Efficiency-Enhancing Signalling in the Samaritan's Dilemma

Johan Lagerlöf*

WZB — Social Science Research Center Berlin

Reichpietschufer 50, D-10785 Berlin, Germany

lagerloef@medea.wz-berlin.de

Phone: +49-30-25491 421; fax: +49-30-25491 444

May 23, 2002

Abstract

Suppose an altruistic person, A, is willing to transfer resources to a second person, B, if B comes upon hard times. If B anticipates that A will act in this manner, B will save too little from both agents' point of view. This is the Samaritan's dilemma. The logic of the dilemma has been employed in an extensive literature, addressing a wide range of both normative and positive issues. This paper shows, however, that the undersaving result is mitigated if we relax the standard assumption of complete information. The reason for this is that if A is uncertain about how big B's need for support is, B will have an incentive to signal that he is in great need by saving more that he otherwise would have done. //[Doc: SD-16.tex]//

JEL classification: D64; D82

Keywords: Altruism; Saving; Efficiency; Signalling

*I thank two anonymous referees, J. Björnerstedt, D. Coen-Pirani, Y. Kang, N.-P. Lagerlöf, B. Persson, D. Soskice, J. Stennek, K. Wärneryd, J. Weibull, and seminar participants at the Stockholm School of Economics, the WZB, ESEM -98, EEA -98, and the Universities of Rochester (the Wallis Institute), Copenhagen and Alicante, as well as Free University of Berlin for helpful comments. Part of this research was conducted while I was visiting the Wallis Institute; I am grateful to them for their hospitality. Financial support from Jan Wallander and Tom Hedelius' Research Foundation, *Svenska Institutet*, *Finanspolitiska Forskningsinstitutet*, and the European Commission (under the TMR Programme; contract no. ERBFMRXCT980203) is gratefully acknowledged.

Suppose an altruistic person, hereafter called A, is willing to transfer resources to a second person, B, if B comes upon hard times. Then, if B today is to decide how much to save for tomorrow, and if B is well aware of A's altruistic concern for him, B will typically save too little as compared to what is socially optimal. This is what Buchanan (1975) has called the "Samaritan's dilemma". The dilemma arises because A is unable to commit not to help B out. Moreover, A's willingness to bail B out if he undersaves serves as an implicit tax on B's savings. For if B saves an extra dollar, then A will transfer, say, ten cents less to B than otherwise. This implicit tax distorts B's saving incentives. As a result, given the equilibrium level of A's support, B would be better off if he consumed less today and more tomorrow. And, since A has altruistic concerns for the welfare of B, this would make also A better off.[1]

The Samaritan's-dilemma effect has been employed in a large number of papers, addressing a wide range of both normative and positive issues. For instance, the inefficiency result has been used to justify and/or explain the existence of compulsory social insurance systems (Thompson, 1980; Veall, 1986; Kotlikoff, 1987; Lindbeck and Weibull, 1988; Hansson and Stuart, 1989). The argument is that a government can force people to save and insure more than they would do voluntarily, thereby making free riding and the Samaritan's-dilemma-type inefficiency impossible. As another example, Bruce and Waldman (1991) and Coate (1995) argue that the Samaritan's dilemma provides an efficiency rationale for in-kind govermental transfers.[2] In those models the government provides a transfer on two occasions over time. Since a cash transfer would be used in an inefficient manner, all parties can benefit if the government gives the first transfer in a tied fashion, such as in the form of an illiquid investment.

Yet another example is from the macroeconomics literature. O'Connell and Zeldes (1993) study an infinite-horizon OLG-model with altruism from children towards their parents. If, as is assumed in the standard litera-

[1] Formal analyses of the Samaritan's dilemma can be found in, e.g., Bernheim and Stark (1988) and Lindbeck and Weibull (1988). In the latter paper it is shown that the dilemma also arises in the case where both agents are altruistic towards each other.

[2] In the context of intra-family transfers, Becker and Murphy (1988, p. 7) also make this argument. They do not, however, provide a formal model.

ture, parental saving is non-strategic, this kind of model is characterized by dynamic inefficiency, that is, the growth rate of the population exceeds the (endogenous) real interest rate. O'Connell and Zeldes demonstrate, however, that the strategic undersaving effect will make the economy dynamically efficient. The reason for this is that less saving leads to a smaller capital stock, which in turn implies a larger marginal product of capital and thus a higher interest rate.

The Samaritan's dilemma also has bearing on the so-called rotten-kid theorem (Becker, 1974). This result concerns a situation where a selfish child can take an action that affects the income of the whole family. The theorem states that if the child's parent is sufficiently altruistic towards the child to transfer resources to it, then the child will choose an action that maximizes the income of the whole family. Hence, the presence of parental altruism induces the child to internalize the externality, and the resource allocation in the family is efficient. One of the conditions needed for this result to hold is that the transfer from parent to child indeed is positive.[3] However, Bruce and Waldman (1990) consider a two-period setting where the child in the first period takes an action affecting the income of the whole family *and* makes a saving decision. They show that if the parent makes an operative transfer (i.e., if the constraint that the transfer is non-negative is not binding) in the second period, then the child indeed chooses the action that maximizes family income. But, because of the Samaritan's dilemma, in this case the child also saves an amount that is too low relative to the efficient level. As a consequence, "rotten kids actually act rotten in at least one dimension, with the result being that the family unit does not achieve the Pareto frontier" (Bruce and Waldman, 1990, p. 157).

Most of the existing literature on the Samaritan's dilemma assumes a setting with complete information: the agents know with certainty their own payoffs and the other agents' payoffs. This is typically an unrealistic assumption. Indeed, in many of the real world situations that are meant to be captured by the models in this literature, there is reason to believe that

[3]There are also other conditions, which are left implicit by Becker. For example, Bergstrom (1989) shows that utility must be transferable for the result to hold.

there is a substantial degree of incomplete information.[4] Yet there is a strong a priori reason to believe that, if one allowed for incomplete information, the undersaving result in the standard Samaritan's-dilemma model should be if not eliminated so at least reduced.

To see this, consider a situation like the one described in the introductory paragraph. Suppose, however, that B (i.e., the recipient of the transfer) has private information about some characteristic of himself that is relevant for his payoff. Moreover, suppose that this characteristic can be represented by a parameter x and that x is such that the larger its magnitude, the lower is B's marginal utility of consumption tomorrow. For example, x could be a measure of an exogenous income that B will receive tomorrow. Since A cares about the welfare of B, A would be willing to make a larger transfer to B tomorrow if A believed that x were small. The assumptions about x also imply that the smaller is x, the more B wants to save (everything else being equal). B thus has an incentive to make A believe that x is small, and B may try to do so by using his savings as a signalling device (à la Spence, 1973). In particular, B has an incentive to save *more* than in the standard setting with complete information. One should thus expect this mechanism to counteract the incentives to undersave in the traditional Samaritan's-dilemma model.[5]

The purpose of this paper is to investigate whether the above intuition holds true in a formal analysis and, if so, to see how far the counteracting mechanism can take us. Section 1 of the paper starts out by presenting

[4] The reader may object that the perhaps most common application of the Samaritan's dilemma concerns the family, and members of a family often know each other's preferences fairly well. Although this may be true, the reason why they know each other's preferences should be that within a family there are ample opportunities to communicate with each other or observe each other's behaviour. It is precisely this communication (in the form of costly signalling) that I will argue counteracts the undersaving result. In fact, in the model that I will specify and solve, the receiver of the signal learns the sender's preferences perfectly.

[5] There is a related literature on signalling and altruism in theoretical biology; see Grafen (1990) and Maynard Smith (1991) for seminal contributions and Godfray and Johnstone (2000) for a survey. Of particular interest to the present paper is Maynard Smith's so-called Sir Philip Sydney game, in which the beneficiary of a transfer of resources, for example a nestling, has private information about its true need. By begging, the nestling can send costly signals about its need to the parent. To the best of my knowledge, the Samaritan's-dilemma effect does not appear in the papers in this literature, nor can the arguments of the present paper be found there.

an example which in the simplest possible way shows how the mechanism works. In this example, B can only make the binary choice whether to "save" or to "squander"; A then chooses one of three different transfer levels. In Sections 2 and 3 a somewhat less stylized but still simple model is considered where A and B can choose among a continuum of transfer and saving levels, respectively. Section 2 formulates the analytical benchmark: a model of the Samaritan's dilemma characterized by complete information. It turns out that in this model the inefficiency result is obtained if the degree of altruism is neither too low nor too high.

Section 3 extends the model in Section 2 by assuming that B has private information about a parameter in his payoff function. This parameter can take one of two distinct values; that is, B can be one of two *types* ("low" or "high"). The analysis is restricted to values of the degree of altruism such that, for both types of B, the undersaving result would obtain if B's type were common knowledge. It is shown that if we impose a commonly used equilibrium refinement (namely, the intuitive criterion), this model has a unique equilibrium outcome. The question is then asked whether this equilibrium outcome is efficient. It turns out that while the behaviour of the low type of B is unaffected by the presence of incomplete information, the high type's saving choice is indeed distorted upwards. In particular, if the degree of A's altruism is (roughly speaking) sufficiently low, then the high type will save *exactly* the amount that is efficient for him.[6] For intermediate values of the altruism parameter the high type will undersave, although here, too, he will save more than he would have done under complete information. Finally, if the altruism parameter is large enough, the signalling incentive may actually make the high type save *more* than the efficient amount.

Section 4 concludes with a discussion of the results and their robustness. Most of the proofs are found in the Appendix.[7]

[6] Crucial for this result is the assumption that A's transfer cannot be negative: she cannot *take* resources from B.

[7] Because of space constraints, several proofs have also been relegated to Lagerlöf (2002), which is available from the author on request. In particular, Lagerlöf (2002) contains all proofs that are omitted from Section 2. It also proves the claims about pooling equilibria made in Section 3 as well as a result stated in footnote 20. All other proofs that are not in the main body of the text can be found in the Appendix.

1 A First Example: "Rich Man, Poor Man"

Consider the following simple game played between two individuals, A and B (see also Fig. 1). B can be of two types: either he is "rich" or "poor". While B knows his type from the outset of the game, A does *not* know B's type. A and B make one decision each. First B chooses whether to "save" or to "squander". Second, A observes whether B has saved or squandered and then chooses whether to give B a "big support" (abbreviated "bs" in the figure), a "small support" (ss), or "no support" (ns). The payoffs are such that if B has chosen to save, A wants to give B no support regardless of whether B is poor or rich. If B has squandered, then A wants to give B a big support if he is poor and a small support if he is rich. B, on the other hand, prefers a big support to a small support and a small support to no support, regardless of his type and regardless of whether he has saved or squandered.

Before analyzing the game with incomplete information depicted in Fig. 1, let us make the observation that if A knew with certainty whether B is poor or rich (and if A's knowing this were common knowledge), then we would have an example of the Samaritan's dilemma: B would squander and then get a support from A, an outcome which is not (Pareto) efficient. To see this, start with the case where it is common knowledge that B is poor. It is straightforward to verify that then there is a unique (subgame perfect) equilibrium outcome, namely the one where B squanders and A gives him a big support. This outcome is dominated, however, by the outcomes where B saves and then gets either a small or a big support. Similarly with the case where it is common knowledge that B is rich. Then there is again a unique equilibrium outcome, namely the one where B squanders and gets a small support. This outcome is also inefficient, since it is dominated by the outcomes where B saves and then gets either a small or a big support.

Let us now solve for the equilibria of the game where A does *not* know whether B is poor or rich. As indicated in Fig. 1, A puts the prior probability $\mu \in (0,1)$ on the event that B is poor and the complementary probability $1-\mu$ on the event that B is rich. The solution concept that I employ is that of perfect Bayesian equilibrium.[8] First notice that in any such equilibrium,

[8] This solution concept requires that, after having observed that B has either saved or

if B has saved, then A will choose "no support". This is because doing so is A's best action regardless of B's type. Similarly, if B has squandered, A will never choose "no support" in an equilibrium. This means that, if being rich, B will squander in any equilibrium. Having established this, let us first look for an equilibrium in which the poor type saves and the rich type squanders; this is the only possible separating equilibrium.[9] In this kind of equilibrium, A will be able to infer B's true type perfectly. Thus, A's best response is to give B a small support if B has squandered and no support if B has saved. Given this behaviour of A, it is indeed optimal for B to save if he is poor and squander if he is rich. Hence, this is an equilibrium.

