

## EFFICIENCY-ENHANCING SIGNALLING IN THE SAMARITAN'S DILEMMA\*

*Johan Lagerlöf*

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. This paper shows that the undersaving result is mitigated if we relax the standard assumption of complete information, because 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 than he otherwise would have done.

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. Since *A* has altruistic concerns for the welfare of *B*, this would also make *A* better off.<sup>1</sup>

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 governmental transfers.<sup>2</sup> In those models the government provides a transfer on two occasions over time. Since a cash transfer would be used in an inefficient

\* I thank two anonymous referees, the editor D. de Meza, J. Björnerstedt, D. Coen-Pirani, Y. Kang, N.-P. Lagerlöf, B. Persson, D. Soskice, J. Stennek, M. Waldman, 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 Wallander and Hedelius' Research Foundation, *Svenska Institutet*, *Finanspolitiska Forskningsinstitutet* and the European Commission (under the TMR Programme; contract No. ERBFMRXCT980203) is gratefully acknowledged.

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

<sup>2</sup> 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.

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 literature, parental saving is non-strategic, this kind of model is characterised 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 has bearing also 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 maximises the income of the whole family. Hence, the presence of parental altruism induces the child to internalise 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.<sup>3</sup> 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 maximises 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 there is a substantial degree of incomplete information.<sup>4</sup> 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, at least reduced.

<sup>3</sup> 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.

<sup>4</sup> 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.

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  $x$  to be 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 (Spence, 1973). In particular,  $B$  has an incentive to save *more* than in the standard setting with complete information. One should thus expect this effect to counteract the incentives to undersave in the traditional Samaritan's-dilemma model.<sup>5</sup>

In this paper I will explore the logic explained above and show how it can counteract the undersaving result in the traditional formulation of the Samaritan's dilemma. Section 1 of the paper formulates a relatively simple model of the Samaritan's dilemma characterised by incomplete information about  $B$ 's preferences. Section 2.1 provides an analytical benchmark by solving a complete information version of that model. Section 2.2 returns to the incomplete information model and solves for its equilibria. In particular, it is shown that if we impose a commonly used equilibrium refinement (namely, the intuitive criterion), this model has a unique equilibrium outcome. Section 2.3 asks the question 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. That Section also provides conditions for when the high type saves too little, too much, or exactly the right amount. Section 3 concludes with a discussion of the results and their robustness.

## 1. The Samaritan's Dilemma with Incomplete Information

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 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. Also  $A$ 's endowment equals  $\omega$ . In the second period, after having observed  $s$ ,  $A$  chooses how much of her endowment to transfer to  $B$ ,  $t \in [0, \omega]$ .  $A$  consumes the residual amount,  $c_A = \omega - t$ , herself.  $B$ 's second-period consumption consists of his savings plus the transfer from  $A$ :  $c_{2B} = s + t$ .

<sup>5</sup> There is a related literature on altruism and signalling in theoretical biology; see Grafen (1990) and Maynard Smith (1991) for seminal contributions and Goodfray 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.

While  $B$  is assumed to have preferences over only his own consumption in period 1 and 2,  $A$  is altruistic in the sense that she has preferences over both her own consumption and  $B$ 's utility level  $U_B$ .  $B$ 's and  $A$ 's preferences are described, respectively, by the following utility functions:

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

$$\begin{aligned} U_A(s, t | \beta) &= \log(c_A) + \alpha U_B(s, t | \beta) \\ &= \log(\omega - t) + \alpha \log(\omega - s) + \alpha \beta \log(s + t), \end{aligned} \quad (2)$$

where  $\beta$  is a parameter that measures the relative importance of  $B$ 's second-period consumption and  $\alpha$  is a parameter that represents the altruistic concern of  $A$  for the welfare of  $B$ .<sup>6</sup> The magnitude of  $\beta$  is private information to  $B$ . It is common knowledge, however, that  $\beta \in \{\beta_L, \beta_H\}$ , with  $0 < \beta_L < \beta_H < 1$ , and that  $\Pr(\beta = \beta_H) = \mu \in (0, 1)$ .  $B$  will be referred to as the 'low type' if  $\beta = \beta_L$ , and as the 'high type' if  $\beta = \beta_H$ .

In order to simplify the exposition, I will assume throughout that  $A$ 's degree of altruism is relatively strong:  $\alpha\beta_L \geq 1$ . This assumption serves the purpose of limiting the number of cases we have to investigate, since it rules out the possibility that the non-negativity constraint on  $A$ 's transfer is binding. The assumption is not, however, necessary to obtain efficiency-enhancing signalling in this model.

