives signal s. A mixed strategy for Sender is a probability distribution $p_1(s; t)$ over S for each $t \in T$, and a mixed strategy for Receiver is a probability distribution $p_2(a; s)$ over A for each signal s received. A Nash equilibrium for the game is thus a pair of probability distributions $(p_1(\cdot; t), p_2(\cdot, s))$ 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 $r(t)$ 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 $(p_1(\cdot; t), p_2(\cdot; s))$, if there is a strictly positive probability that Sender will transmit s, i.e., if

$$\sum_{t \in T} \rho(t) p_1(s; t) > 0.$$

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, we know from our argument in §12.1 that a best response for Receiver maximizes Receiver's expected return, with a probability distribution over T given by

$$\mathrm{P}[t|s] = \frac{p_1(s; t)\rho(t)}{\sum_{t' \in T} p_1(s; t')\rho(t')}.$$

We thus require of p_1 and p_2 that

a. For every state $t \in T$ and all signals $s' \in S$ such that $p_1(s'; t) > 0$, s' maximizes

$$\sum_{a \in A} u(t, s', a) p_2(a; s)$$

over all $s \in S$.

b. For every signal $s \in S$ along the path of play, and all actions $a' \in A$ such that $p_2(a'; s) > 0$, a' maximizes

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

over all $a \in A$.

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. See the discussion of zero probability information sets in §5.17 and §12.1.

13.3 Introductory Offers

A product comes in two qualities, high and low, at unit costs c_h and c_l, with $c_h > c_l > 0$. Consumers purchase one unit per period, and a consumer only learns the quality of a firm's product by purchasing it in the first period. Consumers live for two periods, and a firm cannot change its quality between the first and second period. Thus, a consumer can use the information concerning product quality gained in the first period to decide whether to buy from the firm again in the second period. Moreover, firms can discriminate between first- and second-period consumers and offer different prices in the two periods, for instance, by extending an *introductory offer* to a new customer.

Suppose the value of a high-quality good to the consumer is h, the value of a low-quality good is zero, a consumer will purchase the good only if this does not involve a loss, and a firm will sell products only if it makes positive profits. We say that the industry is in a *truthful signaling equilibrium* if the firms' choice of sale prices accurately distinguishes high-quality from low-quality firms. If the firms' choices do not distinguish high from low quality, we have a *pooling equilibrium*. In the current situation, this means that only the high-quality firms will sell. Let δ be the consumer's discount factor on second-period utility.

a. Show that if $h > c_h + (c_h - c_l)\delta$, there is a truthful signaling equilibrium, and not otherwise.
b. What is the high-quality firm's price structure in a truthful signaling equilibrium?
c. Show that each consumer gains $h - c_l$ in the truthful signaling equilibrium, and firms gain $c_l - c_h + \delta(h - c_h)$ per customer.

13.4 Web Sites (for Spiders)

In the spider *Agelenopsis aperta*, individuals search for desirable locations for spinning webs. The value of a web is $2v$ to its owner. When two spiders come upon the same desirable location, the two invariably compete for it.

Spiders can be either strong or weak, but it is impossible to tell which type a spider is by observation. A spider may rear onto two legs to indicate that it is strong, or fail to do so, indicating that it is weak. However, spiders do not have to be truthful. Under what conditions will they in fact signal truthfully whether they are weak or strong? Note that if it is in the interest of both the weak and the strong spider to represent itself as strong, we have a "pooling equilibrium," in which the value of the signal is zero, and it will be totally ignored—hence, it will probably not be issued. If only the strong spider signals, we have a truthful signaling equilibrium.

Assume that when two spiders meet, each signals the other as strong or weak.[2] Based on the signal, each spider independently decides to attack or withdraw. If two strong spiders attack each other, they each incur a cost of c_s, and each has a 50% chance of gaining/keeping the territory. Thus, the expected payoff to each is $v - c_s$. If neither spider attacks, each has a 50% chance of gaining the territory, so their expected payoff is v for each. If one spider attacks and the other withdraws, the attacker takes the location, and there are no costs. So the payoffs to attacker and withdrawer are $2v$ and 0, respectively. The situation is the same for two weak spiders, except they have a cost c_w. If a strong and a weak spider attack each other, the strong wins with probability 1, at a cost b with $c_s > b > 0$, and the weak spider loses, at a cost $d > 0$. Thus, the payoff to the strong spider against the weak is $2v - b$, and the payoff to the weak against the strong is $-d$. In addition, strong spiders incur a constant cost per period of e to maintain their strength. The table shows a summary of the payoffs for the game.