Although the outcome for the rich type in this equilibrium is the same as in the equilibrium of the corresponding complete-information game, this is not true for the poor type. For the poor type we can in the incomplete-information game sustain the outcome where B saves and A does not give B any support. This outcome is efficient among the outcomes that are relevant for the poor type. Interestingly, it is thus the *less* able type who behaves prudently and saves an amount that is efficient for him, whereas the more able type squanders his income. The reason why the poor type chooses to save is that if he squandered, then he would be perceived as the rich type and get only a small support, an outcome in which he would get a lower payoff than he gets in the equilibrium. In other words, the possibility that B is rich exerts an externality on the poor type. This externality is bad for B, since the poor type now gets only a payoff of 2 instead of a payoff of 3 as in the complete-information model. The externality is good for A, however, since she gets a payoff of 4 instead of 2 when B is poor.

For $\mu \in [.5, 1)$ there also exists a pooling equilibrium where both types of B squander, and A gives B a big support if B has squandered and no support in the out-of-equilibrium event that B has saved. Finally, for $\mu \in (.5, 1)$ there is a semi-pooling equilibrium where the poor type squanders with probability

squandered, A forms some beliefs about B's type. These beliefs must be consistent with Bayes' rule and B's equilibrium strategy, whenever Bayes' rule is defined. Whether B has saved or squandered, A makes a decision that, given her beliefs, maximizes her payoff. B is also required to make a decision that maximizes his payoff given A's behaviour.

[9] A separating equilibrium is an equilibrium in which one type saves and the other squanders.

$(1-\mu)/\mu$ and saves otherwise, and the rich type always squanders. If B has squandered, A gives B a big or a small support with equal probability; if B has saved, then A gives no support.[10]

For those μ's that give rise to multiple equilibria, it is not clear which one of the three equilibrium outcomes is the most reasonable prediction of the game. Yet, if we believe in the assumption that A does not know B's type perfectly and if we are not certain that the players will behave according to the pooling equilibrium, then it is tempting to conclude that the theoretical case for undersaving is weaker than what one might think if one only looked at the traditional formulation of the Samaritan's dilemma. One might object, however, that the example we have analyzed is rather special. For instance, it is not clear how restrictive the specification of the players' payoffs is. Moreover, standard equilbrium refinements that put restrictions on the players' beliefs about out-of-equilbrium events do not have any bite in this simple example, whereas one should expect this to be the case in many alternative settings. I will, therefore, in the following two sections further explore the signalling effect, but in a model that is somewhat less stylized than the above example.

2 The Benchmark: Complete Information

2.1 The Model

There are two individuals, A and B, and two time periods, 1 and 2. A lives only in period 2 while B lives in both periods. At the beginning of the first

[10] Notice that two of the three kinds of equilibria (namely, the separating and the semi-pooling) have the following property: if the probability that B is rich is very small, then, from an ex ante perspective, the outcome will be very close to efficiency. In a working paper version of the present paper (Lagerlöf, 2000), this point was explored both within the context of the example studied here and the model described and analyzed in the following two sections. The observation is interesting, as it suggests that adding only a small amount of uncertainty to a Samaritan's-dilemma game with complete information can substantially change the prediction of the game, and in particular it can restore efficiency almost fully. Still, for this to be true it is important exactly how the small amount of uncertainty is introduced: the argument does not work for the case where there is a small probability that B is poor rather than rich. Moreover, as was argued in the Introduction, in many of the real-world situations that are meant to be captured by the Samaritan's dilemma there is reason to believe that there is a *substantial* degree of incomplete information.

period, B is endowed with exogenous income $\omega > 0$. B's decision concerns how much of this income to save for period 2, $s \in [0, \omega]$. The residual amount, $c_{1B} = \omega - s$, constitutes B's first-period consumption. A's endowment also equals ω. In the second period, after having observed s, A chooses how much of her endowment to transfer to B, $t \in [0, \omega]$.[11] A consumes the residual amount, $c_A = \omega - t$, herself. B's second-period consumption consists of his savings plus the transfer from A: $c_{2B} = s + t$.

B has preferences over his own consumption in period 1 and 2, described by the following utility function:

$$\begin{aligned} U_B(s,t) &= \log(c_{1B}) + \beta \log(c_{2B}) \\ &= \log(\omega - s) + \beta \log(s + t), \end{aligned}$$

where $\beta \in (0,1)$ is a fixed parameter. A is altruistic in the sense that she has preferences over both her own consumption and B's utility level U_B. These preferences are described by the following utility function:

$$\begin{aligned} U_A(s,t) &= \log(c_A) + \alpha U_B(s,t) \\ &= \log(\omega - t) + \alpha \log(\omega - s) + \alpha\beta \log(s + t). \end{aligned} \tag{1}$$

Here $\alpha > 0$ is a fixed parameter that represents the altruistic concern of A for the welfare of B.[12] The structure of the model and in particular the individuals' preferences are common knowledge.

2.2 Analysis

The model described in the preceding subsection constitutes an extensive form game. I will solve for the subgame perfect equilibria of this game through backward induction. Let us thus begin by considering A's problem in period 2. A then maximizes $U_A(s,t)$ as given in equation (1) with respect

[11] Hence a non-negativity constraint is imposed on the transfer, $t \geq 0$: A cannot *take* income from B. This assumption seems natural and it is common in the literature. In Section 3 we will see that the non-negativity constraint is crucial for some of the results concerning efficiency in the model with incomplete information.

[12] The important sense in which α represents A's altruistic concern for B is that, for $\alpha > 0$, A puts a positive weight on B's *marginal* utility of consumption, and this weight is increasing in α.

to t subject to the constraint $t \in [0, \omega]$. Denote the solution to this problem by $\widehat{t}$. It is easy to verify that

$$\widehat{t} = \begin{cases} \frac{\alpha\beta\omega - s}{1+\alpha\beta} & \text{for } s \leq \alpha\beta\omega \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that for any $s < \alpha\beta\omega$, $\widehat{t}$ is decreasing in s. That is, if B increases his savings, A will make a smaller transfer to him. One may think of this effect as an implicit tax on savings. In the analysis that follows we shall see that the implicit tax distorts B's saving incentives and typically makes him consume too much in period 1, as compared to what is socially optimal.

Now consider period 1. Anticipating $\widehat{t}$, B chooses s. B's indirect utility is given by

$$U_B\left(s, \widehat{t}\right) = \log\left(\omega - s\right) + \begin{cases} \beta\log\left(\omega + s\right) + \beta\log\left(\frac{\alpha\beta}{1+\alpha\beta}\right) & \text{for } s \leq \alpha\beta\omega \\ \beta\log\left(s\right) & \text{otherwise.} \end{cases} \tag{3}$$

There are two cases to investigate: (i) $\alpha\beta \geq 1$ and (ii) $\alpha\beta < 1$. The analysis of case (ii) is rather cumbersome and is therefore deferred to Lagerlöf (2002). Case (i)—the one where A's concern for B is relatively great—is very straightforward. Since in this case the non-negativity constraint on t is not binding for any $s \in [0, \omega]$, B solves

$$\max_{s\in[0,\omega]} \log\left(\omega - s\right) + \beta\log\left(\omega + s\right) + \beta\log\left(\frac{\alpha\beta}{1+\alpha\beta}\right).$$

By assumption $\beta < 1$; hence this problem has the solution $s^* = 0$. In other words, B saves nothing and relies fully on the anticipated transfer from A. Substituting $s^* = 0$ into equation (2) yields the equilibrium outcome of t, $t^* = \alpha\beta\omega/\left(1 + \alpha\beta\right)$.

Proposition 1 summarizes the results of the analysis above as well as the results for the case $\alpha\beta < 1$ derived in Lagerlöf (2002). Before stating the proposition, however, we must introduce some more notation. Let the function $\varphi\left(\beta\right)$ be defined by

$$\varphi\left(\beta\right) = \left[\left(1+\beta\right)^{\frac{1+\beta}{\beta}} - \beta\right]^{-1}. \tag{4}$$

The expression in (4) is a critical value of α that will be used when describing what the equilibrium outcomes are in different subsets of the parameter space. It can be shown (see Lagerlöf, 2002) that the function φ is decreasing, $\varphi' < 0$. One may also verify that $\lim_{\beta \to 0} \varphi(\beta) = 1/e$ (where $e^{-1} \approx 0.37$), and that $\lim_{\beta \to 1} \varphi(\beta) = 1/3$. Thus, for any $\beta \in (0,1)$, $\varphi(\beta)$ approximately equals one third.

Proposition 1. *For any $\alpha \neq \varphi(\beta)$ there exists a unique subgame perfect equilibrium, the outcome of which is*

$$(s^*, t^*) = \begin{cases} \left(\frac{\beta}{1+\beta}\omega, 0\right) & \text{for } \alpha < \varphi(\beta) \\ \left(0, \frac{\alpha\beta}{1+\alpha\beta}\omega\right) & \text{for } \alpha > \varphi(\beta). \end{cases}$$

For $\alpha = \varphi(\beta)$ there exists a continuum of subgame perfect equilibria. The outcome of any such equilibrium, however, is either $(s^*, t^*) = (\beta\omega/(1+\beta), 0)$ or $(s^*, t^*) = (0, \alpha\beta\omega/(1+\alpha\beta))$.

Fig. 2 illustrates the results stated in the proposition. The critical value of α defined in equation (4), φ, is depicted in the diagram as a function of β. For values of α below this critical value, B saves the fraction $\beta/(1+\beta)$ of his income and A does not make a transfer. For values of α above the critical value, however, the behaviour of A and B is quite different: B saves nothing and A transfers the fraction $\alpha\beta/(1+\alpha\beta)$ of her income to B. For values of α exactly at the critical value, $\alpha = \varphi(\beta)$, B is indifferent between saving nothing and saving the fraction $\beta/(1+\beta)$ of his income, and any randomization between these two choices may be sustained as part of an equilibrium. For any outcome of such a randomization, A will make a transfer to B according to equation (2), i.e., either not make any transfer or transfer the fraction $\alpha\beta/(1+\alpha\beta)$ of her income.

2.3 Efficiency

Resources can be allocated in two dimensions in this model. First, given a value of t, one may reallocate resources intertemporally by varying s. Second, given a value of s, one may reallocate resources inter-individually by varying

t. In this subsection I will ask the question whether the resource allocation induced by a subgame perfect equilibrium is Pareto efficient. The following result is proven in Lagerlöf (2002).

Proposition 2. *Suppose that $\alpha \neq \varphi(\beta)$. Then the allocation induced by the unique subgame perfect equilibrium is Pareto efficient if and only if either $\alpha < \varphi(\beta)$ or $\alpha \geq (1-\beta)^{-1}$. The allocation (s^*, t^*) induced by a subgame perfect equilibrium when $\alpha = \varphi(\beta)$ is Pareto efficient if and only if $(s^*, t^*) = (\beta\omega/(1+\beta), 0)$.*

That is, if A's degree of altruism takes on any value $\alpha \in \left(\varphi(\beta), (1-\beta)^{-1}\right)$, then the equilibrium outcome is not Pareto efficient (cf. Fig. 2). Recall that for these values of α, B saves nothing but receives a transfer $t^* = \alpha\beta\omega/(1+\alpha\beta)$ from A. If B made a ceteris paribus increase in his savings, however, he would be better off. Moreover, since A has altruistic concerns for the welfare of B, this would make also A better off. The reason why B saves too little is the implicit tax on his savings: if B saved more, A would have an incentive to make the transfer smaller. Hence, crucial for the inefficiency result is that A can observe how much B has saved and that she cannot precommit to a transfer level.