## 2. Efficiency-enhancing Signalling

### 2.1. A Benchmark: Complete Information

Before solving the incomplete information model described in the previous Section, let us first, as an analytical benchmark, investigate the equilibrium behaviour in a complete information version of that model. Let us thus assume that the parameter  $\beta$  can take on only one value (i.e.,  $\beta_L = \beta_H = \beta \in (0, 1)$ ), and that this is common knowledge. All other parts of the model are exactly as described in Section 1.

We can solve this benchmark relatively easily by backward induction. In period 2,  $A$  maximises  $U_A(s, t | \beta)$  with respect to  $t$ , subject to the constraint  $t \in [0, \omega]$ . Denoting the solution to this problem by  $\hat{t}$ , one can readily verify that<sup>7</sup>

$$\hat{t} = \frac{\alpha\beta\omega - s}{1 + \alpha\beta}. \quad (3)$$

Note that  $\hat{t}$  is decreasing in  $s$ : 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

<sup>6</sup> The important sense in which  $\alpha$  represents  $A$ 's altruistic concern for  $B$  is that, for  $\alpha > 0$ ,  $A$  puts a positive weight on  $B$ 's marginal utility, and this weight is increasing in  $\alpha$ .

<sup>7</sup> The assumption that  $\alpha\beta \geq 1$  together with the feasibility constraint on  $s$ ,  $s \leq \omega$ , guarantee that the non-negativity constraint on  $t$  is not binding.

follows we will see that the implicit tax distorts  $B$ 's saving incentives and may make him consume too much in period 1, as compared to what is socially optimal.

In period 1, anticipating  $\hat{t}$ ,  $B$  makes his saving decision.  $B$  thus chooses  $s$  so as to maximise

$$U_B(s, \hat{t} | \beta) = \log(\omega - s) + \beta \log(\omega + s) + \beta \log\left(\frac{\alpha\beta}{1 + \alpha\beta}\right), \quad (4)$$

subject to the constraint  $s \in [0, \omega]$ . Since  $\beta < 1$ , 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 (3) yields the equilibrium outcome of  $t$ ,  $t^* = \alpha\beta\omega/(1 + \alpha\beta)$ .

Is the equilibrium outcome  $(s^*, t^*)$  Pareto efficient? To answer this question, first suppose that  $\alpha \geq (1 - \beta)^{-1}$ . One can verify that then  $(s^*, t^*)$  is the unique solution to the problem of maximising  $U_A(s, t | \beta)$  with respect to  $t$  and  $s$ , subject to the constraint  $(t, s) \in [0, \omega]^2$ . Hence, for  $\alpha \geq (1 - \beta)^{-1}$ , the equilibrium outcome is indeed Pareto efficient. Next, suppose that  $\alpha < (1 - \beta)^{-1}$ . Then, differentiating  $U_B$  with respect to  $s$  and evaluating at  $(s, t) = (s^*, t^*)$ , one has

$$\frac{\partial U_B(s, t | \beta)}{\partial s} \Big|_{(s,t)=(s^*,t^*)} = \frac{1}{\alpha\omega} [1 - \alpha(1 - \beta)] > 0.$$

That is, keeping the level of  $A$ 's transfer fixed,  $B$  would be better off if he saved more. Moreover, since  $\partial U_A/\partial s = \alpha(\partial U_B/\partial s)$ , also  $A$  would be better off if  $B$  saved more given the fixed level of the transfer. Hence, the outcome  $(s^*, t^*)$  is not Pareto efficient for  $\alpha < (1 - \beta)^{-1}$ .

Proposition 1 summarises the results derived above.

**PROPOSITION 1.** *There exists a unique subgame perfect equilibrium of the benchmark model, the outcome of which is  $(s^*, t^*) = (0, \alpha\beta\omega/(1 + \alpha\beta))$ . This equilibrium outcome is Pareto efficient if and only if  $\alpha \geq (1 - \beta)^{-1}$ .*

Under our assumption that  $\alpha\beta \geq 1$ , the condition  $\alpha \geq (1 - \beta)^{-1}$  is satisfied for all  $\beta \leq 1/2$ . For  $\beta > 1/2$ , however, we may have  $\alpha < (1 - \beta)^{-1}$ , in which case  $B$  saves too little. The reason why  $B$  undersaves in the first period 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.