Type 1,Type 2	Action 1,Action 2	Payoff 1,Payoff 2
strong,strong	attack,attack	$v - c_s$,$v - c_s$
weak,weak	attack,attack	$v - c_w$,$v - c_w$
strong,weak	attack,attack	$2v - b$,$-d$
either,either	attack,attack	$2v$,0
either,either	attack,attack	0,0

Each spider has eight pure strategies: signal that it is strong or weak (s/w), attack/withdraw if the other spider signals strong (a/w), attack/withdraw if

[2]Note that this is a signaling game in which there is *bilateral signaling*: Sender sends a signal to Receiver, Receiver simultaneously sends a signal to Sender, and they each choose actions simultaneously. The conditions for a Nash equilibrium in such games are straightforward generalizations of the conditions developed in §13.2.

the other spider signals weak (a/w). We may represent these eight strategies as saa, saw, swa, sww, waa, waw, wwa, www, where the first indicates the spider's signal, the second indicates the spider's move if the other spider signals strong, and the third indicates the spider's move if the other spider signals weak (for instance, swa means "signal strong, withdraw from a strong signal and attack a weak signal"). This is a complicated game, since the payoff matrix for a given pair of spiders has sixty-four entries, and there are four types of pairs of spiders. Rather than use brute force, let us assume there is a truthful signaling equilibrium and see what that tells us about the relationships among v, b, c_w, c_s, d, e, and the fraction p of strong spiders in the population.

Suppose $v > c_s, c_w$, and the proportion p of strong spiders is determined by the condition that the payoffs to the two conditions of being strong and being weak are equal.

a. What strategies are used in a truthful signaling equilibrium?
b. Use (a) to find the proportion p of strong spiders in a truthful signaling equilibrium. Find bounds for v in terms of e, c_w, and c_s for there to exist both strong and weak spiders in equilibrium.
c. What conditions on the parameters must hold for the equilibrium to foster truthful signaling?
d. Show that as long as both strong and weak spiders exist in equilibrium, an increase in the cost e of being strong leads to an increase in payoff to all spiders, weak and strong alike. Explain in words why this "counterintuitive" result is true.
e. Show that for some range of values of the parameters, an increase in the payoff v to the location can entail a *decrease* in the payoff to the spiders. For what value of the parameters is this the case? What value v^* maximizes the payoff to the spiders? Explain in words why this strange-seeming situation can occur.

13.5 Sex and Piety: The Darwin-Fisher Model of Sexual Selection

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 (§4.17). 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 (§6.14). 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. More recent 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 which we will call *pious fasting*. While 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, since with probability $k \geq 0$ a choosy women 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

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

While male and female young adults have equal fractions of each trait—since 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 $n^f p^f a t^m/(1-t^m+at^m)$ 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

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

and

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

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 $t^m r_t$. Since 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 $\beta_t = t^m - t$ and $\beta_p = p^f - p$, then the change Δt in the frequency of the pious trait can be written as

$$\Delta t = \frac{t^m r_t + t^f}{2} - t = \frac{1}{2}\left(\beta_t + \frac{d\beta_p}{p(1-p)}\right). \tag{13.1}$$

What about the change in p across generations? The fraction of mated, choosy females is simply p^f, since 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 nx_{-p} are nonpious and choosy, while $n(1-u)x_{tp}$ 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

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

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 Δp in the frequency of the choosy trait can be written as

$$\Delta p = \frac{p^{m'} + p^f}{2} - p = \frac{1}{2}\left(\beta_p + \frac{d\beta_t}{t(1-t)}\right). \tag{13.2}$$

Let us first investigate (13.1) and (13.2) 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

$$t = \frac{(a-1)p(1-u) - u}{u(a(1-u) - 1)}. \tag{13.3}$$

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 (13.1) and (13.2), we find that stability requires that the denominator of (13.3) 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 Fig. 13.1. 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.

Suppose, however, $k > 0$. If we then solve for $\Delta t = \Delta p = 0$ in (13.1) and (13.2), we easily derive the equation

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

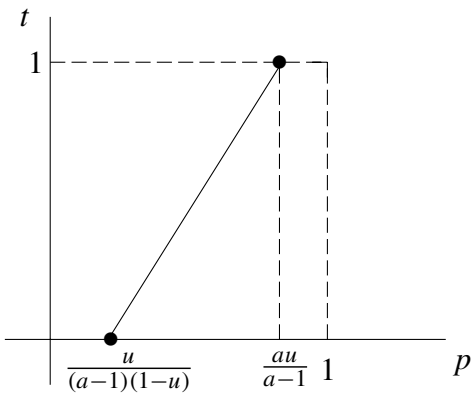

Figure 13.1. Equilibria in Darwin-Fisher sexual selection model when there is no selection against choosy females.

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*. Since 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.

