Examples for and questions about: An Economic (?) Theory of Promises[1]

Bengt Holmstrom and David Kreps

September 1995

0. Introduction

The subject is the use of phrases such as "I promise" given in the course of economic exchange. The *such as* in the preceding sentence is ambiguous, and a large part of this project is to try to figure out just what (in economic terms) is the essence of a promise, and what are economic alternatives to promises.

To give the story away at the start, it is relatively easy — embarrassingly so, perhaps — to cook a story in which (cheap-talk) promises are given. The mathematics of these stories turns out to be messy, but if you don't mind some less-than-complete characterizations, we can put together a story.

The story has problems, however. Chief among them is that there would seem to be alternatives to "promises" that are at least as efficient as promises, at least in simple settings. We conclude (at this stage) that the simple models we are about to sketch are inadequate to explain the prevalence of promises, although they might be helpful in understanding a bit about the structure of equilibria based on promises, once we accept that promises are used. At the end, we will try to indicate what more adequate explanations might look like; as you will see, they aren't very "economic," hence the question mark in the title.

I. Theme: A minimal model of promises

As far as computations and formal models go, we will embed our examples in the following simple situation:

(I.1) There are two parties, A and B. B can enter into an "exchange" with

[1] Preliminary. We are grateful to John Macmillan and Ennio Stachetti for helpful conversations and ideas.

A, where, at some future date, A is called upon to perform some service for B. If A performs that service well, B will be better off for having entered into the exchange and having obtained A's services. But if A performs the service badly, B will be worse off than if he had not entered into the arrangement. A is better off for making this exchange whether she performs it well or not, than if B does not enter into the exchange. But she (A) is best off if she doesn't perform the service well.

In a picture, we have the standard "Trust" game depicted in figure 1. B must decide whether to trust A or not (enter into the exchange or not), and then if B enters into the exchange, A must decide whether to honor or abuse B's trust. Note that the payoffs depicted in figure 1 correspond to the story in the previous paragraph *if* $1 > x > 0$ and $y < 0$.

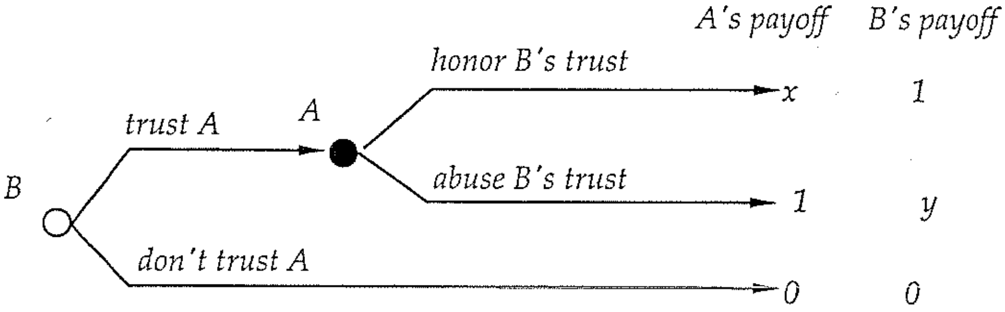


FIG. 1. The trust game.

(I.2) *The standard analysis.* This simple situation has been studied exhaustively by theorists, with the following a summary of standard equilibrium analysis.

(I.2a) As long as $1 > x$ and $y < 0$, and if this game is played once, then A will abuse B's trust if B trusts A (in equilibrium). Hence B will not trust A.

(I.2b) But if A and B engage in this ritualized exchange repeatedly, say at dates $t = 0, 1, 2, \ldots$, and if A discounts her payoffs at some rate $\alpha > 0$ that is close to one, then we can sustain *trust–honor* in a supergame equilibrium: B will begin by trusting A and will continue to do so as long as A honors B's trust; but if A ever abuses B, then B will forego trusting A for N periods. And A honors trust if given and if she has not abused B anytime in the last N periods; otherwise she will abuse B. This isn't quite subgame perfect (although it is easy to modify things so that it is), but it gives a Nash equilibrium: The crucial calculation is: Along the path of play prescribed by these strategies, A gets x per period for a discounted value of $x/(1-\alpha)$. If A ever abuses B's trust, A gets 1 immediately, 0 for the next N periods, and then (assuming A returns to the equilibrium) x thereafter, for a payoff of

$$1 + \alpha^N[x/(1-\alpha)].$$

As long as

$$x \geq \frac{1-\alpha}{1-\alpha^N},$$

which will happen for large enough N if $x \geq 1-\alpha$,[2] A would (weakly) prefer not to abuse B; by a standard dynamic programming argument, one can show that A cannot find a strategy that improves on the one posited, and hence A is playing a best response to B. Along the path A always honors B's trust, and so B's best response is to trust A, hence B is playing a best response, and we have a Nash equilibrium.

(I.2c) It is unimportant that the same B plays against A repeatedly, but only that A plays this game at a sequence of dates and discounts the future as a low enough rate (with large enough α), as long as Bs who face A in the future can observe far enough (N periods) back how A has behaved.

(I.3) This is all (nowadays) the stuff of first year theory courses (perhaps even less). What we wish to do is to find some way to inject into the story the notion

[2] We allow $N = \infty$.

that A will greet B with the statement, "I promise not to abuse you if you trust me," and have this be something more than window dressing economically.

(I.4) One easy way to do this is to attach some direct cost to A of breaking a promise once given. We think this raises some interesting issues, but for the time being we instead presume that A's promise, if she gives it, is cheap-talk. That is, it costs her nothing to say it, and it will cost her nothing *directly* whether she keeps or breaks her promise.

(I.5) If this sort of promise is going to be more than window dressing, then it is clear that the model will have to be fancied up. We are led by the following consideration:

If the promise has any economic role to play, it must be that B doesn't always expect to hear it. Moreover, it must be that B can't anticipate whether he will hear it or not with information he possesses at the moment it is stated.[3] This means that A must possess (at the time of utterance of the speech) some information that B lacks.

(I.6) *The minimal model.* Which leads to the following minimal model. We suppose that A encounters B in the game depicted in figure 1 at dates $t = 0, 1, \ldots$. The value of y is fixed, but at date t the value of x is x_t, drawn independently from some fixed distribution F. We will assume that x_t has finite expectation $\mathrm{E}x > 0$ and support that is bounded above by 1 (with no mass at 1, if in fact 1 is in the support), so that abuse is always better in the short run for A than is honor. Let $\underline{x}$ denote the lower bound of the support of F. We will generally

[3] At this point, we'd better cover ourselves and say: This is meant to be an exercise in general persuasion and not a tight logical argument. It can't be a tight logical argument because the operative term *is more than window dressing* doesn't have a tight definition. Notwithstanding that, we don't mean to claim that there isn't some definition of *more than window dressing* for which our arguments will not apply. But we hope to persuade you that a reasonable interpretation of the phrase leads to the conclusions we reach.

assume B can never observe x_t (although at times we will discuss how things change if he can see x_t ex post).