## 2.2. Analysis of the Incomplete Information Model

Let us now return to the main model where  $B$  has private information about the magnitude of  $\beta$ . This is a very standard signalling game with two types. 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.

I will confine attention to pure strategy equilibria of the game, and I start with characterising the separating equilibria, that is, those equilibria in which the two types behave differently, thereby making it possible for  $A$  to infer  $B$ 's type perfectly. Let  $s_L^*$  and  $s_H^*$  denote the low, respectively the high, type's saving level in such an equilibrium. And let  $t_L^*$  and  $t_H^*$  denote the equilibrium transfers to the low, respectively the high, type. The following lemma will be helpful in the analysis.

LEMMA 1. *In any separating equilibrium,  $s_L^* = 0$ .*

*Proof.* Let  $\tilde{\mu}(s)$  denote the probability  $A$  assigns to the event that  $B$  is the high type conditional on having observed a saving level  $s$ , and let  $E(\beta | s) \equiv \tilde{\mu}(s)\beta_H + [1 - \tilde{\mu}(s)]\beta_L$ . Obviously,  $\tilde{\mu}(s) \in [0, 1]$ , so  $E(\beta | s) \in [\beta_L, \beta_H]$ . One can easily verify that  $A$ 's optimal transfer, conditional on having observed a saving level  $s$ , is given by  $\hat{t}$  in (3) but with  $E(\beta | s)$  substituted for  $\beta$ . Thus, the low type's indirect utility in a separating equilibrium, given that he saves  $s$ , equals

$$\log(\omega - s) + \beta_L \log(\omega + s) + \beta_L \log \left[ \frac{\alpha E(\beta | s)}{1 + \alpha E(\beta | s)} \right];$$

see (4). For a fixed  $E(\beta | s)$ , this expression is decreasing in  $s$ . Furthermore, it is increasing in  $E(\beta | s)$ : if possible, the low type would like to be perceived as the high type in the eyes of  $A$ . In any separating equilibrium, however,  $B$ 's type will, by definition, be revealed, so  $E(\beta | s_L^*) = \beta_L$ . Moreover, if the low type chose some  $s \neq s_L^*$ ,  $E(\beta | s)$  would possibly differ from  $\beta_L$ , but it could not become lower than it already is. But these things taken together mean that if we had  $s_L^* > 0$  in a separating equilibrium, then the low type could gain by deviating to some  $s \in [0, s_L^*)$ .

We are thus looking for an equilibrium with  $s_L^* = 0$  and where  $s_H^*$  is positive. The analysis is facilitated by Figure 1, which shows the saving-transfer space. The two straight lines in the Figure represent  $A$ 's optimal transfer as given by  $\hat{t}$  in (3); the lower transfer line corresponds to  $A$ 's believing that  $\beta = \beta_L$ , and the upper one corresponds to  $A$ 's believing that  $\beta = \beta_H$ . 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 curves 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. An important characteristic of the indifference curves is that they satisfy the so-called single-crossing property: for any given  $s$ , the low type's indifference curve has a larger slope than the high type's;<sup>8</sup> this is what will make it possible to have existence of separating equilibria. In order to understand the logic of a separating equilibrium, it is also important to recall that each type wants to be perceived as the high type in the eyes of  $A$ , since then the transfer is larger. Furthermore, if one moves northwest along any one of the transfer lines, both types are made better off; see (4).

<sup>8</sup> One can also 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 1, be strictly concave functions of  $s$ , which tend to infinity as  $s$  tends to  $\omega$ . 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  $\omega$ , the slopes approach infinity.

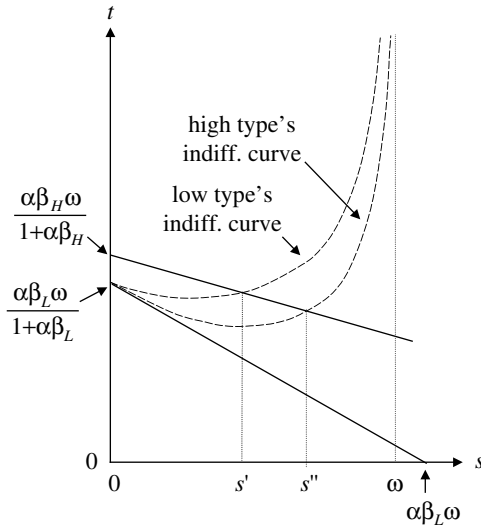


Fig. 1. *Separating Equilibria*. All saving levels between  $s'$  and  $s''$  can be sustained as part of a separating equilibrium but only an equilibrium with  $s = s'$  survives the dominance argument