13.6 Biological Signals as Handicaps

Zahavi (1975), based on close observation of avian behavior, proposed an alternative to the Darwin-Fisher sexual selection mechanism—a notion of costly signaling which 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, since an unhealthy or genetically unfit male lacks the resources to mount such a display. The idea was treated with skepticism for many years, since 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 §13.2. 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, since 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; i.e., $ds/dt > 0$. 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

$$U_s(s) = u_s(t, s, g(s)) + u_a(t, s, g(s))\frac{dg}{ds} = 0. \tag{13.4}$$

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

$$\frac{dg}{ds} = -\frac{u_s(g(s), s, g(s))}{u_a(g(s), s, g(s))}, \tag{13.5}$$

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

Differentiating the first-order condition (13.4) totally with respect to t, we find

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

Since $U_{ss} < 0$ by the second-order condition for a maximum, and since $ds/dt > 0$, 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,

$$\frac{d}{dt}\left[\frac{u_s(t, s, g(s))}{u_a(t, s, g(s))}\right] = \frac{U_{st}}{u_a} < 0. \tag{13.6}$$

We can now rewrite (13.4) as

$$u_a(t, s, g(s))\left[\frac{u_s(t, s, g(s))}{u_a(t, s, g(s))} + g'(s)\right] = 0. \tag{13.7}$$

Since the fraction in this expression is increasing in t, and the expression is zero when $t = g(s)$, this shows $s = t$ 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 principal, suppose $u(t, s, a) = a^r t^s$, so (13.5) becomes $g'/g = -(1/r)\ln g$, which has solution $\ln g = ce^{-s/r}$. Using $g(s_{\min}) = t_{\min}$ 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, and the physical process of signal development should involve an evolutionary dynamic intimately related to receiver neurophysiology (Dawkins and Guilford 1991; Guilford and Dawkins 1991, 1993). In contrast with the Darwin-Fisher model of sexual selection, we would not expect a small amount of costly female choice to undermine a signaling equilibrium, since there are direct fitness benefits to females in locating vigorous males.

13.7 The Shepherds Who Never Cry Wolf

Since we value truthfulness, one might have the impression that when both a truthful signaling and a nonsignaling equilibrium exist, the truthful signaling equilibrium should entail higher payoffs for at least some of the players. But that need not be the case. Here is a counterexample.

Two shepherds take their flocks each morning to adjoining pastures. Sometimes a wolf will attack one of the flocks, causing pandemonium among the threatened sheep. A wolf attack can be clearly heard by both shepherds, allowing a shepherd to come to the aid of his companion. But unless the wolf is hungry, the cost of giving aid exceeds the benefits, and only the shepherd guarding the threatened flock can see if the wolf is hungry.

There are three pure strategies for a threatened shepherd: never signal (N), signal if the wolf is hungry (H), and always signal (A). Similarly, there are three pure strategies for the shepherd who hears a wolf in the other pasture: never help (N), help if signalled (H), and always help (A).

We make the following assumptions. The payoff to a day's work when no wolf appears is 1 for each shepherd. The cost of being attacked by a hungry wolf and a nonhungry wolf is a and $b < a$, respectively. The cost of coming to the aid of a threatened shepherd is d, and doing so prevents the loss to the threatened shepherd, so his payoff is still 1. Finally, it is common knowledge that the probability that a wolf is hungry is $p > 0$.

We assume the shepherds' discount rates are too high, or wolf visits too infrequent, to support a repeated-game cooperative equilibrium using trigger strategies, so the game is a one-shot. If the shepherds are self-interested, of course neither will help the other, so we assume that they are brothers, and the total payoff to shepherd 1 (the threatened shepherd) is his own-payoff π_1 plus $k\pi_2$, where π_2 is the own-payoff of shepherd 2, and similarly, the total payoff to shepherd 2 (the potential helper) is $\pi_2 + k\pi_1$.

If $ka > d > kb$, a shepherd prefers to aid his threatened brother when, and only when, the wolf is hungry (why?). So we assume this is the case. We also assume that $a - dk > c > b - dk$, which means that a threatened shepherd would only want his brother to come to help if the wolf is hungry (why?). So there ought to be a signaling equilibrium in this case. Note, however, that this signaling equilibrium will exist whether p is small or large, so for very large p, it might be worthwhile for a brother *always* to help, thus saving the cost c of signaling to his brother, and saving the cost kc to himself. This, in fact, is the case. While this can be proved in general, you are asked in this problem to prove a special case.