(I.7) *Always trust, always honor (ATAH).* Suppose the game is strictly that in figure 1, where B must decide (at date t) whether to trust A or not with no communication from A. We can sustain the "always trust, always honor" (hereafter, ATAH) outcome as an equilibrium as long as

$$\underline{x} \geq 1 - \alpha\frac{Ex}{1-\alpha}.$$

The logic here is: The most severe punishment that B can mete out is relevant (since we want to sustain the always-trust equilibrium), and this is $N = \infty$. Then the worst case for sustaining the equilibrium is where A has a current high cost of honor (x_t is very low). The expected value to A of honor this period and continuing with the equilibrium is $x_t + \alpha \mathrm{E}x/(1-\alpha)$, so to get honor as a (weak) best response in all circumstances, we need

$$\underline{x} + \alpha\frac{\mathrm{E}x}{1-\alpha} \geq 1,$$

(since A can get 1 by abusing B in the current round) which gives the inequality above. For the sake of later discussion, we let

$$\underline{x}^* = 1 - \alpha\frac{Ex}{1-\alpha}.$$

That is, the condition for ATAH to be sustainable as a Nash equilibrium is that $\underline{x} \geq \underline{x}^*$.

To keep an example in mind, suppose that $\mathrm{E}x = .5$. Then, as a function of α, we get $\underline{x}^*$ as graphed in figure 2. Note that we start with $\alpha = 1/2$, as otherwise we cannot support ATAH, even if x_t has a degenerate distribution. In particular, for $\alpha = .9$, the $\underline{x}^* = -3.5$.

To fix matters very firmly, we will use until futher notice the parameterization where $\alpha = .9$, and $x_t = -3$ with probability .1 and $8/9$ with probability .9.

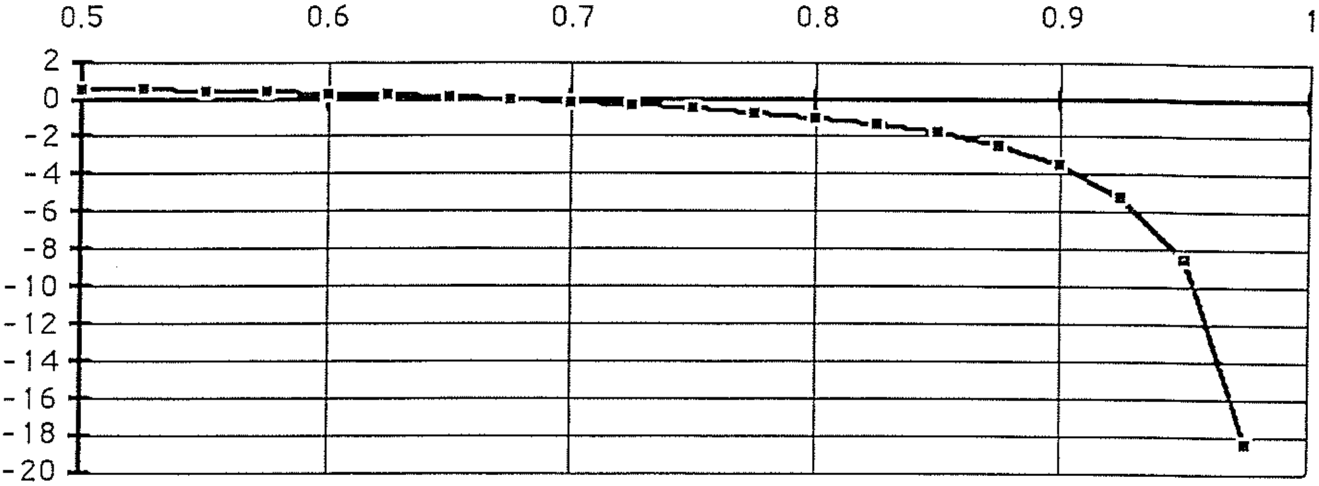


FIG. 2. Values for $\underline{x}^*$.
We graph the function $\underline{x}^*$ as a function of α, for the case where Ex = .5. The value of α is on the abscissa, and $\underline{x}^*$ is measured on the ordinate.

Thus Ex = .5, and ATAH can be supported. For later purposes, we note that this equilibrium gives player B a payoff of 1 per period (the most he can hope for) and thus is certainly on the efficient frontier of all equilibria.

(I.8) *A simple equilibrium with promises.* Now suppose that player A observes x_t prior to the time that B must decide whether to trust A or not, and A is able to make the simple statement "I promise to honor your trust" before B acts. This statement is pure cheap talk, in terms of payoffs. We consider the following pair of strategies: A observes x_t and promises not to abuse B's trust as long as $x_t \geq 0$. B trusts A (along the equilibrium path) as long as A makes this promise, and only if A makes this promise. A keeps her promise along the path of play; she honors B's trust as long as she has promised to do so. If B trusts A when a promise has not been given, A abuses B's trust. If A abuses B after trust has been given, B never again trusts A.

For large enough discount factors, these strategies constitute a Nash equilibrium. B's actions have no affect on A's future strategy, and B is choosing short-run best responses, so (of course) B's strategy is a best response to A's.

As for A, we first compute her equilibrium value a priori (before learning the value of x_t): It is the solution to

$$v = \int \max(x,0)\, F(dx) + \alpha v,$$

which gives

$$v = \frac{\mathrm{E}x^+}{1-\alpha}, \quad \text{where } \mathrm{E}x^+ = \int \max(x,0)\, F(dx).$$

A has four possible contingent deviations:

(a) If A learns that $x_t = x \geq 0$, by following the prescribed strategy she nets (contingently) $x+\alpha v$. One deviation she can attempt is to refrain from promising, which nets $0 + \alpha v$, which is no better as long as $x \geq 0$.

(b) Or if A learns that $x_t = x \geq 0$, she can give the promise and then abuse B. This nets here 1 immediately and zero thereafter. Following the equilibrium is at least as good if $x + \alpha v \geq 1$, so a sufficient condition for this is that $\alpha v \geq 1$, or $\alpha \geq 1/(1+\mathrm{E}x^+)$.

(c) If A learns that $x_t = x < 0$, her contingent payoff from following the strategy above is αv. If she deviates by making the promise and then honoring B's trust (which will be given), she nets $x + \alpha v$, which is worse.

(d) And if A learns that $x_t = x < 0$, makes the promise, and then abuses B, she nets 1. The same condition on α that works for (b) works here.

We can also write out the equilibrium payoff for B, assuming B is long-lived and applies the same discount factor α: B nets

$$u = \frac{1-p_0}{1-\alpha}, \quad \text{where } p_0 = \mathrm{Prob}(x_t < 0).$$

(I.9) *Can simple promises be Pareto-improving?* For our particular parametric example, the simple promise equilibrium nets 8 for A and 9 for B. For the example, ATAH nets 5 for A and 10 for B, so the simple promise equilibrium is better for A than is ATAH, but worse for B. We are interested in the question, Is there some equilibrium without promises that does as well (in the Pareto sense) as this one? (We do not allow for side-payments, yet.)