If the degree of altruism is either sufficiently low ($\alpha < \varphi(\beta)$) or sufficiently high ($\alpha \geq (1-\beta)^{-1}$), then the equilibrium outcome is Pareto efficient. The intuition for this is straightforward. For $\alpha < \varphi(\beta)$, B does not receive any transfer and must rely only on his own savings. Since A's degree of altruism is relatively small it is, given the level of B's savings, indeed optimal for A not to transfer any income to B. This is anticipated by B, so his saving choice is not distorted. Similarly for the case $\alpha > (1-\beta)^{-1}$. Here, since A cares so much about the welfare of B, it is in *both* A's and B's interest that B does not save anything himself.

3 Efficiency-Enhancing Signalling

3.1 The Model

Let us now add asymmetric information to the model studied in the previous section. In particular, let us assume that B has private information about the exact magnitude of the parameter β and that he learns about this in the beginning of the game.[13] The parameter β may be either "low" (β_L) or "high" (β_H), where $0 < \beta_L < \beta_H < 1$. If $\beta = \beta_L$, then B will be referred to as the "low type"; and if $\beta = \beta_H$, then B will be referred to as the "high type".[14] A places the prior probability $\mu \in (0,1)$ on the event that B is the high type and the prior probability $1-\mu$ on the event that B is the low type. The magnitude of the parameter μ is common knowledge.

All other model features are identical to the model described in Section 2.1. In particular there are two time periods. B lives in both of them while A lives only in period 2. A and B are each endowed with exogenous income $\omega > 0$. In period 1, B first learns his type and then chooses how much of his income to save for period 2, $s_i \in [0,\omega]$ for $i = L, H$; s_L is the amount of savings chosen by the low type and s_H is the amount of savings chosen by the high type. A does not know B's type but observes his actual savings, denoted s. In period 2, A chooses how much of her income to transfer to B, $t(s) \in [0,\omega]$. Denote A's posterior beliefs that B is the high type, on having observed s, by $\widetilde{\mu}(s)$. Moreover, A's and B's utility functions, given B's type i, are given by

$$U_A(s,t \mid \beta_i) = \log(\omega - t) + \alpha \log(\omega - s) + \alpha\beta_i \log(s + t),$$

[13] Hence, the private information does not concern B's second-period income, as was suggested in the Introduction (indeed, in the models described in this section and in Section 2, B does not have a second-period income). This assumption is made for the sake of tractability: it simplifies the analysis considerably if there is uncertainty about a parameter multiplied with—instead of in the argument of—the utility function. Yet one should expect the qualitative results of the analysis to be similar if one instead assumed that B has private information about an exogenous second-period income. Section 4 contains a discussion of what results one should expect if one assumed private information about other parameters than β.

[14] The high type here corresponds to the poor type in the example in Section 1. Recall that there it was in B's interest to be perceived as being poor; here it will be in B's interest to be perceived as having a high β.

$$U_B(s,t \mid \beta_i) = \log(\omega - s) + \beta_i \log(s+t). \tag{5}$$

Again, $\alpha > 0$ is a fixed parameter that represents A's altruistic concern for B.

The equilibrium concept that I will employ is that of perfect Bayesian equilibrium, where this is defined in the usual way: both players must make optimal choices at all information sets given their beliefs, and the beliefs are formed using Bayes' rule when that is defined. Henceforth I will simply write "equilibrium" when I mean perfect Bayesian equilibrium. As my notation indicates I will only consider pure-strategy equilibria. For notational convenience, let us write $t(s_i) = t_i$, and let us denote an outcome of a pure-strategy equilibrium by $(s_L^*, t_L^*, s_H^*, t_H^*)$.

3.2 Analysis

I shall restrict the analysis to the subset of the parameter space satisfying the following assumption.

Assumption 1. $\alpha \in \left(\varphi(\beta_L), (1-\beta_L)^{-1}\right)$.

Imposing Assumption 1 means that we only consider the subset of the parameter space where, for both types, the equilibrium outcome is not Pareto efficient in the corresponding complete-information model (cf. Proposition 2 and Fig. 2). Throughout the remainder of the paper, all the results that are reported presuppose that Assumption 1 holds, even when this is not explicitly stated.[15]

To start with, let us characterize the separating equilibria. The following lemma states that when $\alpha > \varphi(\beta_L)$ the low type chooses not to save.

Lemma 1. *Suppose that $\alpha > \varphi(\beta_L)$. Then in any separating equilibrium $s_L^* = 0$.*

[15] This assumption is natural to make given that we want to explore the possibly efficiency-enhancing effect of the counteracting signalling. We should keep in mind, however, that making this assumption actually gives rise to a bias in favour of smaller inefficiencies since, if we have a first-best outcome without signalling, incomplete information can only make things worse. Hence, if we were to investigate also the subset of the parameter space not satisfying Assumption 1, we would find more of oversaving.

The intuition for this result is straightforward. If possible, and regardless of his true type, B would like to be perceived as the high type in the eyes of A. In any separating equilbrium, however, B's type will, by definition, be revealed. Hence, the best thing the low type can do in such an equilibrium is to behave optimally taking into account that A will know B's type when making her transfer decision. We know from the analysis of the benchmark model that, under Assumption 1, this optimal behaviour is not to save anything.[16]

We are thus looking for an equilibrium with $s_L^* = 0$ and where s_H^* is positive. The analysis will be facilitated by Fig. 3a, which shows the saving-transfer space. The two straight lines in the figure represent A's optimal transfer as given by equation (2); the lower straight line corresponds to A's believing that $\beta = \beta_L$, and the upper one corresponds to A's believing that $\beta = \beta_H$. At values of s larger than $\alpha\beta_L\omega$ respectively $\alpha\beta_H\omega$, the non-negativity constraint on A's transfer is binding and the optimal transfer is zero. The figure also depicts two indifference curves through the point $(s,t) = (0, \alpha\beta_L\omega/(1+\alpha\beta_L))$, one for the low type and one for the high type of B. Hence, these represent the levels of utility associated with the types' choosing the low type's equilibrium amount of savings and receiving the low type's equilibrium transfer.[17] Notice that each type is made better off if one moves northward in the diagram, i.e., if B receives a higher t for any given s. A move in the northwest direction is not necessarily making B better off. It turns out, however, that if one moves northwest along the upper straight line, both types are made better off.

As indicated in the figure, the value of s for which the low type's indifference curve intersects the upper straight line is denoted s', and the value of s for which the high type's indifference curve intersects the same line is denoted s''. The first point of intersection, which will turn out to be the

[16] This logic does not apply to the high type. The reason is that if the high type chose some $s \neq s_H^*$, he would possibly, depending A's beliefs, be perceived as the low type.

[17] One can show that the two types' indifference curves through the point $(s,t) = (0, \alpha\beta_L\omega/(1+\alpha\beta_L))$ must, as drawn in Figure 3a, be strictly concave functions of s, which tend to infinity as s tends to ω. Moreover, at $s = 0$, the slopes of these functions are negative but still larger than the slope of the lower straight line; and, as s approaches ω, the slopes approach infinity. Finally, for any given s, the low type's indifference curve has a larger slope than the high type's (the single-crossing property).

most important one in the subsequent analysis, is implicitly defined by the following identity:

$$\log\left(1-\frac{s^{'}}{\omega}\right)+\beta_L\log\left(1+\frac{s^{'}}{\omega}\right)\equiv\beta_L\log\left(\frac{\beta_L\left(1+\alpha\beta_H\right)}{\beta_H\left(1+\alpha\beta_L\right)}\right). \tag{6}$$

As Fig. 3a is drawn, both $s^{'}$ and $s^{''}$ are located to the left of $\alpha\beta_H\omega$. In the Appendix (Lemma A1 and A2, respectively) it is shown that we always have $s^{'}<\alpha\beta_H\omega$, and that $s^{''}<\alpha\beta_H\omega$ if and only if either (i) $\alpha\beta_H\geq 1$ or (ii) $\alpha\beta_H<1$ and $\alpha>\alpha^*\left(\beta_L,\beta_H\right)$, where $\alpha^*\left(\beta_L,\beta_H\right)$ is implictly defined by

$$\beta_L^{\beta_H}\equiv\beta_H^{\beta_H}\left(1+\beta_L\alpha^*\right)^{\beta_H}\left(1-\beta_H\alpha^*\right). \tag{7}$$

For later use, the following lemma notes some properties of the function α^*.

Lemma 2. $\partial\alpha^*\left(\beta_L,\beta_H\right)/\partial\beta_L<0$, $\lim_{\beta_L\to 0}\alpha^*\left(\beta_L,\beta_H\right)=1/\beta_H$, *and* $\lim_{\beta_L\to\beta_H}\alpha^*\left(\beta_L,\beta_H\right)=0$.

To start with, suppose that either condition (i) or (ii) holds so that indeed $s^{''}<\alpha\beta_H\omega$, and refer again to Fig. 3a. In order to sustain a separating equilibrium there are two necessary conditions. First, the low type must not have an incentive to choose the high type's amount of savings. Second, the high type must not have an incentive to choose the low type's amount of savings. Since A will learn B's type perfectly in any separating equilibrium, A will (after having observed s_L^* or s_H^*) make a transfer according to either one of the two straight lines in the figure. Hence, by mimicking the high type, the low type can get a transfer according to the upper straight line. For the low type not to have an incentive to do this we must have $s_H^*\geq s^{'}$; otherwise the low type could, by saving s_H^*, obtain a saving-transfer pair that gives him a higher utility than he will get if saving only $s_L^*=0$. Similarly, for the high type not to have an incentive to mimic the low type, we must have $s_H^*\leq s^{''}$.

If A's out-of-equilibrium beliefs are chosen appropriately, the two necessary conditions $s_H^*\geq s^{'}$ and $s_H^*\leq s^{''}$ are also sufficient for having $s_L^*=0$ and any $s_H^*\in\left[s^{'},s^{''}\right]$ as part of a separating equilibrium. For instance, one may let (recall that $\widetilde{\mu}$ is A's posterior beliefs that B is the high type)

$$\widetilde{\mu}\left(s\right)=\begin{cases}0 & \text{for } s\in\left[0,s_H^*\right)\\ 1 & \text{for } s\in\left[s_H^*,\omega\right].\end{cases} \tag{8}$$

These posterior beliefs are consistent with the equilibrium requirements and they guarantee that the types do not have an incentive to deviate from s_L^* respectively s_H^*.

The important feature of the beliefs in (8) is that they put a sufficiently high probability on the event that B is the low type whenever $s < s_H^*$; this is what guarantees that the high type does not, in the equilibria in which $s_H^* > s^{'}$, want to deviate to some $s \in \left[s^{'}, s_H^*\right)$. There is, however, a very good reason to regard beliefs that have this feature as unreasonable. Namely, choosing some $s > s^{'}$ is a strictly dominated action for the low type, regardless of which posterior beliefs $\widetilde{\mu}(s) \in [0,1]$ A holds: choosing $s = 0$ always gives him a higher utility (see Fig. 3a). One may plausibly argue that A should realize this and therefore assign zero probability to the event that B is the low type when observing some saving level $s > s^{'}$. But if we accept this argument and require that $\widetilde{\mu}(s) = 1$ for all $s > s^{'}$, then it will be impossible to sustain any $s_H^* > s^{'}$ as the high type's saving level in a separating equilibrium. Below I will refer to this line of reasoning as the "dominance argument".[18] The only equilibrium saving level of the high type that survives the dominance argument is $s_H^* = s^{'}$. Hence, when the parameters are such that $s^{''} < \alpha\beta_H\omega$, the dominance argument gives us a unique separating equilibrium outcome, in which $s_L^* = 0$ and $s_H^* = s^{'}$.