Assume $k = 5/12$ (note that $k = 1/2$ for full brothers, but the probability that two brothers that *ostensibly* have the same father *in fact* have the same father is probably about 80% in human populations). Also assume $a = 3/4$, $b = 1/4$, $c = 19/48$, and $d = 1/4$. Finally, assume $p = 3/4$. After verifying that the above inequalities hold, do the following:

a. Show that there is a signaling equilibrium, and find the payoffs to the shepherds.
b. Show that there is pooling equilibrium in which a threatened shepherd never signals, and a shepherd always helps his threatened brother. Show that this equilibrium is Pareto-superior to the signaling equilibrium.
c. There is also a mixed strategy Nash equilibrium (truthful signaling occurs, but not with certainty) in which the threatened shepherd sometimes signals, and the other shepherd sometimes helps without being asked. Find this equilibrium and its payoffs, and show that the payoffs are slightly better than the signaling equilibrium but not as high as the pooling equilibrium.

13.8 My Brother's Keeper

Consider the following elaboration on the theme of §13.7. Suppose the threatened shepherd, whom we will call the Sender, is either healthy, needy, or desperate, each of which is true with probability 1/3. His brother, whom we will call the Donor, is either healthy or needy, each with probability 1/2. Suppose there are two signals that the threatened shepherd can give: a low-cost signal costing 0.1, and a high-cost signal costing 0.2. If he uses either one, we say he is "asking for help." We assume that the payoff for each brother is his own fitness plus 3/4 of his brother's fitness. The Sender's fitness is 0.9 if healthy, 0.6 if needy, and 0.3 if desperate, minus whatever he pays for signaling. The Donor's fitness is 0.9 if healthy and 0.7 if needy. However, the Donor has a resource that ensures his fitness is 1 if he uses it, and the fitness of the Sender is 1 (minus the signaling cost) if he transfers it to the Sender. The resource is perishable, so either he or his brother must use it in the current period.

a. Show that after eliminating "unreasonable" strategies (define carefully what you mean by "unreasonable"), there are six pure strategies for the Sender, in each of which a healthy sender never signals: Never Ask; Signal Low If Desperate; Signal High If Desperate; Signal Low If Desperate or Needy; Signal Low If Needy, High If Desperate; and Signal High If Needy or Desperate. Similarly, there are ten strategies for Donor: Never Help; Help If Healthy and Signal Is High; Help If Healthy and Asked; Help If Healthy; Help If Signal Is High; Help If Healthy and Asked, or Needy and Signal Is High; Help If Healthy or

Signal Is High; Help If Asked; Help If Healthy or Asked; and Help Unconditionally.

b. * If you have a lot of time on your hands, or if you know a computer programming language, derive the 6×10 normal form matrix for the game.

c. * Show that there are seven pure strategy equilibria. Among these there is one completely pooling equilibrium: Never Ask, Always Help. This, of course, affords the Sender the maximum possible payoff. However, the pooling equilibrium maximizes the sum of the payoffs to both players, so it will be preferred by both if they are equally likely to be Sender and Donor. This is asocial optimum even among the mixed strategy equilibria, but that is even harder to determine—my Normal Form Game Solver and Gambit are useful here.

d. Show that the truthful signaling strategies (Signal Low If Needy, High If Desperate, Help If Healthy and Asked or Needy and Signal Is High) form a Nash equilibrium, but that this equilibrium is strictly Pareto-inferior to the pooling (nonsignaling) equilibrium.

This model shows that there can be many signaling equilibria, but all may be inferior to complete altruism (Never Ask, Always Help). This is doubtless because the coefficient of relatedness is so high (3/4 is the coefficient of relatedness between sisters in many bee species, where the queen mates with a single haploid male).

Simulating the model gives an entirely surprising result, as depicted in Fig. 13.2. For this simulation, I created seven hundred players, each randomly programmed to play one strategy as Sender and another as Donor. The players were randomly paired on each round, and one was randomly chosen to be Sender, the other Donor. After every ten rounds, the strategies with the highest scores reproduced, and their offspring replaced those with the lowest scores. Figure 13.2 shows the outcome for the two strongest strategies. For the Donor, this involved using Help If Healthy or If Asked, and for Sender, either Signal Low If Desperate or Needy, or Signal Low If Desperate. After 20,000 rounds, the only remaining strategy (except for occasional mutations), is the latter, the other fifty-nine strategies having disappeared. This is the signaling equilibrium that is best for the Donor but whose total fitness is inferior to the pooling equilibrium Never Ask, Always Help. Nor is this a fluke outcome: I ran the simulation ten times with different seeds to the random number generator, and this equilibrium emerged every time. The implication is clear: *a signaling equilibrium can*