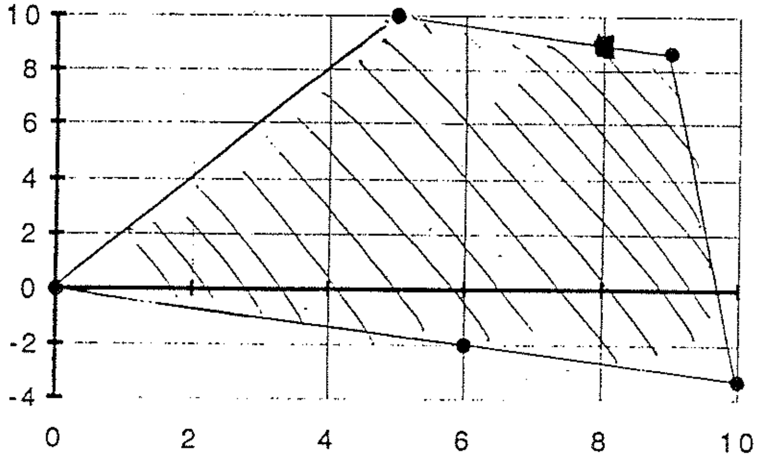
By *an equilibrium* (or, more generally, an outcome) *without promises*, we mean one in which B must decide (at date t) whether to trust A uncontingently on the value of x_t. That is, the essence of a promise by A (in our story so far) is that it allows A (in incentive compatible fashion) to warn B that circumstances are not propitious for trust.

(I.9a) In our parametric example, as long as $y < -1/3$, it is very easy to see that the simple-promise equilibrium is Pareto undominated by any feasible outcome without promises. The argument is simple: B has two (uncontingent) strategies in any stage: to trust A or not. A has four stage-strategies: To honor trust (if given) regardless of x_t, to abuse regardless of x_t, to honor if $x_t = -3$ only, and to honor if $x_t = 8/9$ only. The stage-expected payoffs for each of the eight possibility stage-strategy combinations are recorded in figure 3(a). Any feasible, let alone equilibrium payoff vector (for the repeated game) will be a positive affine combination of these eight vectors with weights summing to 10, and in figure 3(b) we graph the polygon of such combinations, under the assumption that $y = -1/3$; note that $(8,9)$ (indicated by the sqare) is just on the boundary of the set. Since $y < -1/3$, this strictly overstates what is feasible (as long as our combination involves any use of abuse, which $(8,9)$ does). This gives the result.

(I.9b) If y is just a bit bigger than $-1/3$, it will still be the case that the simple-promise equilibrium is Pareto-undominated by any payoff that can be obtained in an equilibrium without promises (without contingent actions by B). A sloppy rendition of the argument runs as follows: In order to get $(8,9)$ (even allowing for public randomizations, which will convexify the set of equilibrium payoffs), we need to be able to implement the point $(9, 26/3^{+})$. To obtain this payoff, A must be allowed to abuse B when $x_t = -3$, yet still be trusted. That is, B must

	trust A	don't trust A
honor regardless	5,10	0,0
abuse regardless	10,-10/3	0,0
honor if x = -3	6,-2	0,0
honor if x = 8/9	9,26/3	0,0

(a) Stage game expected payoffs



(b) The feasible region without promises

FIG. 3. Pareto undomination of the simple-promise equilibrium.

be willing to trust A no matter what transpires earlier. But in response to this strategy, A would abuse B in all cases. Put differently, if A is to honor B's trust when $x_t = 8/9$ (which is essential to this payoff), the continuation payoff for A if A honors trust must exceed the continuation payoff for A if A abuses trust. In fact, we can bound the difference in the continuation payoffs (the computation isn't important, but the difference must be at least .1234567890...). Since the same difference must apply when $x_t = -3$, the overall equilibrium payoffs to A and B must be strictly within the feasible (no-incentive-compatibility-constraint) sets (and we can put a bound on how much inside). Thus if $(8, 9)$ is just inside the boundary of the feasible set (which it is if y is a bit greater than $-1/3$), it will necessarily be outside the equilibrium set.

The sloppy argument just given raises the question, What is the set of equilibrium-feasible payoffs without promises? After very helpful discussion with Ennio Stacchetti, we're fairly sure that an Abreu-Pearce-Stacchetti style value iteration algorithm will produce the set; which in principle is a calculation one could do. We are a bit less sure, but still optimistic, that a similar computation would give all the equilibrium-feasible payoffs with promises. But neither calculation is straightforward [unless we are missing a trick], and a brute-force approach was not attempted.

(I.10) *What do promises add?* Now that we know that promises can give otherwise infeasible payoffs, our task is to explain why this happens and what they add.

Since we are unable to supply the precise sets of equilibrium payoffs without and then with promises, what follows should be regarded with some suspicion, But it seems to us that there are (at least) three qualitative ways in which the simple-promise equilibrium would add to what is feasible between A and B.

(I.10a) The first is that promises allow A to signal to B instances in which giving trust will be (jointly) too costly to be worthwhile. There are two ways in which the costs might be manifest: If B trusts A and A honors that trust, A receives x_t. Even if we assumed that A could verifiably reveal to B the value x_t and B could compensate A, as long as $x_t + 1 \leq 0$, the giving of trust that is honored is worse than foregoing trust in the current period. Second, if B trusts A and A abuses B, A gets 1 and B, y. As long as $y + 1 \leq 0$, this is worse than foregoing trust. So if $x_t < -1$ and $y < -1$, it is clearly in interests of neither A nor B to have B trust A. A, through the expedient uses of promises, can reveal to B (in self-enforcing fashion) that it is in neither's interests to have trust this period.

(I.10b) In addition, without promises and insofar as x_t cannot be reliably revealed to B, incentive-compatibility (making sure that A doesn't abuse B in all instances and blame it on low values of x_t) will lower what is feasible in equilibrium. Promises, by removing the need to enforce honor with continuation payoffs in case of abuse, will therefore be of some aid.

(I.10c) The third reason why promises can be valuable in the current setting is well outside the theoretical constructs considered so far. Think of a case in which y is fairly close to zero, close enough so that (perhaps) the simple-promise equilibrium is Pareto dominated by the set of no-promise equilibria. Even so, to sustain any efficient no-promise equilibrium outcome except for ATAH will take a fairly complex set of strategies, public randomizations (or something equivalent), and the like. The simple-promise equilibrium is very straightforward and (one presumes) would be relatively easy to learn.

(I.11) *Efficiency, sidepayments, and liquidated damages.* In the forgoing discussion, we have skirted the issue of sidepayments between A and B. The possibility of sidepayments, especially contingent on whether A abuses or honors B's trust, will pose significant interpretational problems for our results, problems that we will confront only at the end (when we lapse into arm-waving). But since the problems presented by sidepayments are bound to come up, it is probably better for us to head off those questions by taking a small detour and acknowledging the problems here.