Let us now turn to the case where $s^{''} \geq \alpha\beta_H\omega$ (or, equivalently, the subset of the parameter space where $\alpha\beta_H < 1$ and $\alpha < \alpha^*(\beta_L, \beta_H)$), which is illustrated in Fig. 3b. Here, again, a necessary condition for sustaining a separating equilibrium is that the two types do not have an incentive to mimic each other: s_H^* must be between $s^{'}$ and s°, where s° is the value of s at which the high type's indifference curve crosses the s-axis from below. I will again invoke the dominance argument from above, which means that A's out-of-equilibrium beliefs must put all probability on the event that B is the high type when $s > s^{'}$. As a consequence, for some $s_H^* \in \left[s^{'}, s^{\circ}\right]$ to be part of a separating equilibrium, this saving level and the transfer that it induces must give the high type a utility that is at least as high as what

[18] This equilibrium refinement is implied by (but is weaker than) the so-called intuitive criterion, suggested by Cho and Kreps (1987). The intuitive criterion is very often applied in signalling games. Later I will use it here, too, in order to rule out pooling equilibria.

he would get by choosing any other $s > s'$, given that by choosing any such s he will be recognized as the high type. Hence, if the high type saves some amount greater than $\alpha\beta_H\omega$ in an equilibrium, this amount must be equal to $\beta_H\omega/(1+\beta_H)$, since this is the high type's optimal saving level if not expecting any transfer. Moreover, if the high type saves some amount smaller than or equal to $\alpha\beta_H\omega$ in an equilibrium, this amount must equal s', since the high type's utility increases as one moves northwest along the upper straight line.

Before proceding with the analysis, let us define the following two parameter regimes: Regime I is where $(1+\beta_H)^{-1} \leq \alpha$, and Regime II is where $(1+\beta_H)^{-1} > \alpha$. In Regime I, the high type's optimal saving level if not expecting any transfer, $\beta_H\omega/(1+\beta_H)$, is smaller than or equal to the saving level where the high type's transfer scheme meets the horizontal axis, $\alpha\beta_H\omega$. It follows from this and the arguments in the previous paragraph that in Regime I only $s_H^* = s'$ can be part of a separating equilibrium that survives the dominance argument. In Regime II, $\beta_H\omega/(1+\beta_H) > \alpha\beta_H\omega$.[19] Whether we in Regime II can sustain $s_H^* = s'$ or $s_H^* = \beta_H\omega/(1+\beta_H)$ as part of a separating equilibrium that survives the dominance argument thus amounts to asking which of these saving levels gives the high type the highest utility, given that A will correctly infer B's type and transfer resources to him accordingly. Hence, we need to know the sign of ΔU, where

$$\Delta U \equiv U_B\left(s', \frac{\alpha\beta_H\omega - s'}{1+\alpha\beta_H} \mid \beta_H\right) - U_B\left(\frac{\beta_H\omega}{1+\beta_H}, 0 \mid \beta_H\right).$$

Given that the size of the transfer is increasing in α, one would expect that ΔU is positive if and only if α exceeds some threshold. Lemma 3 below confirms this.

[19] One also has that $\beta_H\omega/(1+\beta_H) < s^\circ$. To see this, notice that the high type's indifference curve through the point $(s,t) = (\beta_H\omega/(1+\beta_H), 0)$ must be tangent to the s-axis at $s = \beta_H\omega/(1+\beta_H)$ (cf. Fig. 3b); otherwise $s = \beta_H\omega/(1+\beta_H)$ would not maximize the high type's utility given that he is not expecting a transfer. Moreover, this indifference curve can of course not cross the one of the high type that is drawn in Fig. 3b. It follows that $s^\circ > \beta_H\omega/(1+\beta_H)$.

Lemma 3. *There exists a function* $\alpha^{**}(\beta_L, \beta_H, \omega)$*, implicitly defined by*

$$\begin{aligned} & \log\left(1 - \frac{s'(\alpha^{**})}{\omega}\right) + \beta_H \log\left(1 + \frac{s'(\alpha^{**})}{\omega}\right) \\ \equiv \quad & \beta_H \log\left(\frac{1 + \alpha^{**}\beta_H}{\alpha^{**}}\right) - (1 + \beta_H)\log(1 + \beta_H) \end{aligned}$$

(where $s'(\alpha^{**})$ *is* s' *evaluated at* α^{**}*), such that* $\Delta U \lesseqgtr 0$ *as* $\alpha \lesseqgtr \alpha^{**}(\beta_L, \beta_H, \omega)$*. Moreover,* $\alpha^{**}(\beta_L, \beta_H, \omega) > \varphi(\beta_H)$ *for all* $\beta_L \in (0, \beta_H)$ *and all* $\omega > 0$*, and*

$$\lim_{\beta_L \to 0} \alpha^{**}(\beta_L, \beta_H, \omega) = \lim_{\beta_L \to \beta_H} \alpha^{**}(\beta_L, \beta_H, \omega) = \varphi(\beta_H). \tag{9}$$

Hence, summing up the analysis so far, if (i) the parameters are such that $s'' \geq \alpha\beta_H\omega$, (ii) we are in Regime II, and (iii) $\alpha < \alpha^{**}(\beta_L, \beta_H, \omega)$, then in any separating equilibrium that survives the dominance argument we have $s_H^* = \beta_H\omega/(1 + \beta_H)$. These three conditions are met if and only if $\alpha < \widehat{\alpha}(\beta_L, \beta_H, \omega)$, where

$$\widehat{\alpha}(\beta_L, \beta_H, \omega) \equiv \min\left\{\alpha^*(\beta_L, \beta_H), \alpha^{**}(\beta_L, \beta_H, \omega), \frac{1}{1 + \beta_H}\right\}.$$

If, on the other hand, at least one of the three conditions is violated (i.e., if $\alpha > \widehat{\alpha}(\beta_L, \beta_H, \omega)$), then in any separating equilibrium that survives the dominance argument we have $s_H^* = s'$.

We have not yet considered the existence of pooling equilibria. For the sake of brevity, this will not be done here. In Lagerlöf (2002), however, I show that the intuitive criterion, (Cho and Kreps, 1987), rule out all pooling equilibria. As explained in footnote 18, this equilibrium refinement implies the dominance argument used above, although it is stronger. Since the intuitive criterion rules out all pooling equilibria, it gives us a unique equilibrium outcome.

Proposition 3. *Suppose* $\alpha \neq \widehat{\alpha}(\beta_L, \beta_H, \omega)$*. Then there is a unique equilibrium outcome that satisfies the intuitive criterion. If* $\alpha > \widehat{\alpha}(\beta_L, \beta_H, \omega)$*, then this outcome is* $(s_L^*, t_L^*, s_H^*, t_H^*) = \left(0, \frac{\alpha\beta_L\omega}{1+\alpha\beta_L}, s', \frac{\alpha\beta_H\omega - s'}{1+\alpha\beta_H}\right)$*; and if* $\widehat{\alpha}(\beta_L, \beta_H, \omega) < 0$*, this outcome is* $(s_L^*, t_L^*, s_H^*, t_H^*) = \left(0, \frac{\alpha\beta_L\omega}{1+\alpha\beta_L}, \frac{\beta_H\omega}{1+\beta_H}, 0\right)$.

In Fig. 4, the results are illustrated in a diagram that depicts α on the vertical and β_L on the horizontal axis. Recall that Assumption 1 is satified in the region above the graph of $\varphi(\beta_L)$ and below the graph of $(1-\beta_L)^{-1}$, where $\varphi(\beta_L)$ is slightly downwardsloping, starts at $1/e$, and ends at $\varphi(\beta_H)$. The figure also shows the graphs of $\alpha^*(\cdot, \beta_H)$, $\alpha^{**}(\cdot, \beta_H, \omega)$, and $(1+\beta_L)^{-1}$ (in drawing these graphs, I made use of Lemmas 2 and 3). Proposition 3 tells us that when we are in the region below all these three graphs (the shadowed area in the figure), i.e., when $\alpha < \widehat{\alpha}(\beta_L, \beta_H, \omega)$, the unique equilibrium outcome has the high type saving $\beta_H \omega (1+\beta_H)^{-1}$. Otherwise, in the region where $\alpha > \widehat{\alpha}(\beta_L, \beta_H, \omega)$, the unique equilibrium outcome has the high type saving s'. An important question for the subsequent analysis will be whether the shadowed area is non-empty, as presumed in Fig. 4. Although this paper cannot offer any analytical result that answers this question, plotting the relevant graphs for various parameter values with the help of the software Mathematica indicates that the area is indeed always non-empty.

3.3 Efficiency

Let us now investigate whether the unique equilibrium outcome that we derived above is efficient and, if not, whether B saves too little or too much. In a game with incomplete information the concept of efficiency is not straightforward. Here I will use the following definitions. Following Holmström and Myerson (1983) I say that an outcome (s_L, t_L, s_H, t_H) is *incentive feasible* if

$$U_B(s_i, t_i \mid \beta_i) \geq U_B(s_j, t_j \mid \beta_i), \qquad \forall i, j \in \{L, H\} \text{ with } i \neq j$$

and $(s_L, t_L, s_H, t_H) \in [0, \omega]^4$. An outcome $\left(s_L', t_L', s_H', t_H'\right)$ *ex post dominates* an outcome (s_L, t_L, s_H, t_H) if

$$U_i\left(s_j', t_j' \mid \beta_j\right) \geq U_i\left(s_j, t_j \mid \beta_j\right), \qquad \forall j \in \{L, H\} \text{ and } \forall i \in \{A, B\}, \quad (10)$$

with at least one strict inequality. And an outcome (s_L, t_L, s_H, t_H) is *ex post incentive efficient* if there is no other incentive feasible outcome that ex post dominates (s_L, t_L, s_H, t_H).

It turns out that no outcome of a separating equilibrium can be ex post incentive efficient. This is because in a separating equilibrium the signalling

mechanism does not affect the low type's saving choice (cf. Lemma 1). The high type's choice, however, is distorted upwards in a separating equilibrium: $s_H^* > 0$; thus, it is conceivable that s_H^* is part of an outcome that is "efficient for the high type". In the following I will investigate if and when the high type's saving level in the unique equilibrium outcome indeed is "efficient for the high type". Formally, I say that an outcome $\left(s_L, t_L, s_H^{'}, t_H^{'}\right)$ *ex post dominates* an outcome (s_L, t_L, s_H, t_H) *for the high type* if (10) is satisfied for $j = H$ with at least one strict inequality. And an allocation (s_L, t_L, s_H, t_H) is *ex post incentive efficient for the high type* if there is no other incentive feasible outcome that ex post dominates this outcome for the high type.

There are two conditions that are necessary for $(s_L^*, t_L^*, s_H^*, t_H^*)$ to be ex post incentive efficient for the high type as well as the outcome of a separating equilibrium:

$$s_H^* \in \arg\max_{s\in[0,\omega]} U_B\left(s, t_H^* \mid \beta_H\right), \tag{11}$$

$$t_H^* \in \arg\max_{t\in[0,\omega]} U_A\left(s_H^*, t \mid \beta_H\right). \tag{12}$$

The first condition guarantees that the intertemporal allocation of resources is efficient, and the second condition is necessary for $(s_L^*, t_L^*, s_H^*, t_H^*)$ to indeed be the outcome of a separating equilibrium. Straightforward algebra shows that, in Regime I, conditions (11) and (12) are met only for $(s_H^*, t_H^*) = \left(\widetilde{s}_I, \widetilde{t}_I\right)$, where

$$\left(\widetilde{s}_I, \widetilde{t}_I\right) \equiv \left(\frac{1-\alpha\left(1-\beta_H\right)}{1+\alpha\left(1+\beta_H\right)}\omega, \frac{\alpha\left(1+\beta_H\right)-1}{1+\alpha\left(1+\beta_H\right)}\omega\right).$$

Similarly, in Regime II conditions (11) and (12) are met only for $(s_H^*, t_H^*) = \left(\beta_H\omega/\left(1+\beta_H\right), 0\right) \equiv \left(\widetilde{s}_{II}, \widetilde{t}_{II}\right)$. One can verify that, in Regime I, $\left(s_L^*, t_L^*, \widetilde{s}_I, \widetilde{t}_I\right)$ is indeed ex post incentive efficient for the high type; and, in Regime II, $\left(s_L^*, t_L^*, \widetilde{s}_{II}, \widetilde{t}_{II}\right)$ is indeed ex post incentive efficient for the high type.

Recall from Proposition 3 that in Regime I the high type saves $s^{'}$. Thus, in this regime the high type will save the efficient amount only in the special case where $s^{'} = \widetilde{s}_I$; if $s^{'} < \widetilde{s}_I$ then the high type saves too little, and if $s^{'} > \widetilde{s}_I$ he saves too much. So how does $s^{'}$ relate to $\widetilde{s}_I$? Lemma 4 below tells us that

this depends on how α relates to a cut-off value $\alpha^{***}(\beta_L, \beta_H)$. The lemma also notes some useful properties of this cut-off value.