As indicated in the Figure, the value of  $s$  for which the low type's indifference curve intersects the upper transfer line is denoted  $s'$ , and the value of  $s$  for which the high type's indifference curve intersects the same line is denoted  $s''$  (given our assumption that  $\alpha\beta_L \geq 1$ , one obviously must have  $s'' < \alpha\beta_H\omega$ , as drawn in the Figure). The first point of intersection, which will turn out to be the most important one in the subsequent analysis, is implicitly defined by the following identity:

$$\log\left(1 - \frac{s'}{\omega}\right) + \beta_L \log\left(1 + \frac{s'}{\omega}\right) \equiv \beta_L \log\left[\frac{\beta_L(1 + \alpha\beta_H)}{\beta_H(1 + \alpha\beta_L)}\right]. \tag{5}$$

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 by 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  $s_L^* = 0$  and any  $s_H^* \in [s', s'']$  to be part of a separating equilibrium. For example, we can let

$$\tilde{\mu}(s) = \begin{cases} 0 & \text{for } s \in [0, s_H^*) \\ 1 & \text{for } s \in [s_H^*, \omega] \end{cases} \tag{6}$$

(as  $\tilde{\mu}(s)$  denotes  $A$ 's posterior beliefs that  $B$  is the high type conditional on having observed a saving level  $s$ ). These beliefs are consistent with the equilibrium requirements and they guarantee that the types do not have an incentive to deviate from  $s_L^*$  respectively  $s_H^*$ . As already mentioned, the reason why a separating equilibrium can exist is that the two types' indifference curves satisfy the single-crossing property: the marginal utility of choosing a particular saving level differs between the two types, so a high type can separate himself from the low type by choosing a saving level that would be too costly for the low type to emulate.

The important feature of the beliefs in (6) 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  $s \in [s', s_H^*]$ . 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  $\tilde{\mu}(s) \in [0, 1]$   $A$  holds: choosing  $s = 0$  always gives him a higher utility (see Figure 1). One may plausibly argue that  $A$  should realise 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  $\tilde{\mu}(s) = 1$  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'. The only equilibrium saving level of the high type that survives the dominance argument is  $s_H^* = s'$ . Hence, the dominance argument gives us a unique separating equilibrium outcome, in which  $s_L^* = 0$  and  $s_H^* = s'$ .

We have not yet considered the existence of pooling equilibria. For the sake of brevity, this will not be done here. In Lagerlöf (2003), however, I show that the intuitive criterion (Cho and Kreps, 1987) rules out all pooling equilibria. This equilibrium refinement implies the dominance argument used above, although it is stronger, and it is very often applied in signalling games. Since the intuitive criterion rules out all pooling equilibria, it gives us a unique equilibrium outcome.

**PROPOSITION 2.** *There is a unique equilibrium outcome that satisfies the intuitive criterion, which is given by*

$$(s_L^*, t_L^*, s_H^*, t_H^*) = \left( 0, \frac{\alpha\beta_L\omega}{1 + \alpha\beta_L}, s', \frac{\alpha\beta_H\omega - s'}{1 + \alpha\beta_H} \right).$$

### 2.3. 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 concept of *ex post* incentive efficiency, suggested by Holmström and Myerson (1983). Following them, I say that an outcome  $(s_L, t_L, s_H, t_H)$  is *incentive feasible* if the low type prefers  $(s_L, t_L)$  to  $(s_H, t_H)$ , the high type prefers

$(s_H, t_H)$  to  $(s_L, t_L)$ , and  $(s_L, t_L, s_H, t_H) \in [0, \omega]^4$ . And an outcome  $(s'_L, t'_L, s'_H, t'_H)$  *ex post dominates* an outcome  $(s_L, t_L, s_H, t_H)$  if both players, for each realisation of  $B$ 's type  $j$ , prefer  $(s'_j, t'_j)$  to  $(s_j, t_j)$  (where the preference must be strict in at least one of these four comparisons). An outcome  $(s_L, t_L, s_H, t_H)$  is then said to be *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'. That is, the equilibrium outcome could possibly satisfy a weaker version of *ex post incentive efficiency* in which the above definition of 'ex post dominates' requires the two players to prefer  $(s'_j, t'_j)$  to  $(s_j, t_j)$  for  $j = H$  but not necessarily for  $j = L$ . In the following I will investigate if and when the high type of  $B$  behaves efficiently in this sense in the unique equilibrium outcome.