(I.11a) We measure efficiency by sum-of-payoffs, under the hypothesis that payoffs are in monetary equivalents and sidepayments are theoretically possible.[4] We do not worry overmuch about the bargaining strength of the two sides, and (correspondingly) won't worry about the distribution of the sum-of-payoffs — insofar as (noncontingent) sidepayments are feasible, distributional issues are a side issue.

Then it is clear that overall efficiency depends on x_t and y as shown in figure 4. That is, we compare

$$1+x_t, \quad 1+y, \quad \text{and} \quad 0,$$

[4] To take this seriously, we ought to renormalize payoffs so that the benefits to B of honored trust and the benefits to A of abuse are not necessarily equal. But for now, we won't bother with this.

finding that trust and honor is efficient if the first is largest, trust and abuse is efficient if the second is largest, and no trust is efficient if the third is largest.

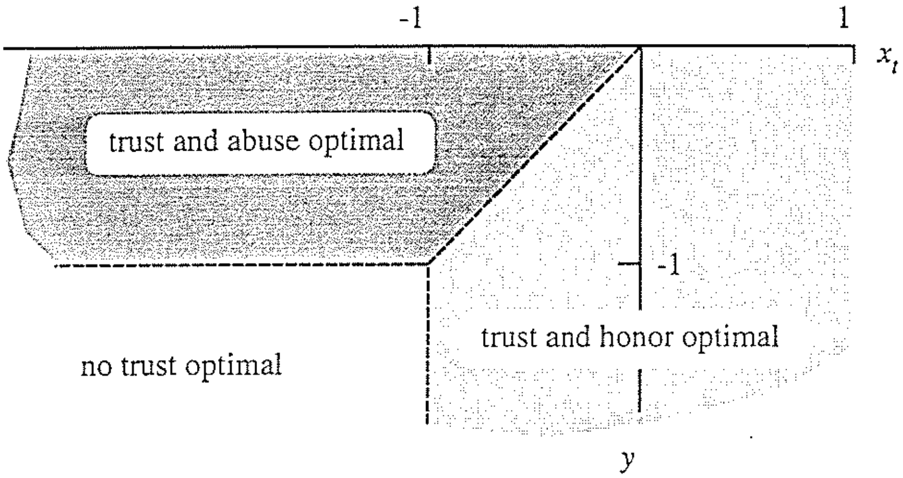


FIG. 4. Efficient outcomes depending on x_t and y.

(I.11b) To set the benchmark, in cases where x_t can be revealed to B ex post, there is nothing standing in the way of a fully efficient outcome (subject to the usual individual rationality constraints). So we will proceed assuming that x_t cannot be revealed to B.

(I.11c) *The case* $y \leq -1$. We *tend* to think in terms of cases where $y \leq -1$; i.e., where the injury done to B by abuse (relative to a base of no trade) is greater than the benefit B gets from trade.[5] Insofar as B could deal with other trading partners, and the benefit from dealing with this particular A reflects some slight comparative advantage in this trading relationship over others, this would tend to be true. We will proceed under this assumption, and then come back to it.

Under this assumption, trust and abuse is never efficient. Now in the simple promises equilibrium, we get trust (and honor) when $x_t > 0$. In our parame-

[5] In a more general formulation, the idea is simply that the benefits that A gets from abusing B less the cost to B from abuse should be such that this outcome is inefficient relative to no trade between the two.

terization, we had zero probability of $x_t \in (-1,0)$, so as long as $y \leq -1$, the simple-promise equilibrium is fully efficient (and will do better than any non-contingent choice by B, as long as the support of x_t intersects both $(-\infty,-1)$ and $(0,1]$). But if there is positive probability of $-1 < x_t < 0$, simple promises falls short of the socially efficient outcome.

With sidepayments contingent on whether a promise is given or not, this is easy to fix: We simply suppose that B compensates A with 1 if A promises good behavior, unless and until A abuses B after giving a promise. In this case, A, looking at x_t, will choose to give the promise as long as $1+x_t > 0$; her incentives to carry out the promise then will turn on the value of continuing relationship, much as before. (The cut-off discount factor is larger than before, since in this case she can get 2 immediately by giving a promise and then abusing B, while the value to her of the ongoing relationship doesn't increase by $1/\alpha$ or even by 1.)

(I.11d) *If* $y > -1$. What if $y > -1$? If sidepayments can be made contingent on honor/abuse, the obvious thing to do is to have a "liquidated damages" arrangement, in which A pays B the amount $-y$ if A abuses B, and B pays 1 to A if A is honorable. (To leave B with some surplus, simply shift these amounts up and down, respectively, by the same constant.) In the equilibrium, B always trusts A (no trust is never efficient if $y > -1$), so A's choices are between honoring B's trust, netting x_t+1, or abusing B, netting $1+y$. That is, with liquidated damages and a reward for honor, A fully internalizes the consequences of her actions, and we have full efficiency.[6]

(I.11e) To close off this excursion into contingent payments and liquidated damages, let us say the following: It is clear that efficiency can be enhanced within our models with these sorts of payments. Especially when $y < -1$, the enhancement is easy: A must be compensated for the simple act of giving a promise.

[6] What if y changes from period to period? We'll get to this later.

Notwithstanding this, we will proceed with the main lines of our analysis in a schizophrenic manner: we will assume that sidepayments are not possible, but at the same time we will measure efficiency with a sum of the payoffs approach.

(I.12) *Other no-promises equilibria.* In our previous analysis, we argued that promises can improve the situation over what we can get without contingent choices (whether to trust) by B in two ways: by comparing with ATAH, and by looking at the set of feasible payoff vectors. Ideally, we should be computing the set of all *equilibrium* payoff vectors, assuming that B doesn't get any information about x_t. As we noted, in theory this can be done, at least for pure strategy equilibrium (except for convexification with publicly observable randomizations); one would use the general results of Abreu, Pearce, and Stachetti (about self-generation) to compute the equilibrium set by value iteration. But as a practical matter, the computations are difficult, and it seems hard to guess the structure of efficient equilibria because of A's possession of private information which, by assumption, cannot even be revealed ex post.

Notwithstanding these difficulties, it will be helpful to expand our understanding of what can be achieved when B must act at date t without any signal from A about x_t, and so we proceed in this subsection to describe a one-dimensional family of equilibria of this character. But to reiterate (since it is easy to lose this in the welter of details to come): There is no reason to believe that any of these equilibria are efficient in the set of all equilibria.

To keep computational complexity to a minimum, we will assume that A and B have access to a publicly observable randomizing variable at the end of each period, which will be used in the implementation of these equilibria.