Lemma 4. *There exists a function $\alpha^{***}(\beta_L, \beta_H)$, implicitly defined by*

$$\log\left[\frac{1+(1+\beta_H)\,\alpha^{***}}{2\alpha^{***}}\right]+\beta_L\log\left[\frac{\beta_L\left[1+(1+\beta_H)\,\alpha^{***}\right]}{2\beta_H\left(1+\beta_L\alpha^{***}\right)}\right]\equiv 0, \quad (13)$$

such that $s' \lessgtr \widetilde{s}_I$ as $\alpha \lessgtr \alpha^{***}(\beta_L, \beta_H)$. Moreover, $\lim_{\beta_L\to 0}\alpha^{***}(\beta_L, \beta_H) = \lim_{\beta_L\to\beta_H}\alpha^{***}(\beta_L, \beta_H) = (1-\beta_H)^{-1}$; and, for β_L sufficiently close to β_H, $\alpha^{***}(\beta_L, \beta_H) < (1-\beta_L)^{-1}$.

It remains to consider Regime II. It follows immediately from Proposition 3 and the algebra above that, if $\alpha < \widehat{\alpha}(\beta_L, \beta_H, \omega)$, in Regime II the equilibrium outcome $(s_L^*, t_L^*, s_H^*, t_H^*)$ is indeed ex post incentive efficient for the high type. For $\alpha > \widehat{\alpha}(\beta_L, \beta_H, \omega)$, however, the high type will undersave, since $s' < \widetilde{s}_{II}$.

Summing up, we have the following proposition.

Proposition 4. *The unique equilibrium outcome is ex post incentive efficient for the high type if $\alpha < \widehat{\alpha}(\beta_L, \beta_H, \omega)$ or $\alpha = \alpha^{***}(\beta_L, \beta_H)$; it involves undersaving on the part of the high type if $\widehat{\alpha}(\beta_L, \beta_H, \omega) < \alpha < \alpha^{***}(\beta_L, \beta_H)$; and it involves oversaving on the part of the high type if $\alpha > \alpha^{***}(\beta_L, \beta_H)$.*

Fig. 5 illustrates the results stated in the proposition. This figure is similar to Fig. 4, but it also shows the graph of $\alpha^{***}(\cdot, \beta_H)$. From Lemma 4 we know that, at least if β_L is sufficiently close to β_H, this graph goes through the region of the parameter space where Assumption 1 is satisfied. Hence, there is a non-empty region of the parameter space (namely, the checked area in the figure) where the high type saves too much![20] Proposition 4 also tells us that on the lower border of this region, i.e., where $\alpha = \alpha^{***}(\beta_L, \beta_H)$ and where Assumption 1 is satsified, the unique equilibrium outcome is indeed ex post

[20] In Lagerlöf (2002) it is shown that, for all β_L and β_H, $\alpha^{***}(\beta_L, \beta_H) > 1$; hence, in order to obtain the result that the high type oversaves, A must care more about B than about herself.

incentive efficient for the high type, although this is of course a knife-edge phenomenon. In the region where $\alpha < \widehat{\alpha}(\beta_L, \beta_H, \omega)$ (the shadowed area in the figure), however, we have generically that the unique equilibrium outcome is indeed ex post incentive efficient for the high type. In the intermediate part of the parameter space, where $\alpha < \alpha^{***}(\beta_L, \beta_H)$ and $\alpha > \widehat{\alpha}(\beta_L, \beta_H, \omega)$, the high type will undersave. Yet, even here, the high type's saving choice will be distorted upwards: he will save more than he would have done under complete information.

Of particular interest among these results is the one saying that for a subset of the parameter space the high type saves *exactly* the efficient amount. What is the logic behind this? To see this, notice that here as in other signalling games the intuitive criterion will select the separating equilibrium in which the high type separates at the lowest possible cost. If we had not excluded negative transfers, the cost would always had been minimized at the lowest possible amount of savings, s'. The presence of the non-negativity constraint, however, creates a convexity in the transfer schedule (cf. Figs. 3a and 3b), which might make the high type prefer to save an amount somewhere on the horizontal part of the schedule; if so, the high type's saving choice will not be distorted, since he does not expect any transfer.

Clearly, B is more likely to be better off by saving this larger amount instead of only s' when A's degree of altruism is relatively low, since then the transfer associated with s' will be relatively small and thus less attractive for him. What is not that obvious, though, is whether Assumption 1 will still be met for such a low degree of altruism. The algebra (and the computer plottings referred to in Section 3.2), however, tell us that there are indeed parameters such that the high type saves the efficient amount and Assumption 1 is satisfied, although this cannot happen when β_L is very close to zero or β_H, as is seen from Fig. 5 (and proven in Lemma 3).

4 Concluding Discussion

The Samaritan's dilemma—i.e., the idea that, in the presence of altruism, people may choose to save (or work or insure) to a too small extent—certainly appeals to our intuition. In the alternative formulation of the Samaritan's

dilemma considered in this paper there is an additional effect present, which counteracts the undersaving effect. The logic of this new force should also, once we have become aware of it, be very intuitive. For the new force to indeed work in the "right" direction (i.e., to counteract the undersaving effect), the following condition must hold. Suppose that B has private information about some parameter x and that he, everything else being equal, has an incentive to save *more* when knowing that x is high (respectively, low). Then, believing that x is high (respectively, low) will induce A to make the transfer to B *larger*. Since B wants the transfer to be large, he would like A to believe that x is high (respectively, low); and he can try to make A believe this by saving more.

The condition is met if, as was suggested in the Introduction, x is a measure of B's second-period income or if, as was assumed in the formal model of Section 3, x is a discount factor or a weight on B's second-period utility. One may wonder whether the presence of the counteracting effect is hinging on the assumption that the incomplete information concerns one of these two particular characteristics of B. What if B had private information about the return on his savings or about his first-period income?

If the parameter x is interpreted as the return on B's savings and if we stick to the log-utility specification in the present paper, then it is clear that there would not be any counteracting force present. This is because with log-utility and with A's transfer t being equal to zero, the optimal saving level is independent of the return; and if t is positive, then the optimal saving level is increasing with the return. Yet if the intertemporal elasticity of substitution is constant but sufficiently less than one (or, equivalently, if the degree of relative risk aversion is sufficiently greater than one), then B will have an incentive to save more when knowing that the return is low. And A will of course have an incentive to make her transfer larger when believing that the return is low. Hence, under this assumption, one would again get efficiency-enhancing signalling. The assumption about the intertemporal elasticity of substitution seems reasonable: the log-utility assumption in the present paper was made for the sake of tractability, and there is empirical evidence that this elasticity is indeed less than one.

Private information about B's first-period income does of course not give rise to any opportunity to signal as long as B's utility function is additively separable over time, since then the size of B's income in the first period does not affect A's incentive to transfer income to him in the second period. But if B's marginal utility of second-period consumption is increasing with B's first-period consumption, then the condition above is again satisfied. This requirement on the sign of the cross derivative of the utility function is, for instance, met for the following preferences: $U_B(c_{1B}, c_{2B}) = (c_{1B})^a (c_{2B})^b$ for some $a, b > 0$.

Yet another parameter in the model that there could conceivably be uncertainty about is the altruism parameter, α.[21] In the model analyzed in this paper, A's having private information about α would not give rise to any signalling, since A is acting last in the game. Yet this is not true for the formulation of the Samaritan's dilemma considered in Lindbeck and Weibull (1988). In that model there are two individuals who are altruistic towards each other. They both, simultaneously, make a saving decision in period one. In period two they observe the other one's saving decision and then, simultaneously, decide how much (if anything) to transfer to each other. If they both are equally wealthy, then, in equilibrium, only the individual who is more altruistic will make a positive transfer. Anticipating this, the less altruistic individual will undersave in the first period. If one to this setting added the assumption that one or both of the individuals have incomplete information about the other one's degree of altruism, then one should expect the undersaving to be exacerbated, the reason being that both individuals would like to signal that they are less altrusitic than the other one, and a person whose degree of altruism is indeed low should expect a transfer from the other and will therefore save less on his own.

One particularly interesting result of the present paper says that for a subset of the parameter space the high type of B saves *exactly* the efficient amount. Crucial for this result is the assumption that the transfer cannot be negative, which is indeed both natural and standard in the literature. Given the logic that leads up to the result (see the discussion after Proposition

[21] Uncertainty about the degree of altruism has been modelled by Chakrabarti, Lord, and Rangazas (1993) and Lord and Rangazas (1995).

4), it is clear that it is robust: it would, for instance, hold also for other utility functions, as long as these are not too different from the log-utility specification assumed here. Exactly how and in what direction things would change if one instead assumed, for example, a utility function with a constant but not necessarily unitary intertemporal elasticity of substitution is harder to say, and this question is left for future work.

As noted in the Introduction, the logic of the traditional Samaritan's-dilemma model and in particular the undersaving result has been employed in an extensive literature, addressing various issues. Although these models are not identical to the benchmark model of the present paper, the basic logic is the same. Hence, one should expect the undersaving result also in those other models to be sensitive to the assumption that information is complete. An interesting topic for future research would be to investigate the signalling mechanism in the present paper in a setting that is closer to the ones in the existing literature, in order to find out to what extent those results indeed are sensitive to the complete-information assumption.

Appendix

Proof of Lemma 1: Suppose that $\alpha > \varphi(\beta_L)$ and that $s_L^* > 0$ in a separating equilibrium. I will show that this leads to a contradiction. To start with, consider the case where $s_L^* \in (0, \alpha\beta_L\omega]$. Then the low type receives a transfer from A according to the first line in equation (2) but with β_L substituted for β. The low type's utility is accordingly given by (cf. the first line of equation (3)):

$$V(s_L^*) = \log(\omega - s_L^*) + \beta_L \log(\omega + s_L^*) + \beta_L \log\left(\frac{\alpha\beta_L}{1+\alpha\beta_L}\right).$$

If the low type instead chose $s = 0$, however, he would receive a transfer of at least $\alpha\beta_L\omega/(1+\alpha\beta_L)$, which would give him a utility of $V(0)$. This utility level is strictly greater than $V(s_L^*)$ for all $s_L^* \in (0, \alpha\beta_L\omega]$, since $V(s_L^*)$ is strictly decreasing in s_L^*. Now consider the case where $s_L^* \in (\alpha\beta_L\omega, \omega]$. Then the low type receives a transfer from A according to the second line in equation (2) (i.e., a zero transfer). However, since $\alpha > \varphi(\beta_L)$, the low type is

strictly better off from choosing $s = 0$ than from choosing any $s_L^* \in (\alpha\beta_L\omega, \omega]$. This follows from the proof of Proposition 1 (see Lagerlöf, 2002). We thus have a contradiction, which proves the lemma. □

Lemma A1. *$s' < \alpha\beta_H\omega$.*

Proof of Lemma A1: Since we must have $s' < \omega$, it is obvious that the lemma is true for $\alpha\beta_H \geq 1$. Suppose that $\alpha\beta_H < 1$. Then $s' < \alpha\beta_H\omega$ if and only if the left-hand side of (6) evaluated at $s' = \alpha\beta_H\omega$ is strictly smaller than the righ-hand side. This condition can be written as

$$\log(1 - \alpha\beta_H) + \beta_L \log\left(\frac{\beta_H(1 + \alpha\beta_L)}{\beta_L}\right) \equiv \Upsilon(\alpha, \beta_L, \beta_H) < 0.$$

It is readily verified that Υ is strictly concave in β_H, and that $\partial\Upsilon(\alpha, \beta_L, \beta_H)/\partial\beta_H = 0 \Leftrightarrow \beta_H = \beta_L/[\alpha(1 + \beta_L)]$. Hence, if the above condition is satisfied for $\beta_H = \beta_L/[\alpha(1 + \beta_L)]$, then it is always satisfied. Moreover, Υ is strictly decreasing in α. It thus suffices to show that evaluated at $\alpha = \varphi(\beta_L)$ and $\beta_H = \beta_L[\varphi(\beta_L)(1 + \beta_L)]^{-1}$, $\Upsilon(\alpha, \beta_L, \beta_H) \leq 0$. Straightforward algebra, however, shows that $\Upsilon\left(\varphi(\beta_L), \beta_L, \beta_L[\varphi(\beta_L)(1 + \beta_L)]^{-1}\right) = 0$. □