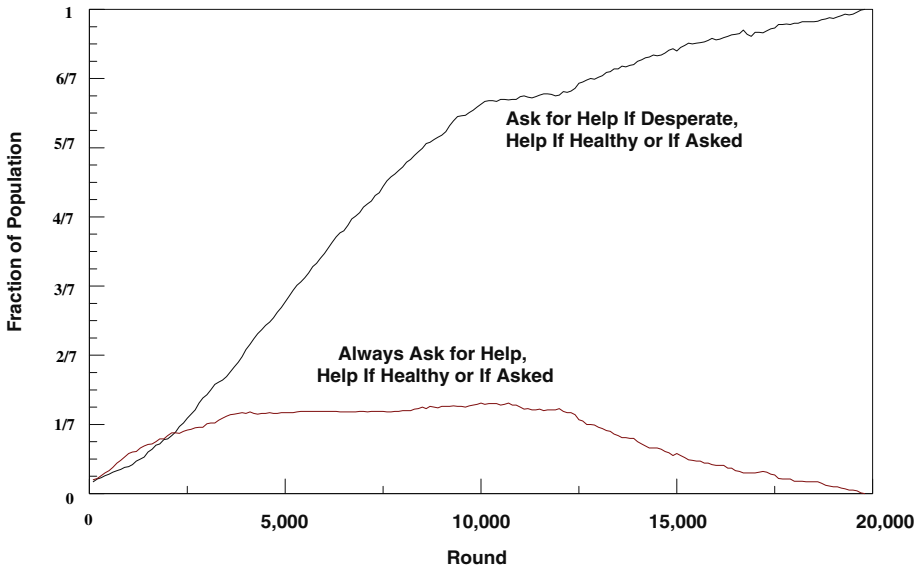

Figure 13.2. A signaling equilibrium in the Brother's Helper game.

emerge from an evolutionary process even when it is inferior to a pooling equilibrium.

13.9 Honest Signaling among Partial Altruists

In a certain fishing community, each fisher works alone on the open sea, earning a payoff that we will normalize to 1. A fisher occasionally encounters threatening weather. If the fisher does not escape the weather, his payoff is zero. If a threatened fisher has sufficient energy reserves, he can escape the bad weather, and his expected payoff is u, where $0 < u < 1$. We call such a fisher *secure*. However, with a certain probability p ($0 < p < 1$) a threatened fisher does *not* have sufficient energy reserves. We say he is *in distress*.

If a threatened fisher sees another fisher on the horizon, he can send a signal to ask for help, at cost t, with $0 < t < 1$. If the fisher is in distress and a potential helper comes to his aid (we assume the potential helper is not threatened), the payoff to the distressed fisher is 1, but the cost to the helper is $c > 0$. Without the help, however, the distressed fisher succumbs to the bad weather and has payoff 0.

To complicate matters, a threatened fisher who is helped by another fisher but who is *not* distressed has payoff v, where $1 > v > u$. Thus, threatened fishers have an incentive to signal that they are in distress even when they are not. Moreover, fishers can tell when other fishers are threatened, but only the threatened fisher himself knows his own reserves, and hence whether or not he is in distress.

We assume that encounters of this type among fishers are one-shot affairs, because the probability of meeting the same distressed fisher again is very small. Clearly, unless there is an element of altruism, no fisher will help a threatened fisher. So let us suppose that in an encounter between fishers, the nonthreatened fisher receives a fraction $r > 0$ of the total payoff, including signaling costs, received by the threatened fisher (presumably because r is the degree of genetic or cultural relatedness between fishers). However, the helper bears the total cost c himself.

For example, if a fisher is in distress and signals for help and receives help, the distressed fisher's payoff is $1 - t$ and the helper's payoff is $r(1 - t) - c$.

The nonthreatened fisher (Fisher 1) who sees a threatened fisher (Fisher 2) has three pure strategies: Never Help, Help If Asked, and Always Help. Fisher 2 also has three strategies: Never Ask, Ask When Distressed, Always Ask. We call the strategy pair {Help If Asked, Ask If Distressed} the *Honest Signaling* strategy pair. If this pair is Nash, we have an Honest Signaling equilibrium. This is called a *separating equilibrium* because agents truthfully reveal their situation by their actions. Any other equilibrium is called a *pooling equilibrium*, since agents' actions do not always reveal their situations.[3]

The reasoning you are asked to perform below shows that when there are potential gains from helping distressed fishers (i.e., $(1+r)(1-t) > c$), then if fishers are sufficiently altruistic and signaling is sufficiently costly but not excessively costly, an Honest Signaling equilibrium can be sustained as a Nash equilibrium. The idea that signaling must be costly (but not too costly) to be believable was championed by Amotz Zahavi (1975) and modeled by Grafen (1990a), Maynard Smith (1991), Johnstone and Grafen (1992, 1993), and others in a notable series of papers. The general game-theoretic point is simple, but extremely important: if a signal is not on balance truthful, it will not be heeded, so if it is costly to emit, it will not be emitted. Of course, there is much out-of-equilibrium behavior, so there is lots of room for duplicity in biology and economics.