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:

$$s^*_H \in \arg \max_{s \in [0, \omega]} U_B(s, t^*_H \mid \beta_H), \tag{7}$$

$$t^*_H \in \arg \max_{t \in [0, \omega]} U_A(s^*_H, t \mid \beta_H). \tag{8}$$

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, for any  $\alpha \leq 1/(1-\beta_H)$ , conditions (7) and (8) are met only for  $(s^*_H, t^*_H) = (s^e, t^e)$ , where

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

Although (7) is only a necessary – not a sufficient – condition for *ex post incentive efficiency*, one can verify that (again assuming that  $\alpha \leq 1/(1 - \beta_H)$ )  $(s^*_L, t^*_L, s^e, t^e)$  is, whenever we can sustain it as an outcome of separating equilibrium, indeed *ex post incentive efficient* for the high type.

Thus, the high type saves the efficient amount if  $s' = s^e$ ; if  $s' < s^e$ , the high type saves too little; and if  $s' > s^e$ , he saves too much. So how does  $s'$  relate to  $s^e$ ? Lemma 2 below (which is proven in the Appendix) tells us that this depends on how  $\alpha$  relates to a cut-off value  $\alpha^*(\beta_L, \beta_H)$ . The lemma also notes some useful properties of this cut-off value.

LEMMA 2. *Suppose that  $\beta_L > 1/2$  and  $\alpha \in [1/\beta_L, 1/(1 - \beta_L)]$ , so that both types undersave in the benchmark. Then there exists a function  $\alpha^*(\beta_L, \beta_H)$  such that  $s' \begin{matrix} \leq \\ > \end{matrix} s^e$  as  $\alpha \begin{matrix} \leq \\ > \end{matrix} \alpha^*(\beta_L, \beta_H)$ , with  $\alpha^*(\beta_L, \beta_H)$  implicitly defined by*

$$\log \left[ \frac{1 + (1 + \beta_H)\alpha^*}{2\alpha^*} \right] + \beta_L \log \left\{ \frac{\beta_L [1 + (1 + \beta_H)\alpha^*]}{2\beta_H (1 + \beta_L \alpha^*)} \right\} \equiv 0. \tag{9}$$

Moreover, this function satisfies  $\lim_{\beta_L \rightarrow \beta_H} \alpha^*(\beta_L, \beta_H) = (1 - \beta_H)^{-1}$ ; and, for  $\beta_L$  sufficiently close to  $\beta_H$ ,  $\alpha^*(\beta_L, \beta_H) < (1 - \beta_L)^{-1}$ .

In Figure 2 the results are illustrated in a diagram that depicts  $\alpha$  on the vertical and  $\beta_L$  on the horizontal axis. The Figure assumes that  $\beta_L > 1/2$ , since for lower values of this parameter the low type would not undersave under complete information, given our assumption that  $\alpha \geq 1/\beta_L$  (see the remarks after Proposition 1). Also recall from Proposition 1 that the equilibrium outcome under complete information involves undersaving only if  $\alpha < 1/(1 - \beta_L)$ . The Figure depicts the graphs of  $\alpha = 1/\beta_L$  and  $\alpha = 1/(1 - \beta_L)$  together with that of  $\alpha^*(\cdot, \beta_H)$ .

Lemma 2 tells us that, at least if  $\beta_L$  is sufficiently close to  $\beta_H$ , the graph of  $\alpha^*(\cdot, \beta_H)$  goes through the region of the parameter space where  $\alpha < 1/(1 - \beta_L)$  holds. Hence, there is a non-empty region of the parameter space (namely, the shadowed area in the Figure) where under complete information the high type undersaves but under incomplete information he saves too much! Lemma 2 also tells us that on the lower border of this region, i.e., where  $\alpha = \alpha^*(\beta_L, \beta_H)$ , the unique equilibrium outcome is indeed *ex post* incentive efficient for the high type, although this is of course a knife-edge phenomenon. In the remaining part of the parameter space (the checked area in the Figure), where  $\alpha < \min[\alpha^*(\beta_L, \beta_H), 1/(1 - \beta_L)]$  and  $\alpha \geq 1/\beta_L$ , the high type undersaves. Yet, even here, the high type's saving choice is distorted upwards: he saves more than he would have done under complete information.