Lemma A2. *$s'' < \alpha\beta_H\omega$ if and only if either (i) $\alpha\beta_H \geq 1$ or (ii) $\alpha\beta_H < 1$ and $\alpha > \alpha^*(\beta_L, \beta_H)$.*

Proof of Lemma A2: Let us first state the formal definition of s'':

$$\log\left(1 - \frac{s''}{\omega}\right) + \beta_H \log\left(1 + \frac{s''}{\omega}\right) \equiv \beta_H \log\left(\frac{\beta_L(1 + \alpha\beta_H)}{\beta_H(1 + \alpha\beta_L)}\right). \qquad \text{(A1)}$$

If $\alpha\beta_H \geq 1$, then it is obvious that $s'' < \alpha\beta_H\omega$. Suppose that $\alpha\beta_H < 1$. Then $s'' < \alpha\beta_H\omega$ if and only if the left-hand side of (A1) evaluated at $s'' = \alpha\beta_H\omega$ is strictly smaller than the right-hand side. Making this substitution and rewriting yield

$$\log(1 - \alpha\beta_H) + \beta_H \log(1 + \alpha\beta_L) < \beta_H \log\left(\frac{\beta_L}{\beta_H}\right).$$

This inequality is satisfied for α's close to β_H^{-1} and it is not for $\alpha = 0$. Moreover, it is readily verified that the left-hand side is strictly decreasing in α for $\alpha\beta_H < 1$. Hence, the cut-off value α^* is well-defined, and $s'' < \alpha\beta_H\omega$ if and only if $\alpha > \alpha^*(\beta_L, \beta_H)$. The identity in (7) that defines α^* is obtained by changing the above inequality to an equality and re-arranging. □

Proof of Lemma 2: The result that $\partial\alpha^*(\beta_L, \beta_H)/\partial\beta_L < 0$ follows immediately from the fact that s'' is strictly decreasing in α; this, in turn, can be seen from (A1). The results that $\lim_{\beta_L \to 0} \alpha^*(\beta_L, \beta_H) = 1/\beta_H$ and $\lim_{\beta_L \to \beta_H} \alpha^*(\beta_L, \beta_H) = 0$ follow from substituting $\beta_L = 0$ respectively $\beta_L = \beta_H$ into (7). □

Proof of Lemma 3: By using the assumed functional form for $U_B(s, t \mid \beta_H)$ (see (5)), we can write

$$\Delta U = \log\left(1 - \frac{s'}{\omega}\right) + \beta_H \log\left(1 + \frac{s'}{\omega}\right) - \beta_H \log\left(\frac{1 + \alpha\beta_H}{\alpha}\right) + (1 + \beta_H)\log(1 + \beta_H).$$

Notice that ΔU is strictly decreasing in s' and, keeping s' fixed, strictly increasing in α. Moreover, s' is a function of α with $\partial s'/\partial\alpha < 0$. Hence, $\partial\Delta U/\partial\alpha > 0$. Also, it can easily be verified that if evaluating ΔU at $\alpha = \varphi(\beta_H)$, the two last terms of ΔU vanish; thus, $\Delta U \mid_{\alpha=\varphi(\beta_H)} < 0$. Moreover, since $s' \to 0$ as $\alpha \to \infty$, for large enough α's $\Delta U > 0$. It follows that the threshold α^{**} is well defined with $\Delta U \lessgtr 0$ as $\alpha \lessgtr \alpha^{**}(\beta_L, \beta_H, \omega)$, and that $\alpha^{**}(\beta_L, \beta_H, \omega) > \varphi(\beta_H)$ for all $\beta_L \in (0, \beta_H)$ and all $\omega > 0$. Moreover, since $s' \to 0$ as $\beta_L \to 0$ and as $\beta_L \to \beta_H$, we get the equalities in (9). □

Proof of Lemma 4: By using the definition of s', one can show that $\widetilde{s}_I \geq s'$ is equivalent to

$$\log\left[\frac{1 + \alpha(1 + \beta_H)}{2\alpha}\right] + \beta_L \log\left[\frac{\beta_L[1 + \alpha(1 + \beta_H)]}{2\beta_H(1 + \alpha\beta_L)}\right] \geq 0. \qquad \text{(A2)}$$

Clearly, inequality (A2) is satisfied if α is sufficiently close to zero. Moreover, the upper constraint on α in Assumption 1, $(1 - \beta_L)^{-1}$, is strictly smaller

than $[\beta_L(\beta_H-\beta_L)]^{-1}$. To prove the first claim of the lemma, it thus suffices to show that (i) the left-hand side of (A2) is strictly decreasing in α for all $\alpha \in \left(0, [\beta_L(\beta_H-\beta_L)]^{-1}\right)$ and (ii) it does not hold for $\alpha = [\beta_L(\beta_H-\beta_L)]^{-1}$. To establish (i), differentiate the left-hand side of (A2) with respect to α; the resulting expression has the same sign as $[\alpha\beta_L(\beta_H-\beta_L)-1]$, which clearly is strictly negative for all $\alpha < [\beta_L(\beta_H-\beta_L)]^{-1}$. To establish (ii), substitute $\alpha = [\beta_L(\beta_H-\beta_L)]^{-1}$ into the left-hand side of (A2). This yields

$$(1+\beta_L)\log\left[\frac{\beta_L(\beta_H-\beta_L)+1+\beta_H}{2}\right]-\beta_L\log\left[(\beta_H-\beta_L+1)\beta_H\right] \equiv g(\beta_L,\beta_H).$$

One can show that $g(\beta_L,1) < 0$ for all $\beta_L \in (0,1)$ (one has $g(0,1) = g(1,1) = 0$, and $g''_{11}(\beta_L,1) > 0$ whenever $g'_1(\beta_L,1) = 0$). Moreover, $g(\beta_L,\beta_H)$ is increasing in β_H. To see this, note that we can write $g'_2(\beta_L,\beta_H) > 0 \Leftrightarrow$

$$(1+\beta_L)^2(\beta_H-\beta_L+1)\beta_H-(2\beta_H+1-\beta_L)\left[\beta_L(\beta_H-\beta_L)+1+\beta_H\right]\beta_L > 0,$$

the left-hand side of which is increasing in β_H and zero evaluated at $\beta_H = \beta_L$. It follows that the threshold α^{***} is well defined with $s' \gtreqless \widetilde{s}_I$ as $\alpha \gtreqless \alpha^{***}(\beta_L,\beta_H)$. Moreover, by substituting $\beta_L = \beta_H$ and $\alpha^{***} = (1-\beta_H)^{-1}$ into (13), one can verify that $\lim_{\beta_L\to\beta_H}\alpha^{***}(\beta_L,\beta_H) = (1-\beta_H)^{-1}$. Similarly with the claim that $\lim_{\beta_L\to 0}\alpha^{***}(\beta_L,\beta_H) = (1-\beta_H)^{-1}$, although here one must also make use of the fact that $\lim_{\beta_L\to 0}\beta_L\log(\beta_L) = 0$. Let us finally show that for β_L sufficiently close to β_H, $\alpha^{***}(\beta_L,\beta_H) < (1-\beta_L)^{-1}$. Since $\lim_{\beta_L\to\beta_H}\alpha^{***}(\beta_L,\beta_H) = (1-\beta_H)^{-1}$, it suffices to show that

$$\lim_{\beta_L\to\beta_H}\frac{\partial\alpha^{***}(\beta_L,\beta_H)}{\partial\beta_L} > \lim_{\beta_L\to\beta_H}\frac{\partial(1-\beta_L)^{-1}}{\partial\beta_L} = (1-\beta_H)^{-2}.$$

Straightforward calculations yield

$$\frac{\partial\alpha^{***}(\beta_L,\beta_H)}{\partial\beta_L} = -\frac{\partial\left[LHS(13)\right]/\partial\beta_L}{\partial\left[LHS(13)\right]/\partial\alpha} = \frac{\log\left[\frac{\beta_L[1+\alpha(1+\beta_H)]}{2\beta_H(1+\alpha\beta_L)}\right]+(1+\alpha\beta_L)^{-1}}{\frac{1-\alpha\beta_L(\beta_H-\beta_L)}{\alpha(1+\alpha\beta_L)[1+\alpha(1+\beta_H)]}}.$$

Hence, using $\lim_{\beta_L\to\beta_H}\alpha^{***}(\beta_L,\beta_H) = (1-\beta_H)^{-1}$, one has

$$\lim_{\beta_L\to\beta_H}\frac{\partial\alpha^{***}(\beta_L,\beta_H)}{\partial\beta_L} = \frac{2}{(1-\beta_H)^2} > (1-\beta_H)^{-2},$$

which always holds. □

References

Becker, Gary. S., (1974). 'A theory of social interactions', *Journal of Political Economy*, vol. 82, no. 6, pp. 1063-1093.

Becker, Gary. S. and Murphy, Kevin M., (1988). 'The family and the state', *Journal of Law and Economics*, vol. 31 (April), pp. 1-18.

Bergstrom, Theodore. C., (1989). 'A fresh look at the Rotten Kid Theorem—and other household mysteries', *Journal of Political Economy*, vol. 97, no. 5, pp. 1138-1159.

Bernheim, B. Douglas and Stark, Oded, (1988). 'Altruism within the family reconsidered: Do nice guys finish last?', *American Economic Review*, vol. 78, no. 5, pp. 1034-1045.

Bruce, Neil and Waldman, Michael, (1990). 'The Rotten-Kid Theorem meets the Samaritan's dilemma', *Quarterly Journal of Economics*, vol. 105 (February), pp. 155-165.

Bruce, Neil and Waldman, Michael, (1991). 'Transfers in kind: Why they can be efficient and nonpaternalistic', *American Economic Review*, vol 81, no. 5, pp. 1345-1351.

Buchanan, James M., (1975). The Samaritan's dilemma, in (Phelps, E. S. ed.) *Altruism, morality and economic theory*, New York: Russel Sage Foundation, pp. 71-85.

Chakrabarti, Subir, Lord, William and Rangazas, Peter, (1993). 'Uncertain altruism and investment in children', *American Economic Review*, vol. 83, no. 4, pp. 994-1002.

Cho, In-Koo and Kreps, David M., (1987). 'Signaling games and stable equilibria', *Quarterly Journal of Economics*, vol. 102 (May), pp. 179-221.

Coate, Stephen, (1995). 'Altruism, the Samaritan's dilemma, and government transfer policy', *American Economic Review*, vol. 85, no. 1, pp. 46-57.

Goodfray, H. Charles. J. and Johnstone, Rufus A., (2000). 'Begging and bleating: The evolution of parent-offspring signalling', *Philosophical Transactions of the Royal Society of London, Series B (Biological Sciences)*, vol. 355, no. 1403, pp. 1581-1591.

Grafen, Alan (1990). 'Biological signals as handicaps', *Journal of Theoretical Biology*, vol. 144, no. 4, pp. 517-546.

Hansson, Ingemar and Stuart, Charles, (1991). 'Social security as trade among living generations', *American Economic Review*, vol. 79, no. 5, pp. 1182-1195.

Holmström, Bengt and Myerson, Roger B., (1983). 'Efficient and durable decision rules with incomplete information', *Econometrica*, vol. 51, no. 6, pp. 1799-1819.

Kotlikoff, Laurence J., (1987). 'Justifying public provision of social security', *Journal of Policy Analysis and Management*, vol 6, no. 4, pp. 674-689.

Lagerlöf, Johan, (2000). 'Incomplete information in the Samaritan's dilemma: The dilemma (almost) vanishes', WZB Discussion Paper No. FS IV 99-12, revised version June 2000, Social Science Research Center Berlin (WZB).

Lagerlöf, Johan, (2002). 'Supplementary material to "Efficiency-enhancing signalling in the Samaritan's dilemma"', mimeo, Social Science Research Center Berlin (WZB).