[3]For more on separating and pooling equilibria, see §12.10, §12.11, and chapter 13.

a. Show that if

$$(1+r)\left[v-u+\frac{pt}{1-p}\right] < c < (1+r)(1-t), \tag{13.8}$$

then Honest Signaling is socially efficient (i.e., maximizes the sum of the payoffs to the two fishers)? HINT: Set up the 3×3 normal form for the game, add up the entries in each box, and compare. For the rest of the problem, assume that these conditions hold.

b. Show that there is always a pooling equilibrium in which Fisher 2 uses Never Ask. Show that in this equilibrium, Fisher 1 Never Helps if

$$c > r[p + (1-p)(v-u)] \tag{13.9}$$

and Always Helps if the opposite inequality holds.

c. Show that if

$$v - u < \frac{c}{r} < 1$$

and

$$v - u < t < 1,$$

Honest Signaling is a Nash equilibrium.

d. Show that if t is sufficiently close to 1, Honest Signaling can be a Nash equilibrium even if it is not socially efficient.

e. Show that if Honest Signaling and {Never Ask, Never Help} are both Nash equilibria, then Honest Signaling has a higher total payoff than {Never Ask, Never Help}.

f. Show that if Honest Signaling and {Never Ask, Always Help} are both Nash equilibria, then Honest Signaling has a higher total payoff than {Never Ask, Always Help}.

13.10 Educational Signaling I

Suppose there are two types of workers, high-ability (h) and low-ability (l), and the proportion of high-ability workers in the economy is $\alpha > 0$. Suppose workers invest in acquiring a level of schooling s, which is both costly to obtain and productive. Specifically, suppose that a high-ability worker incurs a cost $c_h(s)$ of obtaining s years of schooling, while a low-ability workers incurs a cost of $c_l(s)$. We also assume schooling is more costly for low-ability workers than for high, so $c_h'(s) < c_l'(s)$ for all $s \geq 0$.

Schooling is also productive, so the marginal productivity of a high-ability worker with s years of schooling is $y_h(s)$, and the corresponding value for a low-ability worker is $y_l(s)$. We assume $y_h(0) = y_l(0) = 0$ and $y_h'(s) > y_l'(s) > 0$ for all $s \geq 0$, which means that high-ability workers have higher marginal products than low-ability workers, and schooling increases the productivity of high-ability workers more than low. To simplify the diagrams, we assume y_h and y_l are linear functions of s.

Suppose employers cannot observe ability, but they do observe s, and if workers with different abilities obtain different amounts of schooling, they may offer a wage based on s. We assume the labor market is competitive, so all firms must offer a wage equal to the expected marginal product of labor.

A truthful signaling equilibrium in this case involves high- and low-ability workers choosing different amounts of schooling, so employers know their type by their schooling choices. They thus pay wages $y_h(s)$ to the high-ability workers and $y_l(s)$ to the low. Assuming workers know this, high-ability workers will choose s to maximize $y_h(s) - c_h(s)$ and low-ability workers will choose s to maximize $y_l(s) - c_l(s)$. This is depicted in Fig. 13.3. Agents maximize their payoff by choosing the highest indifference curve that intersects their wage curve, which means tangency points between wage curves and indifference curves as illustrated. Moreover, neither type of agent would prefer to get the amount of schooling chosen by the other, since this would involve a lower level of utility; i.e., the equilibrium point for each type lies below the indifference curve for the other type.

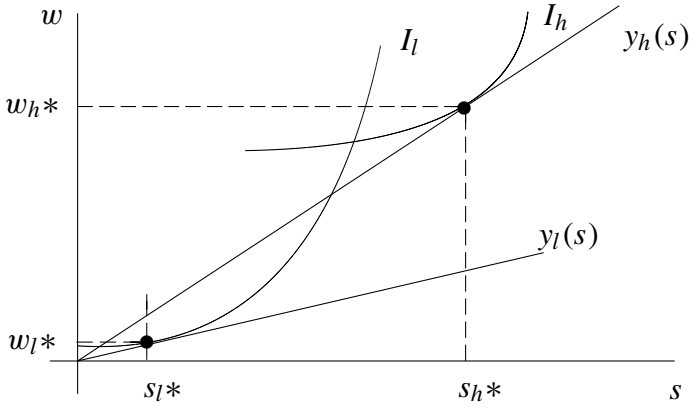

Figure 13.3. A truthful signaling equilibrium.