In these equilibria, there are two "states," a trust state and a nontrust state, which describe the actions of B. In the trust state, B trusts A. If A honors B's trust, the state is trust again in the following period. If A abuses B's trust, the state moves to nontrust with probability $1-\beta$ (which is parametric to the equilibrium, and which is determined, for simplicity, by a publicly observable randomization). A's decision whether to honor B's trust is determined by a comparison of x_t with some x_* (which is also parametric to the equilibrium) — if $x_t \geq x_*$, then A honors B's trust. If the state is nontrust, then B does not trust A, and the state next period is trust with probability β and nontrust with probability $1-\beta$. N.B., x_* will be determined by β and A's optimal-strategy constraint; the equilibrium is essentially parameterized by β alone.

Let p_* be the probability that $x_t < x_*$. Assuming this describes a Nash equilibrium, the payoffs to A starting in the trust and nontrust states, denoted by v and v' respectively, are determined by the following functional equations:

$$v = p_*[1 + \alpha(\beta v + (1-\beta)v')] + (1-p_*)\alpha v + \int_{[x_*,1]} x\, F(dx),$$

$$\text{and } v' = \alpha[\beta v + (1-\beta)v'].$$

(The usual logic applies.) The second of these can be rewritten

$$v' = \frac{\alpha\beta}{1-\alpha(1-\beta)}v,$$

which, substituted into the first, gives

$$v = \frac{p_* + \int_{[x_*,1]} x\, F(dx)}{1-\alpha[p_*\beta + p_*(1-\beta)\alpha\beta/[1-\alpha(1-\beta)] + (1-p_*)]}.$$

As for the payoffs to B, writting u and u' for his values in the trust and no-trust state, we have the functional equations

$$u = (1-p_*)(1+\alpha u) + p_*(y + \alpha(\beta u + (1-\beta)u')$$

$$\text{and } u' = 0 + \alpha(\beta u + (1-\beta)u').$$

After some manipulation, we get

$$u' = \frac{\alpha\beta}{1-\alpha(1-\beta)}u \text{ and } u = \frac{p_* y + (1-p_*)}{(1-\alpha) * (1 + [p_*(1-\beta)/\{1-\alpha(1-\beta)\}] + (1-p_*)]}.$$

Note that we do not get an equilibrium if x_* is set so high that, for the corresponding p_*, we have

$$p_* y + (1 - p_*) < 0;$$

in this case B would refuse to play along.

To compute these equilibria (or, at least, the efficient ones among them), it is easiest to take x_* as parametric and derive β. We won't take you through the derivation, but it turns out that, fixing x_*,

$$v = \frac{\int_{-\infty}^{1} \max(x_*, x)\, F(dx)}{1-\alpha}, \text{ and } \beta = 1 - \frac{1 - x_*}{\alpha[v(1-\alpha) + (1 - x_*)]}.$$

[Well, let us say something about why we can, somewhat miraculously, write down the equilibrium value as a function of one of two defining variables. The key comes in our parameterization, and more specifically in the fact that if $x_t \leq x_*$ (so A will abuse B), A's contingent payoff is independent of x_t. Since we set β so that A is indifferent between abuse and honor at x_*, we can effectively replace outcomes of x_t less than x_* with x_* (this takes a martingale continuation argument to be made precise). Thus we can compute the value under the assumption that we never meet $x_t < x_*$, in which case there is never any abuse, and things are very simple — the formula for u should now be obvious.] This applies as long as β so computed is positive; at the point where β becomes negative, an equilibrium (with critical value x_*) cannot be supported.

(I.12b) For example, in the simple example from before, we have $x_t = 8/9$ with probability .9 and $= -3$ with probability .1. Setting $x_* = -3$ is essentially recovering ATAH.[7] We know that this equilibrium gives values 5 to A and 10 to B. It is relatively straightforward to calculate the equilibrium for $x_* = 8/9$; we get $v = 80/9$ and $\beta = .8765\ldots$. The value of v depends on y, of course;

[7] If you calculate β from the formulae above, you will get the largest value of β needed to enforce honor when $x_* = -3$; an infinite punishment is not needed.

if $y = -2$, we get $u = 6.9136$. Compare with the (first-best, as long as $y \leq -1$) simple-truth equilibrium, which gives payoffs 8 and 9 for A and B.

(I.12c) A comparison of these equilibria (and their value functions) with the simple-promise equilibrium does not give very clean or clear results: The only clear comparison is for the case $x_* = 0$. Then simple-promises is clearly superior: A gets precisely the same payoff in both equilibria, but B is worse off in the $x_* = 0$ equilibrium on two grounds: He pays $-y$ in cases where $x_t < 0$, and his (presumably) profitable subsequent trades with A are forgone, at least for a while. (In terms of the value function u, the first lowers the numerator, and the second lowers the denominator.)

Moving away from $x_* = 0$ muddies the waters. A unambiguously prefers to raise x_*, and if left to choose among these equilibria, she would choose x_* corresponding to a zero value for B (where $p_* y + (1 - p_*) = 0$). But increasing p_* decreases the numerator in B's value function; and while it also decreases the denominator (by raising β), in all cases we've computed B unambiguously prefers lower x_*. But the sum of the two payoff values takes on no particular pattern. Examples can be found where the sum of values rise with x_*, where they fall with x_*, and where they are not monotonic in x_*. (As y gets more and more negative, lower x_* are relatively favored, of course; in all examples we've examined with $y < -1$, the sum falls in x_*, but if this is a proposition, we am far from a proof.)

Finally, there is no clear pattern between the payoffs or their sum in the simple-truth equilibrium vs. the "best" of the x_* equilibrium. Examples can be produced where the sum in the best x_* equilibrium exceeds the sum in simple-truth, for x_* at the top of its distribution; where the sum in the best x_* equilibrium exceeds the sum in simple-truth for x_* at the bottom of its distribution; and where simple-truth dominates. (For the latter, take any case in which $y \leq -1$ and the support of x_t excludes $(-1, 0)$. Then we know simple-truth will be first best.)

(I.13) *Post mortems for the simple model.* As noted earlier, the simple model suggests three roles for promises. The first, and clearest, is that they allow A to communicate in self-enforcing fashion with B information that is in both their interests. We can think of this as relating to two extant literatures:[8] (1) The use of cheap-talk communication in games with some, but not complete, commonality of interests. This goes back to Crawford and Sobel (Econometrica, 1982). For a contextual application, see recent work by Farrell and Gibbons. (2) More specifically the use of cheap talk in repeated interactions: Ben Porath and Kahnemann (1993) analyze a case in which players take actions that (all) others do not see, and they obtain conditions (on the observability by one or by two other parties) under which the folk theorem still holds. Lehrer has done a very general analysis of cases with hidden actions, although we are unsure whether he investigates the role of cheap talk. We are vaguely away of general work on this subject by Matsushima and Matsui (which is to say, Kreps has seen a presentation, but can't find a copy of the preprint). And this has been the subject of some analysis in specific contexts: such as Sobel's credibility game (although this isn't for an infinitely repeated game), and the study of the revelation of insider information, by Benabou and Laroque (1992).