Lindbeck, Assar and Weibull, Jörgen W., (1988). 'Altruism and time inconsistency: The economics of fait accompli', *Journal of Political Economy*, vol. 96, no. 6, pp. 1165-1182.

Maynard Smith, John, (1991). 'Honest signalling: The Philip Sidney game', *Animal Behaviour*, vol. 42, no. 6, pp. 1034-1035.

O'Connell, Stephen A. and Zeldes, Stephen P., (1993). 'Dynamic efficiency in the gifts economy', Journal of Monetary Economics, vol. 31, no. 3, pp. 363-379.

Spence, Michael, (1973). 'Job market signaling', *Quarterly Journal of Economics*, vol. 87 (August), pp. 355-374.

Thompson, Earl A., (1980). 'Charity and nonprofit organizations', in (Clarkson, K. and Martin, D., eds.) *Economics of Nonproprietory Organizations*, Greenwich, CT: JAI Press, Inc, pp. 125-138.

Veall, Michael R., (1986). 'Public pensions as optimal social contracts', Journal of Public Economics, vol. 31, no. 2, pp. 237-251.

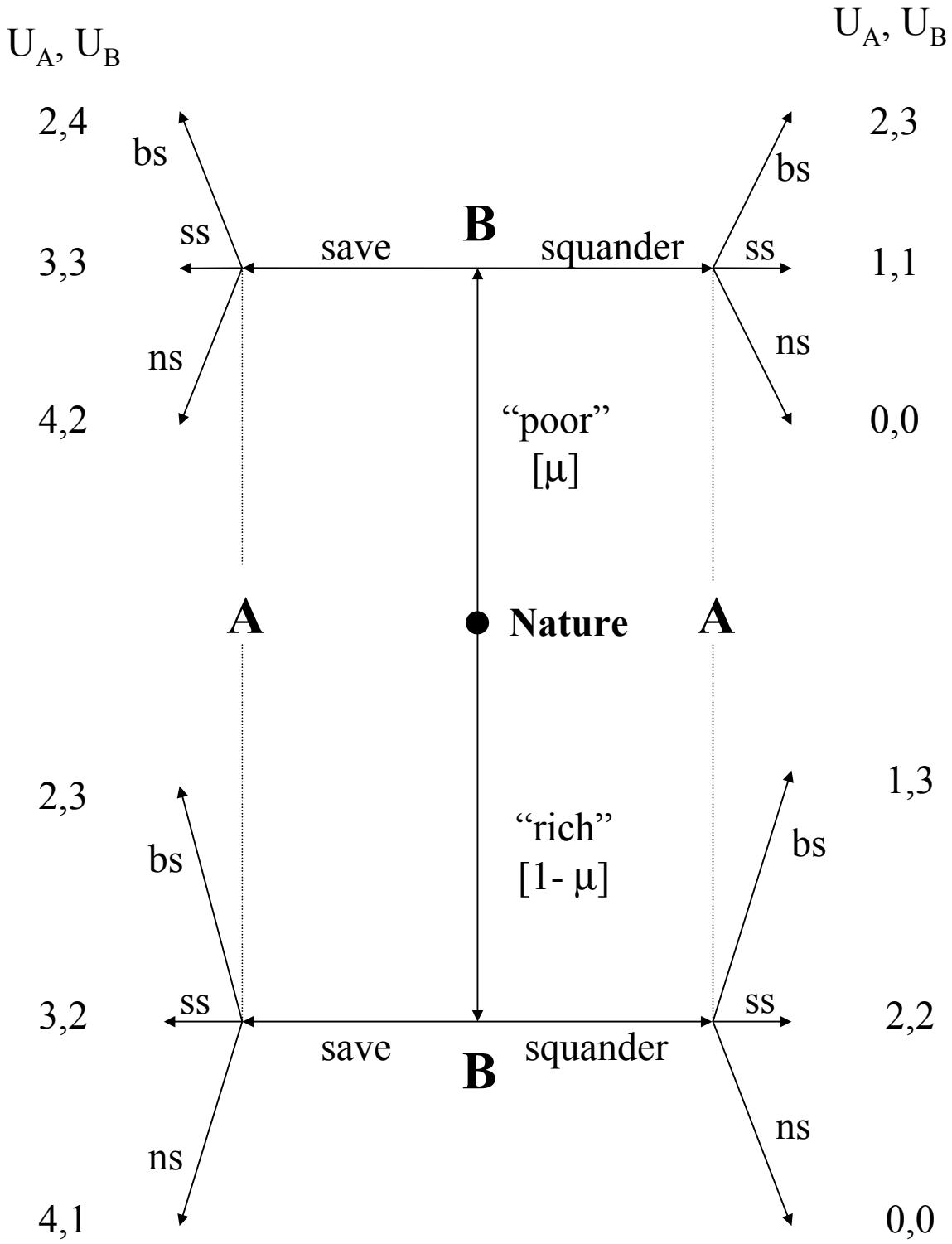

Fig. 1. An example: "Rich Man, Poor Man".

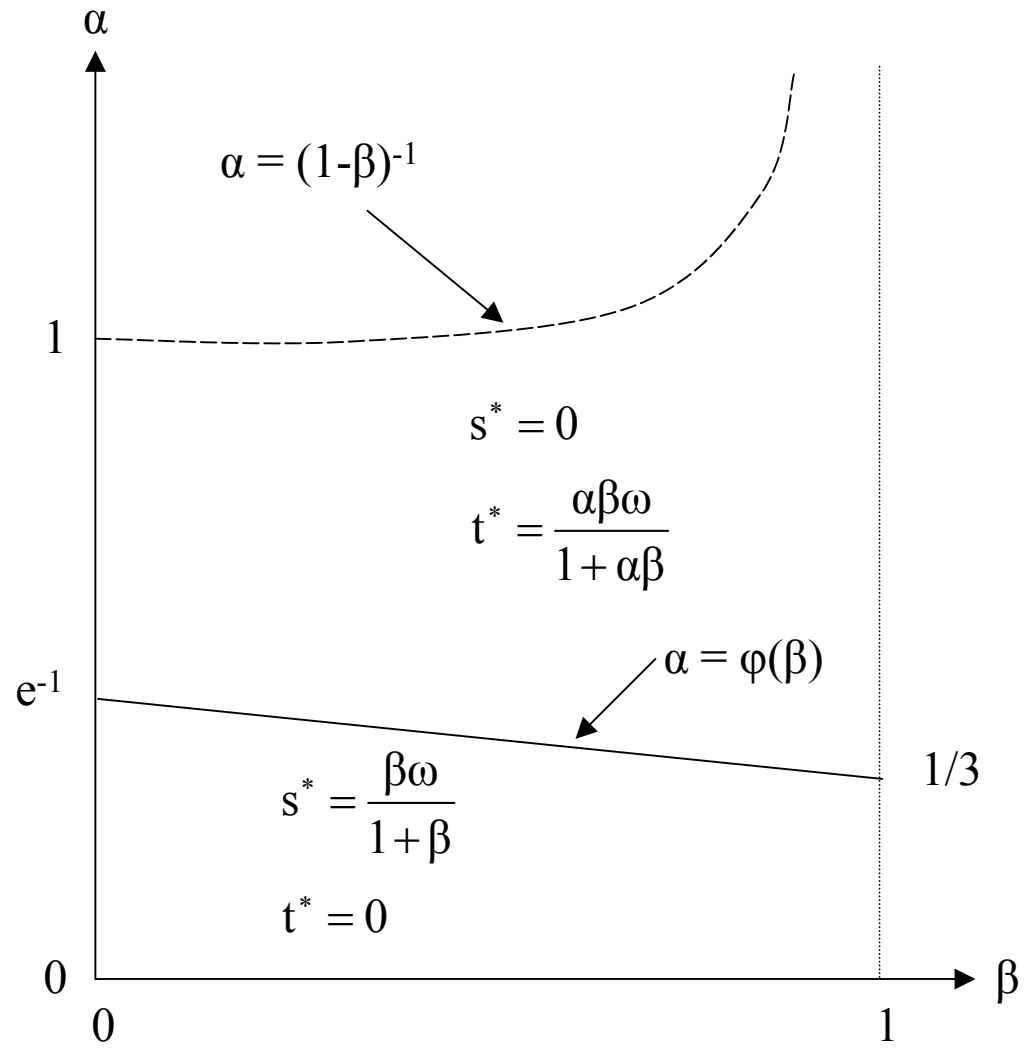

Fig. 2. The Benchmark: Complete information.

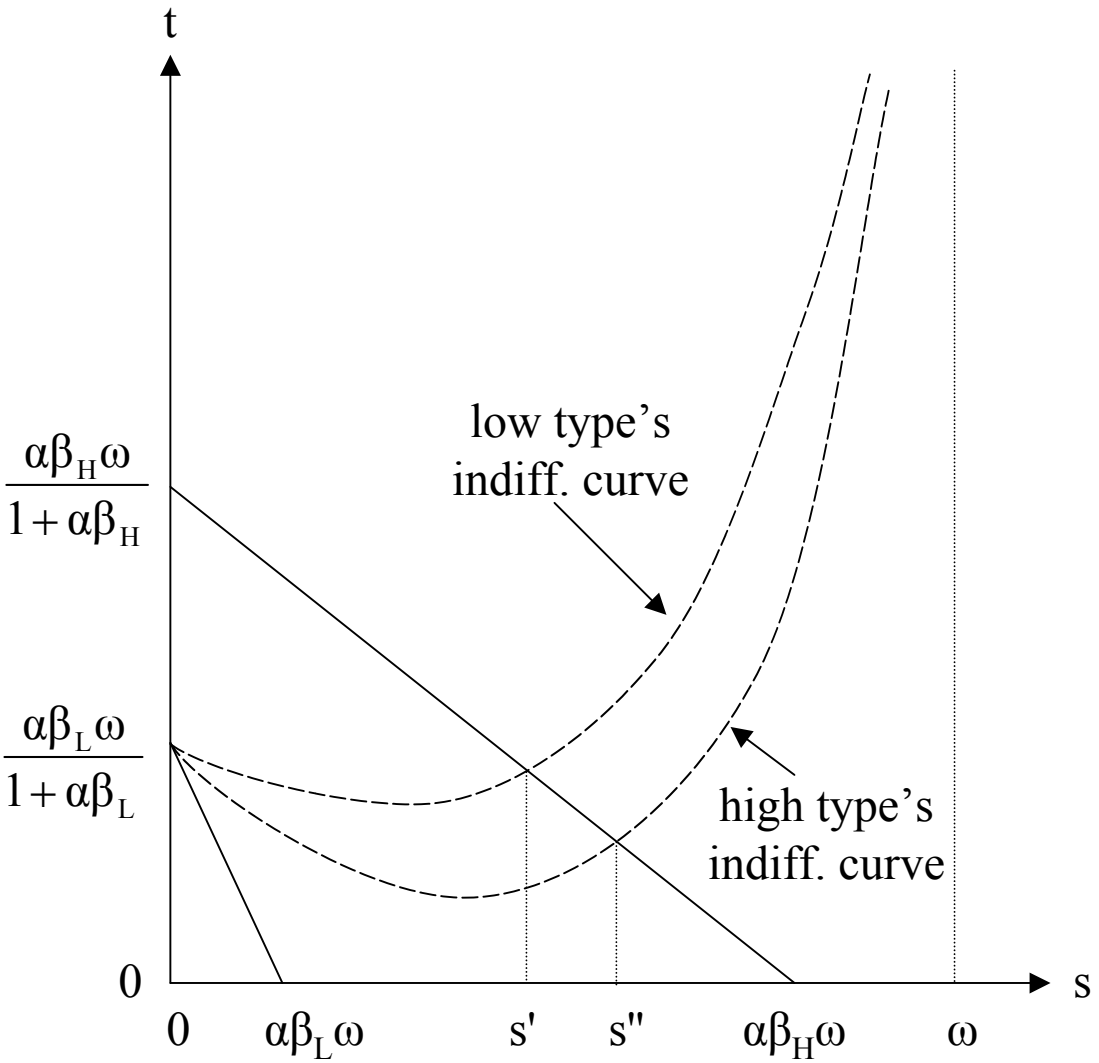

Fig. 3a. Separating equilibria.
The case $s'' \leq \alpha\beta_H\omega$.