a. Explain why there cannot be a truthful signaling equilibrium if the costs of schooling are the same for the two ability levels. Draw a diagram to illustrate your argument. HINT: Indifference curves for the same utility function cannot cross.
b. Modify Fig. 13.3 to illustrate the following assertion: If the optimum schooling level for the high-ability worker lies inside the optimal indifference curve for the low-ability worker, then the low-ability worker will mimic the high-ability worker and destroy the truthful signaling equilibrium.
c. However, high-ability workers may have a response to this: they may be able to increase their educational level to a point sufficiently high that it will no longer benefit the low-ability workers to imitate them. This is called an "Educational Rat Race." Make a diagram illustrating this rat race and another in which it is not worthwhile for high-ability workers to signal their quality.
d. Analyze the case of a pooling equilibrium, in which both high- and low-ability workers choose the same schooling level. Show that in this case employers do not use either the $y_h(s)$ or the $y_l(s)$ schedules, but rather set wages so that

$$w(s) = \alpha y_h(s) + (1 - \alpha) y_l(s) \tag{13.10}$$

for both types of workers. Show that in a pooling equilibrium, high-ability workers maximize their payoff subject to hitting the wage curve $w(s)$, and low-ability workers imitate their choice of educational level. Draw a diagram illustrating this result, and make sure the curves are drawn so neither high- nor low-ability workers have an incentive to switch unilaterally to the truthful signaling equilibrium.

This analysis does not exhaust the possibilities for a signaling equilibrium. There could also exist mixed strategy equilibria in which some low-ability workers imitate the high-ability workers and others do not. There could also be strange Bayesian priors for the employers that would lead to strange pooling equilibria. For instance, if employers believe that a worker who does not choose $s = s_o$ for some given s_o are "crazy" and must be low-ability. Then every worker may choose s_o to get the pooling wage, which is higher than the low-ability wage. Such behavior by employers would be stupid, and they might be driven out of existence in a dynamic adjustment process.

13.11 Education as a Screening Device

Suppose a worker can be of high ability a_h with probability α, or low ability $a_l < a_h$ with probability $1 - \alpha$. Workers know their own ability, but employers do not. Workers can also choose to acquire high as opposed to low education, and this is observable by employers. Moreover, it costs c/a ($c > 0$) for a worker of ability a to acquire high education, so high education is more costly for the low-ability worker. We assume that workers are paid their expected marginal product, and the marginal product of a worker of ability a is just a, so high education does not improve worker productivity—education is at best a screening device, informing employers which workers are high ability. Suppose e_l is the event "worker chose low education" and e_h is the event "worker chose high education." Then, if w_l and w_h are the wage paid to low- and high-education workers, respectively, we have

$$w_k = \mathrm{P}[a_h|e_k]a_h + \mathrm{P}[a_l|e_k]a_l, \qquad k = l, h, \tag{13.11}$$

where $\mathrm{P}[a|e]$ is the conditional probability that the worker has ability a in the event e.

A Nash equilibrium for this game consists of a choice $e(a) \in \{e_l, e_h\}$ of education level for $a = a_h, a_l$ and a set of probabilities $\mathrm{P}[a|e]$ for $a = a_h, a_l$ and $e = e_h, e_l$ that are consistent in the sense that if $\mathrm{P}[e] > 0$, then $\mathrm{P}[a|e]$ is given by Bayesian updating.

a. Show that there is a pooling equilibrium in which $e(a_h) = e(a_l) = e_l$, $w_h = w_l = \alpha a_h + (1-\alpha)a_l$, and $\mathrm{P}[a_l|e_l] = \mathrm{P}[a_l|e_h] = 1 - \alpha$. In other words, employers disregard the education signal, and workers choose low education.
b. Show that there is some range of values for c such that there is a truthful signaling equilibrium in which $e(a_h) = e_h$, $e(a_l) = e_l$, $w_l = a_l$, $w_h = a_h$, $\mathrm{P}[a_l|e_l] = 1$, and $\mathrm{P}[a_l|e_h] = 0$. In other words, despite the fact that education does not increase worker productivity, workers can signal high ability by acquiring education, and employers reward high-ability workers with relatively high wages.
c. In the spirit of trembling hand perfection (§5.16 and §12.1), suppose that with a small probability $\epsilon > 0$ a worker is given a free education, regardless of ability. Show that the pooling equilibrium does not have to specify arbitrarily the probabilities $\mathrm{P}[a_l|e_h]$ off the path of play, since $\mathrm{P}[e_h] = \epsilon > 0$, and since both ability types are equally likely to get a free education, we have $\mathrm{P}[a_l|e_h] = 1 - \alpha$.