Summing up, we have the following proposition.

**PROPOSITION 3.** *Suppose that  $\beta_L > 1/2$  and  $\alpha \in [1/\beta_L, 1/(1 - \beta_L)]$ , so that both types undersave in the benchmark. Then the unique equilibrium outcome of the incomplete information model is such that the low type's saving choice is unaffected relative to the benchmark*

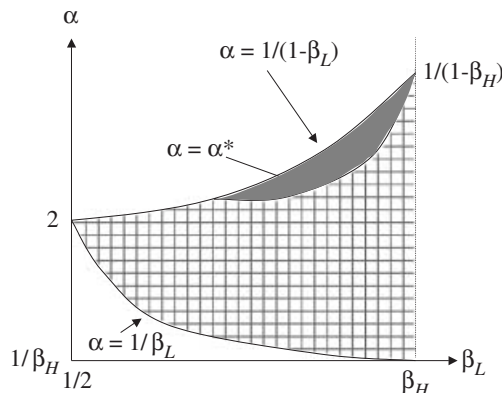


Fig. 2. *Efficiency.* In the shadowed region the high type oversaves and in the checked region he undersaves, although not as severely as under complete information

whereas the high type saves more than in the benchmark. The high type's saving level is 'ex post incentive efficient for the high type' if  $\alpha = \alpha^*(\beta_L, \beta_H)$ ; it involves undersaving if  $\alpha < \alpha^*(\beta_L, \beta_H)$ ; and it involves oversaving if  $\alpha > \alpha^*(\beta_L, \beta_H)$ .

### 3. 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. Also the logic of this new force should, 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) must 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 described in Section 1,  $x$  is a weight on  $B$ 's second-period utility. One may wonder whether the presence of the counteracting effect hinges 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.  $A$  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 the size of  $B$ 's income in the first period then 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 conceivably could be uncertainty about is the altruism parameter,  $\alpha$ .<sup>9</sup> In the model analysed in the present paper,  $A$ 's having private information about  $\alpha$  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 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 altruistic 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.<sup>10</sup>

*WZB and CEPR*

*Date of receipt of first submission: June, 2000*

*Date of receipt of final typescript: June, 2003*

## Appendix: Proof of Lemma 2

By using the definition of  $s'$  in (5), one can show that  $s^e \geq s'$  is equivalent to  $Z(\beta_L, \beta_H, \alpha) \geq 0$ , where  $Z(\beta_L, \beta_H, \alpha)$  is the left-hand side of (9) (but with  $\alpha$  substituted for  $\alpha^*$ ). Clearly, this inequality is satisfied if  $\alpha$  is sufficiently close to zero. Moreover, the upper constraint on  $\alpha$  in Lemma 2,  $(1 - \beta_L)^{-1}$ , 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)  $Z(\beta_L, \beta_H, \alpha)$  is strictly decreasing in  $\alpha$  for all  $\alpha \in (0, [\beta_L(\beta_H - \beta_L)]^{-1})$  and (ii)  $Z(\beta_L, \beta_H, \alpha) \geq 0$  does not hold for  $\alpha = [\beta_L(\beta_H - \beta_L)]^{-1}$ . To establish (i), differentiate  $Z(\beta_L, \beta_H, \alpha)$  with respect to  $\alpha$ ; 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), note that  $Z(\beta_L, \beta_H, [\beta_L(\beta_H - \beta_L)]^{-1}) = g(\beta_L, \beta_H)$ , where

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

Clearly,  $g(\beta_L, 1) < 0$  for all  $\beta_L \in (0, 1)$ . Moreover, one can easily check that  $g(\beta_L, \beta_H)$  is increasing in  $\beta_H$ . It follows that the threshold  $\alpha^*$  is well defined with  $s' \leq s^e$  as

<sup>9</sup> Uncertainty about the degree of altruism has been modelled by, for example, Chakrabarti *et al.* (1993).

<sup>10</sup> If we assumed that the players are interacting over more time periods, so that  $A$  had more than one opportunity to transfer resources to  $B$ , then a similar result should hold also in a model in which altruism is one-sided. For in such an environment  $A$  would early on have an incentive to transfer and save relatively little in order to signal low altruism, in an attempt to limit  $B$ 's tendency to overconsume. I thank Mike Waldman for suggesting this to me.