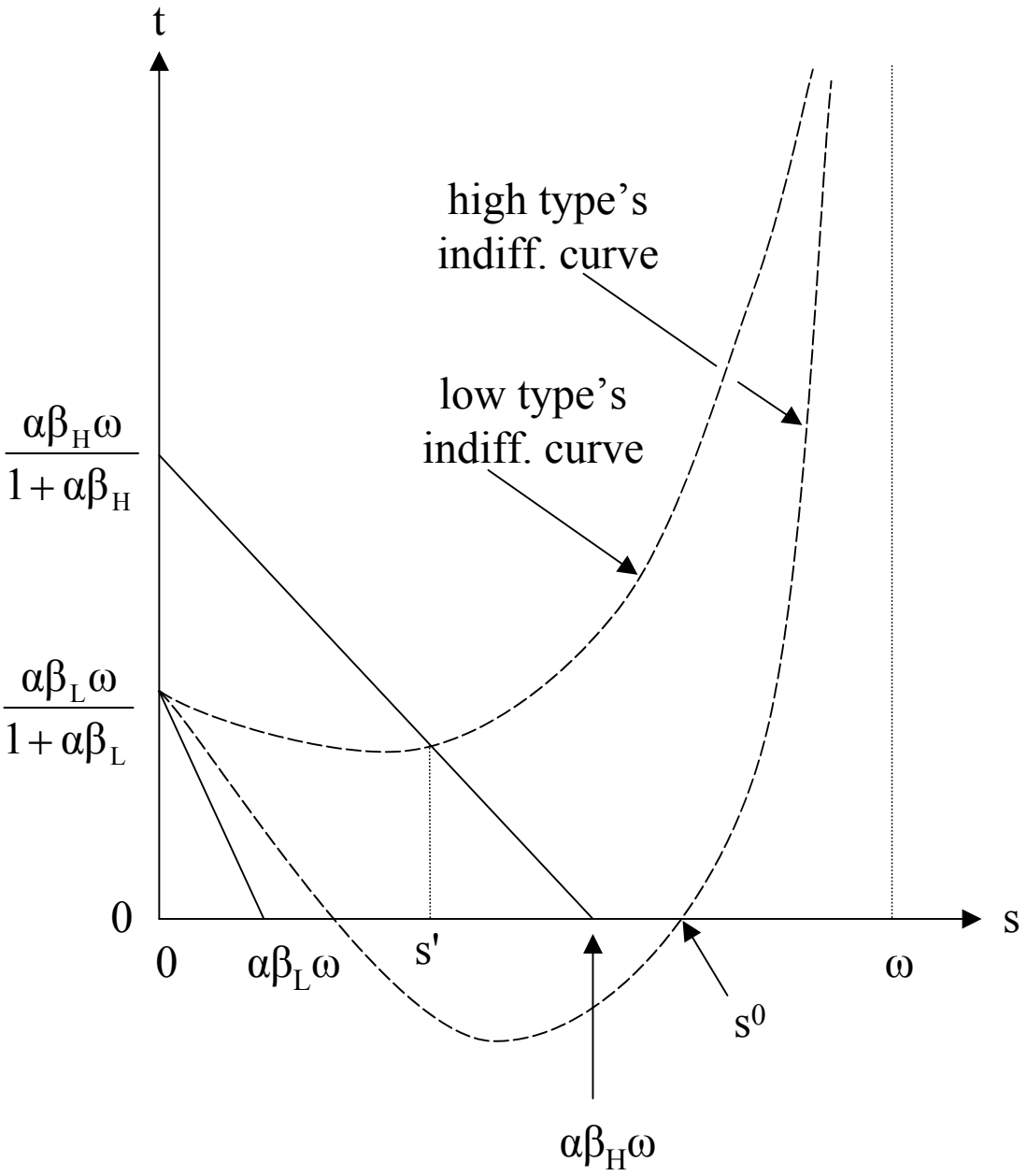

Fig. 3b. Separating equilibria.
The case s''>$\alpha\beta_H\omega$.

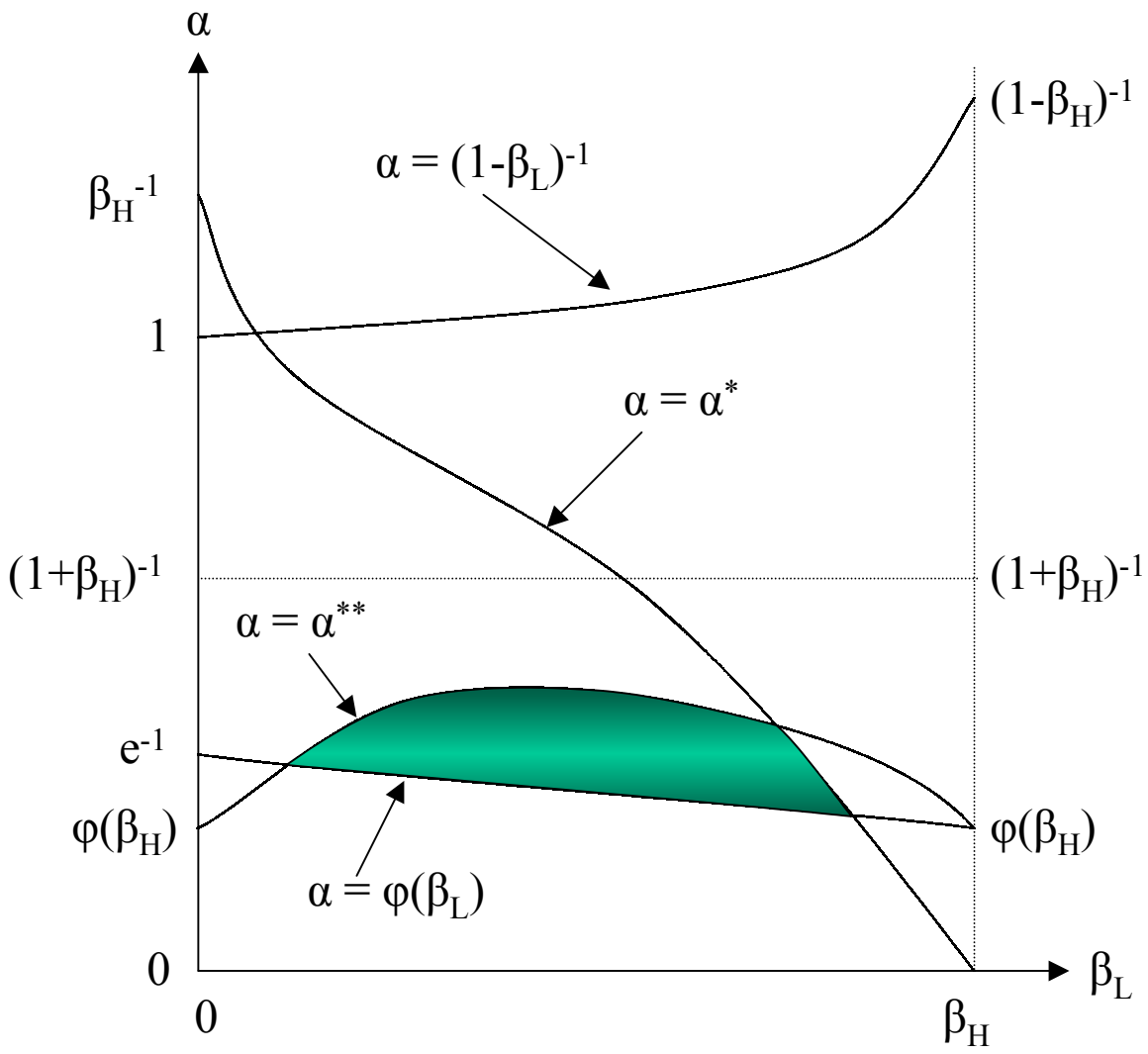

Fig. 4. The unique equilibrium outcome. In the shadowed region the high type saves $s_H^* = \beta_H \omega / (1+\beta_H)$; elsewhere the high type saves $s_H^* = s'$.

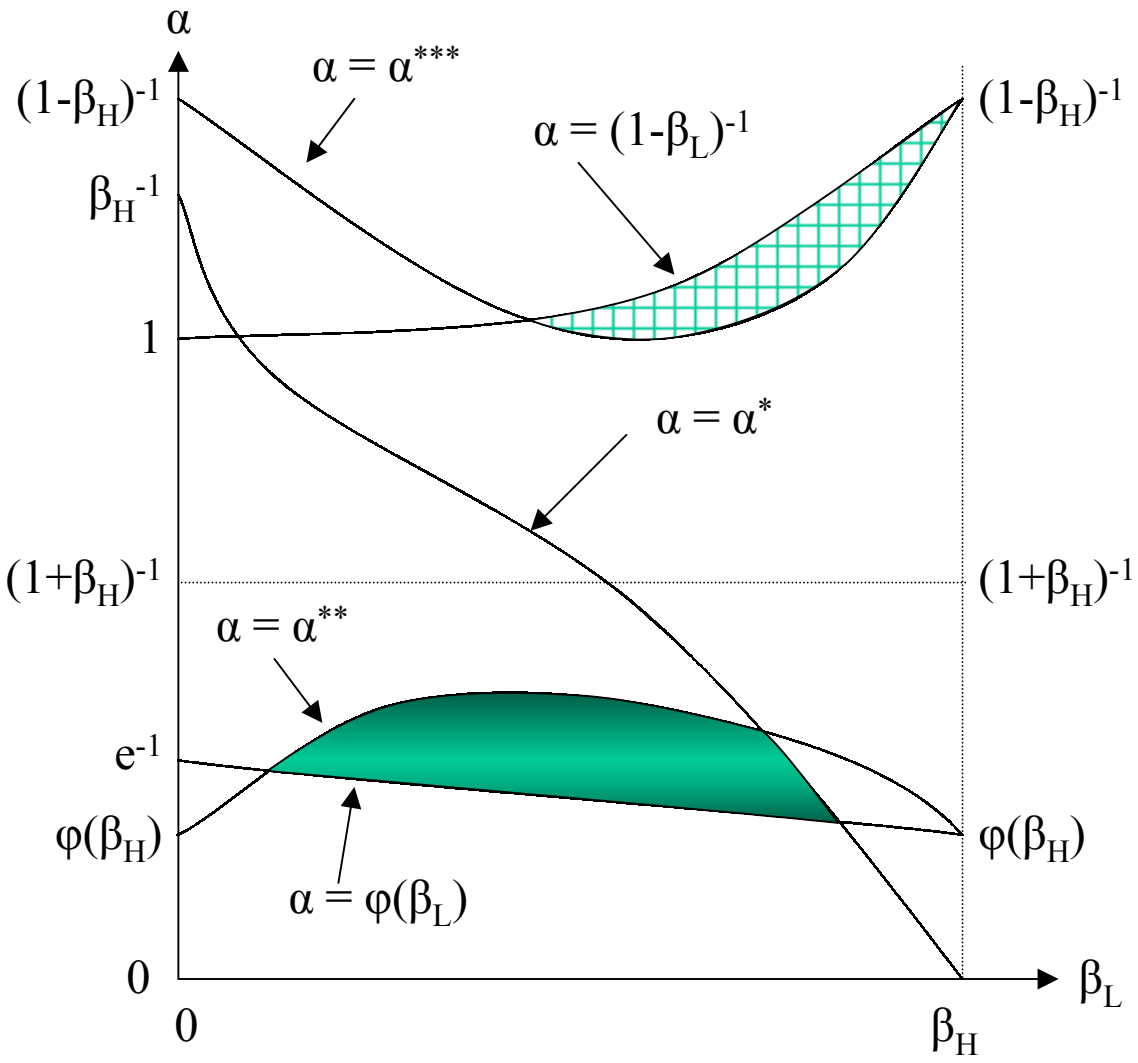

Fig. 5. Efficiency. In the checked region the high type oversaves, whereas in the shadowed region he saves the efficient amount. Elsewhere the high type undersaves, although not as severely as under complete information.