d. Show that if c is sufficiently small, there are two pooling equilibria and no truthful signaling equilibrium. The first pooling equilibrium is as before. In the second pooling equilibrium, both ability types choose to be educated. Specifically, $e(a_h) = e(a_l) = e_h$, $w_l = a_l$, $w_h = \alpha a_h + (1 - \alpha)a_l$, $\mathrm{P}[a_l|e_l] = 1$, and $\mathrm{P}[a_l|e_h] = 1 - \alpha$. Note that this requires specifying the probabilities for e_l, which are off the path of play. The truthful signaling equilibrium is inefficient and inegalitarian, while the pooling equilibrium is inefficient but egalitarian. The pooling equilibrium is not very plausible, because it is more reasonable to assume that if a worker gets education, he is high ability.

e. Show that if we added a small exogenous probability $\epsilon > 0$ that a worker of either type is denied an education, all outcomes are along the path of play, and the posterior $\mathrm{P}[a_l|e_l] = 1 - \alpha$ follows from the requirement of Bayesian updating.

f. Now suppose the educational level is a continuous variable $e \in [0, 1]$. Workers then choose $e(a_h), e(a_l) \in [0, 1]$, and employers face probabilities $\mathrm{P}[a_h|e]$, $\mathrm{P}[a_l|e]$ for all education levels $e \in [0, 1]$.

   Show that for $e \in [0, 1]$, there is a $\bar{e} > 0$ such that for any $e^* \in [0, \bar{e}]$, there is a pooling equilibrium where all workers choose educational level e^*. In this equilibrium, employers pay wages $w(e^*) = \alpha a_h + (1-\alpha)a_l$ and $w(e \neq e^*) = a_l$. They have the conditional probabilities $\mathrm{P}[a_l|e \neq e^*] = 1$ and $\mathrm{P}[a_l|e = e^*] = 1 - \alpha$.

g. Show that when $e \in [0, 1]$, if c is sufficiently large, there is a range of values of e^* such that there is a truthful signaling equilibrium where high-ability workers choose $e = e^*$ and low-ability workers choose $e = 0$. In this equilibrium, employers pay wages $w(e^*) = a_h$ and $w(e \neq e^*) = a_l$. They face the conditional probabilities $\mathrm{P}[a_l|e \neq e^*] = 0$ and $\mathrm{P}[a_l|e \neq e^*] = 1$.

13.12 Capital as a Signaling Device

Suppose there are many producers, each with a project to fund. There are two types of projects, each of which requires capital investment k. The "good" project returns 1 at the end of the period, and the "bad" project returns 1 with probability p $(0 < p < 1)$ at the end of the period, and otherwise returns 0. There are also many lenders. While each producer knows the type of his own project, the lenders only know that the frequency of good projects in the economy is q $(0 < q < 1)$.

We assume the capital market is perfect and all agents are risk neutral (§16.41). Thus, each lender's reservation position is the risk-free interest rate $\rho > 0$, so a producer can always obtain financing for his project if he offers to pay an interest rate r that allows a lender to earn expected return ρ on his capital investment k.

We call a project with capital cost k *socially productive* if its expected return is greater than $k(1 + \rho)$. This corresponds to the idea that while individual agents may be risk averse, the law of large numbers applies to creating a social aggregate, so a social surplus is created on all projects that return at least the risk-free interest rate.

a. Show that for any $p, q > 0$ there is a nonempty interval (k^g_{min}, k^g_{max}) of capital costs k such that no project is funded, despite the fact that a fraction q of the projects are socially productive.
b. Show that for any $p, q > 0$ there is a nonempty interval (k^b_{min}, k^b_{max}) of capital costs k such that all projects are funded, despite the fact that a fraction $1 - q$ of the projects are not socially productive.

This is a sorry state of affairs, indeed! But is there not some way that an owner of a good project could signal this fact credibly? In a suitably religious society, perhaps the requirement that borrowers swear on a stack of Bibles that they have good projects might work. Or if producers have new projects available in each of many periods, we may have a "reputational equilibrium" in which producers with bad projects are not funded in future periods, and hence do not apply for loans in the current period. Or society might build debtors' prisons and torture the defaulters.

But suppose none of these is the case. Then equity comes to the rescue! Suppose each producer has an amount of capital $k^p > 0$. Clearly, if $k^p \geq k$, there will be no need for a credit market, and producers will invest in their projects precisely when they are socially productive (prove it!). More generally,

c. Show that for all $p, q, k > 0$ such that good projects are socially productive and bad projects are socially unproductive, there is a wealth level $k^p_{min} > 0$ such that if all producers have wealth $k^p > k^p_{min}$; a producer's willingness to invest k^p in his project is a perfect indicator that the project is good. In this situation, exactly the good projects are funded, and the interest rate is the risk-free interest rate ρ.

The previous result says that if producers are sufficiently wealthy, there is a truthful signaling equilibrium, in which producers signal the quality of their

projects by the amount of equity they are willing to put in them. But if there are lots of nonwealthy producers, many socially productive investments may go unfunded (Bardhan, Bowles, and Gintis 2000).