$\alpha \geq \alpha^*(\beta_L, \beta_H)$ . Moreover, by substituting  $\beta_L = \beta_H$  and  $\alpha^* = (1 - \beta_H)^{-1}$  into (9), one can verify that  $\lim_{\beta_L \rightarrow \beta_H} \alpha^*(\beta_L, \beta_H) = (1 - \beta_H)^{-1}$ . Let us finally show that for  $\beta_L$  sufficiently close to  $\beta_H$ ,  $\alpha^*(\beta_L, \beta_H) < (1 - \beta_L)^{-1}$ . Since  $\lim_{\beta_L \rightarrow \beta_H} \alpha^*(\beta_L, \beta_H) = (1 - \beta_H)^{-1}$ , it suffices to show that

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

Straightforward calculations yield

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

Hence, using  $\lim_{\beta_L \rightarrow \beta_H} \alpha^*(\beta_L, \beta_H) = (1 - \beta_H)^{-1}$ , one has  $\lim_{\beta_L \rightarrow \beta_H} \frac{\partial \alpha^*(\beta_L, \beta_H)}{\partial \beta_L} = 2(1 - \beta_H)^{-2}$ , which is always greater than  $(1 - \beta_H)^{-2}$ .

## References

- Becker, Gary S. (1974). 'A theory of social interactions', *Journal of Political Economy*, vol. 82(6), pp. 1063–93.
- 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(5), pp. 1138–59.
- Bernheim, B. Douglas and Stark, Oded (1988). 'Altruism within the family reconsidered: do nice guys finish last?', *American Economic Review*, vol. 78(5), pp. 1034–45.
- Bruce, Neil and Waldman, Michael (1990). 'The Rotten-Kid Theorem meets the Samaritan's dilemma', *Quarterly Journal of Economics*, vol. 105 (February), pp. 155–65.
- Bruce, Neil and Waldman, Michael (1991). 'Transfers in kind: why they can be efficient and nonpaternalistic', *American Economic Review*, vol. 81(5), pp. 1345–51.
- Buchanan, James M. (1975). The Samaritan's dilemma, in (E. S. Phelps, ed.), *Altruism, Morality and Economic Theory*, pp. 71–85, New York: Russel Sage Foundation.
- Chakrabarti, Subir, Lord, William and Rangazas, Peter (1993). 'Uncertain altruism and investment in children', *American Economic Review*, vol. 83(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(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(1403), pp. 1581–91.
- Grafen, Alan (1990). 'Biological signals as handicaps', *Journal of Theoretical Biology*, vol. 144(4), pp. 517–46.
- Hansson, Ingemar and Stuart, Charles (1989). 'Social security as trade among living generations', *American Economic Review*, vol. 79(5), pp. 1182–95.
- Holmström, Bengt and Myerson, Roger B (1983). 'Efficient and durable decision rules with incomplete information', *Econometrica*, vol. 51(6), pp. 1799–819.
- Kotlikoff, Laurence J (1987). 'Justifying public provision of social security', *Journal of Policy Analysis and Management*, vol. 6(4), pp. 674–89.
- Lagerlöf, Johan (2003). 'Supplementary material to "Efficiency-enhancing signalling in the Samaritan's dilemma"', mimeo, WZB Berlin, available at [www.johanlagerlof.org](http://www.johanlagerlof.org).
- Lindbeck, Assar and Weibull, Jörgen W (1988). 'Altruism and time inconsistency: the economics of fait accompli', *Journal of Political Economy*, vol. 96(6), pp. 1165–82.
- Maynard Smith, John (1991). 'Honest signalling: the Philip Sidney game', *Animal Behaviour*, vol. 42(6), pp. 1034–5.
- O'Connell, Stephen A. and Zeldes, Stephen P (1993). 'Dynamic efficiency in the gifts economy', *Journal of Monetary Economics*, vol. 31(3), pp. 363–79.

- Spence, Michael (1973). 'Job market signaling', *Quarterly Journal of Economics*, vol. 87 (August), pp. 355-74.
- Thompson, Earl A. (1980). 'Charity and nonprofit organizations', in (K. Clarkson and D. Martin, eds.), *Economics of Nonproprietary Organizations*, pp. 125-38, Greenwich, CT: JAI Press, Inc.
- Veall, Michael R. (1986). 'Public pensions as optimal social contracts', *Journal of Public Economics*, vol. 31(2), pp. 237-51.