The second role for promises is that because they allow for communication, they remove to some extent the need for inefficient "punishment" of A by B when A abuses B. These punishments are not always needed — without promises, A and B have access to ATAH — but in general ATAH is itself inefficient (and, for $\underline{x}$ low enough, ATAH isn't an equilibrium outcome), so that without communication, B must do something to keep A from abusing him all the time and claiming low x_t.

Third, and somewhat related to the second (although not in terms of formal

[8] We would appreciate any additional references.

theory), promises give us relatively simple equilibria. (We'll lose that virtue in a minute.)

The problem with simple promises in the forgoing model is that they themselves are inefficient. Essentially, A has no incentive to give a promise in cases where the cost to him of honoring the promise makes him worse off than in the no-trade outcome but not so worse off that, together, A and B are jointly better off (using a sum of payoffs). Also, for cases where $y > -1$, it might be efficient to have breach of a promise, but except for the possibility of liquidated damages, we haven't got a grip on how to allow for this.

We'll return to these general postmortems in a bit. But before doing so, we want to enrich the basic model a bit.

II. Broken promises

In the very simple models above, promises given are always kept. This is not much in accord with observation of the real world, and as a first variation on the basic model, we consider what it will take to get promises that are not always kept.

One route to this end is to consider more complex equilibria than the ones we've looked at, but without changing our formulations. Rather than do this, we will complicate the formulation, so that it is inevitable that promises will sometimes be broken.

(II.1) The complication runs as follows. Up to this point, we've assumed that A knows x_t at the time when she can offer her promise or not. Hence she knows what her later costs and benefits of honor will be, and (in a pure strategy equilibrium) she can forecast whether she will honor or abuse B.

Suppose instead that she only has some indication of what will be the eventual value of x_t. To keep matters as simple as possible (they will still be fairly complex), suppose that x_t will take on one of two values, $\overline{x}$ or $\underline{x}$; think of

$\underline{x} < -1$ and $0 < \overline{x} < 1$, so that if $x_t = \underline{x}$, it is socially inefficient to honor trust (relative to not having trust offered), but if $x_t = \overline{x}$, A would prefer to be trusted (and honor that trust) to forgoing trust. In fact, we will go a step further and assume that $\underline{x}$ is so low that A will, more or less automatically, abuse B if B trusts her and $x_t = \underline{x}$.

(II.2) We assume that at the time that B must decide whether to trust A or not, A does not observe x_t, but instead receives some private information which allows A to reassess the probability that $x_t = \underline{x}$. We let p_t denote the probability assessed by A on the basis of her information; so that p_t has expectation equal to the prior probability of $\underline{x}$. If B trusts A, then A learns the true value of x_t, and either honors B's trust or not.

(II.3) The form of the equilibrium we will investigate runs as follows. Play begins in a "possible-trust" phase: A observes p_t. If $p_t \leq p_*$, A issues a promise to B that she will honor his trust if he trusts her. If $p_t > p_*$, A does not give this promise. B trusts A if and only if the promise is given. If no promise is given, we move to the next date, remaining in the possible trust phase. If the promise is given, so that B trusts A, A observes x_t and responds accordingly: if $x_t = \underline{x}$, A abuses B; if $x_t = \overline{x}$, A honors B's trust. If A honors B's trust, we again move into the next stage in the possible-trust phase. If A abuses B, however, we enter the next period in the possible-trust phase with probability β; with probability $1 - \beta$ we enter a "punishment phase." In the punishment phase, B does not trust A, regardless of what A says, and the phase in the subsequent period is the "possible-trust" phase with probability β. Random phase transitions are according to a publicly observable randomization.

(II.4) This equilibrium is effectively parameterized by β, which is a surrogate for the length of punishment inflicted on A if she ever abuses B's trust. The smaller is β, the more sure A will have to be that she will be able to honor B's trust, before she will promise to do so. Thus β will endogenously determine p_* (much as β endogenously determined x_* in the previous section).

Letting v be A's value starting in the possible-trust phase, and letting v' be her value starting in the punishment phase, we have the following functional equations for v and v' (in terms of p_* and β):

$$v' = \alpha(\beta v + (1-\beta)v'), \text{ and}$$

$$v = (1 - G(p_*))(0 + \alpha v) + \int_{[0,p_*]} [(1-p)(\overline{x} + \alpha v) + p(1 + \alpha(\beta v + (1-\beta)v')] \, G(dp).$$

The first of these is

$$v' = \frac{\alpha\beta}{1-\alpha(1-\beta)} v.$$

As for the second, we write $\mathbf{I}(p_*)$ for $\int_{[0,p_*]} p\, G(dp)$, and the second functional equation (replacing v') becomes

$$v = \left[\alpha - \frac{\alpha(1-\beta)(1-\alpha)}{1-\alpha(1-\beta)}\mathbf{I}(p_*)\right] v + G(p_*)\overline{x} + \mathbf{I}(p_*)(1-\overline{x}).$$

Thus

$$v = \left[G(p_*)\overline{x} + \mathbf{I}(p_*)(1-\overline{x})\right] \Big/ \left[(1-\alpha)\left(1 + \frac{\alpha(1-\beta)\mathbf{I}(p_*)}{1-\alpha(1-\beta)}\right)\right].$$

A given p_* and β constitute an equilibrium if, when $p_t = p_*$, A is just indifferent between giving the promise or not. If she fails to make the promise, she nets αv. If she does make the promise, she nets

$$p_*(1 + \alpha(\beta v + (1-\beta)v')) + (1-p_*)(\overline{x} + \alpha v)).$$

Equating these two gives the equilibrium condition[9]

$$p_* + (1-p_*)\overline{x} = p_* \frac{\alpha(1-\beta)(1-\alpha)}{1-\alpha(1-\beta)} v.$$

[9] With the following warning: This is the equilibrium condition for an interior value of p_* if the support of p_t extends to a bit to right of p_*. If there is some interval $(p^*, p^* + \epsilon)$ that has null intersection with the support of p_t, then the equality should be an inequality (lhs $\geq$ rhs) with the reverse inequality at the next highest point in the support of p_t, changing p_* but not v.

After manipulation, this gives

$$\beta = 1 - \frac{p_* + (1 - p_*)\overline{x}}{\alpha(p_* G(p_*)\overline{x} + p_* + \overline{x} - p_*\overline{x} - \mathrm{I}(p_*)\overline{x})}.$$

In addition, we have to check that A will be willing to honor B's trust if x_t turns out to be $\overline{x}$: This is the inequality condition

$$\overline{x} + \alpha v \geq 1 + \alpha(\beta v + (1 - \beta)v') = 1 + \frac{\alpha\beta}{1 - \alpha(1 - \beta)} v$$

which is

$$1 - \overline{x} \leq \frac{\alpha(1 - \alpha)(1 - \beta)}{1 - \alpha(1 - \beta)} v.$$

And when we have an equilibrium pair p_* and β, the value function for B (saving you the algebra) is

$$u = \left[G(p_*) + \mathrm{I}(p_*)(y - 1)\right] \Big/ \left[(1 - \alpha)\left(1 + \frac{\alpha(1 - \beta)\mathrm{I}(p_*)}{1 - \alpha(1 - \beta)}\right)\right].$$

(If this is negative, we don't have an equilibrium.)

Although it is not evident from the algebra, it seems fairly clear that β will be rising in p_*. It also seems fairly clear that A will prefer higher cutoffs and B will prefer lower ones. But we have not had time to do the analysis to support this intuition.

(II.5) We close with an example. We continue with the parametric example in which $x_t = 8/9$ with probability .9 and -11 with probability .1. This means that $\mathrm{E}x$ is negative — ATAH is certainly not an equilibrium — although this is unimportant; the main thing is that A will certainly not honor B's trust, no matter what this means for the future, if $x_t = \underline{x}$ (as long as $\alpha = .9$, which we assume).

To set benchmarks, if A manages to learn x_t prior to B having to decide whether to trust A or not, and if A can give a promise, we have the simple promise equilibrium with gives the (first-best) outcome of 8 for A and 9 for B.

And, without communication, we have the $x_* = 8/9$ equilibrium; to enforce honor when $x_t = 8/9$, we need $\beta = .87654\ldots$. The payoff for A is $80/9$. The payoff for B depends on y: We will consider the case $y = -8$, in which case we get $u = .9876543\ldots$, and the sum of the payoffs is $9.876543\ldots$.

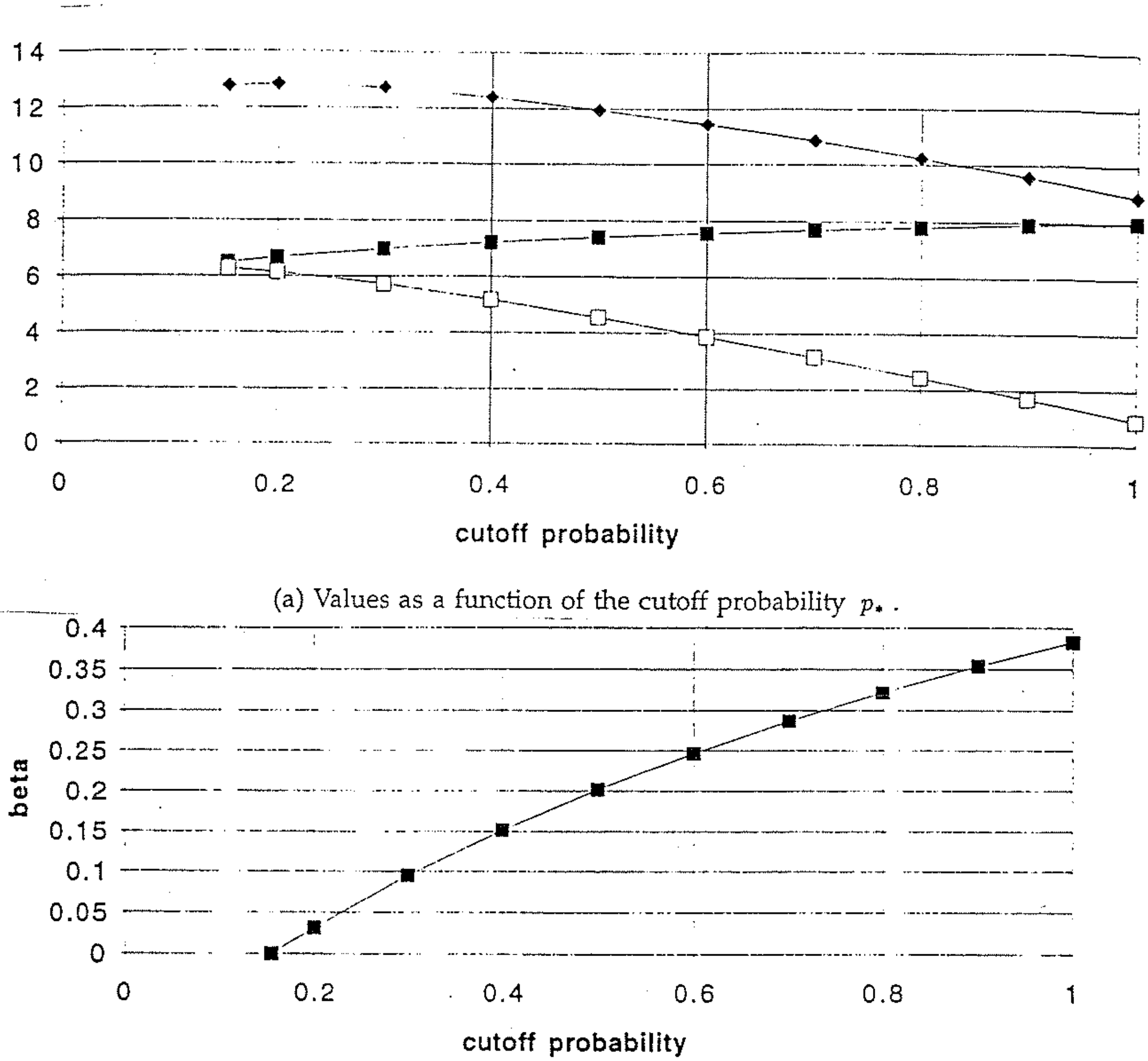
We assume that at the time B must decide whether to trust A (in period t), A has information that permits her to assess probability p_t that x_t will equal -11 (and she will be forced to breach her promise). We assume that p_t has distribution function $G(p) = p^{1/9}$ for $p \in [0,1]$; this has expectation .1. We spare the detailed calculations, and summarize the results in figure 5. In figure 5(a) we graph A's values (solid boxes), B's values (open boxes), and the sum (diamonds), for different cutoff values; in 5(b) the corresponding values of β are graphed, for the range $p_* \geq .1548303$ (approximately). (For $p_* < .1548303$, β becomes negative, so this is the lowest cutoff that can be enforced). The sum of payoffs is largest in the interior, but at a relatively low cutoff level of around .2.

(It is potentially interesting that the values for $p_* = 1$, where A always promises B and B always (in the potential trust phase) trusts A, give lower payoffs than does the $x_* = 8/9$ equilibrium. This happens (as far as we have been able to tell) because the additional constraint that A is *just* content to make the promise at $p_* = 1$ is binding and the constraint that she will be willing to honor B's trust when $x_t = 8/9$ is slack. In fact, if we increase the value of β above .38 or so, A will still want to make the promise for all p_t, and the second constraint remains slack until $\beta = .87654\ldots$. See the previous footnote concerning this point.)

III. Private info about y and promises of varying strength

To this point, we have assumed that A knows y. Hence unless B, hearing a statement from A, is just indifferent between trusting A and not, A will know in equilibrium the consequences of the statements she makes.

Suppose we enrich the story a level further by supposing that, in period



(a) Values as a function of the cutoff probability p_* .

(b) β as a function of p_*

FIG. 5. An example with broken promises.

t, the payoff to B if he trusts A and she breaches is y_t, a random variable, observed by B but not by A. Assume that the $\{y_t\}$ form an i.i.d. sequence, with H the distribution function of y_t.

It might be that there is some role for cheap-talk by B concerning y_t, given at the same time that A makes her speech based on x_t. But we will ignore that possibility and consider instead the following structure of actions: A receives some information about the future value of x_t, and makes a cheap-talk statement about that. B, hearing and interpreting A's statement, and aware of the current value of y_t, decides whether to trust A. A then learns the value of x_t, and decides (anticipating future actions) whether to honor or abuse B (if B has trusted her this round).

We will also keep to the relatively simple case of section II, in which x_t is either $\overline{x}$ or $\underline{x}$, and A's information about x_t takes the form of a probability assessment p_t, drawn according to some distribution function G. We assume that $\underline{x}$ is so low that A is bound to breach if $x_t = \underline{x}$, and we will be constructing equilibria in which A will honor B's trust if $x_t = \overline{x}$.

Up to this point, A's cheap talk have been bivariate; either A promised good behavior (with high probability) or not. And in the equilibria, B trusted A when and only when A promised good behavior. In the new story, things are richer in two ways. First, B may trust A absent a promise, if y_t is close enough to zero, and he may decide against trust even in the event of a promise, if y_t is very far from zero. Second, now there is a role for more than two possible messages from A; A will (in general) want to induce B into trust when A's information p_t is close to zero, and A may want to give stronger assurances that she can be trusted for very low values of p_t.

(III.1) *Form of the equilibrium.* We look for equilibria taking the following form. We begin in a "potential trust" phase. A has a vocabulary of messages $\{m\}$ she can send. We do not preclude the possibility that the language is as rich as the real line; perhaps A's message is "I'm looking at $p_t = p$" for all p in the support of G. B acts myopically in period t (hence this story works best if you think of there being a sequence of B s, although we will compute values for B with a discounted sum); in equilibrium, corresponding to each message m is a (measurable) set of p_t, and B correctly infers from m what is the probability that A will breach. B takes this message-dependent probability and decides, on the basis of y_t, whether to trust A. If he does not trust A, then we proceed to the next period in the potential-trust phase. If he does, then A sees x_t and decides whether to honor B's trust or not. If she honors B's trust, we move to the next date in potential-trust. If she abuses his trust, then a publicly observable randomization is conducted such that with probability β_m we stay in potential trust while with probability $1 - \beta_m$ we move to "punishment-m." While in the punishment-m phase, B refuses to trust A no matter what is y_t (and A would abuse B for sure and without penalty if he did trust her); and at the end of the period we use the β_m randomization to determine whether potential trust is restored or punishment continues. We also allow for one message m^0 which inspires distrust; i.e., if A sends the message m^0, she is announcing that she will abuse B whether her x_t is $\overline{x}$ or $\underline{x}$. In equilibrium, B will certainly refuse to trust A in this instance (we assume that $y_t < 0$ with probability one). And should B be silly enough to trust A after hearing m^0, once she abuses him, the next stage begins in the potential trust phase.

It should be noted that the equilibria of section II take the form of these equilibria, but with only two messages, m^0 and some message which definitely inspires trust. If there is no uncertainty about y_t, we aren't quite forced into an equilibrium of this form — we might have one more message, which makes B just indifferent between entering or not, and B randomizes in consequence. Notwithstanding this somewhat perverse possibility, we assert that in spirit if not in fact, if we are going to have a theory with promises of different strengths (calling for different levels of punishment if they aren't kept), then we essentially must have some uncertainty about how B will respond to them; i.e., uncertainty about y_t is virtually necessary.

Concerning the publicly observable randomizations: In a more realistic formulation, we would suppose that different strength promises would result in punishments for varying but deterministic lengths of time. That is, if A promises good behavior and subsequently abuses B, B forgoes exchange for 5 periods;

whereas if A promises on the souls of her parents that she will behave and then subsequently abuses B, B foregoes exchange for 25 periods. But it is analytically more convenient to set (approximately) $\beta = 1/5$ in the first case and $\beta = 1/25$ in the second. Obviously, with a discrete set of punishment levels, we won't be able to get separating equilibria. But the spirit of what one gets (we assume) isn't much different, and the analytics are much more complex.

Concerning the formal game model: As part of the game, we have to specify the set of messages available to A in advance, and as part of any strategy profile, we have to say how B will react to messages that are unsent in equilibrium. Because the talk is cheap (except endogenously), we will not worry about such things, thinking (formally, at least) that unsent messages lead to infinite withdrawal from trade. Hence as long as some message is sent in equilibrium, unsent messages will be avoided.

Finally, we note in advance the strong stationary character of the equilibrium. It is easy to imagine strategies in which A's use of the messages depends on how trustworthy she has been in the past. But the stationarity we have assumed is essential to our abilities to make progress on the equilibria, so we ignore all such (alternative) possibilities.

(III.2) *Structural properties of the equilibrium.* The resemblance of an equilibrium of the sort described above to a standard signalling model should be clear. Suppose we have an equilibrium, with messages m^1 through m^K, arrayed so that the corresponding β s, which I will denote by β^1 through β^K, are strictly decreasing. (The case where two signals have the same β requires some special pleading which we won't bother with for now.) Corresponding to each m^k is a set of p_t that would choose this message, and we write p^k as the conditional probability that $x_t = \underline{x}$, conditional on this set. (If for some p_t a randomized message is sent, compute p^k in the obvious fashion.) Since (by assumption) B acts myopically, the probability that he will trust A having heard m^k is the probability that

$$y_t p^k + (1 - p^k) \geq 0, \text{ or } 1 - H((p^k - 1)/p^k).$$

(If there is positive probability that $y_t = (p^k - 1)/p^k$, some randomization may be employed. It is therefore evident that the sequence p^k will be strictly decreasing in k; if $p^k \leq p^{k'}$ for $k < k'$, then the signal $m^{k'}$ would never be send, since m^k gives no smaller probability that B will offer trust this period, and strictly lower penalties for failing to honor trust. Moreover:

Proposition. [10] *In any equilibrium of the sort under investigation, for messages ordered as described above, if A will send m^k when $p_t = p$, then she will send a message with no higher index whenever $p_t \leq p$.*

In other words, the "types" of A separate into neat intervals, with the highest cost types sending lower index messages.

(III.3) *Fully separating equilibria?* Following Crawford and Sobel (1982), it is natural to wonder how finely we can divide up the "types" of A in an equilibrium. Recall that in their setting, there was a limit to number of elements in the the cheap-talk partition that could be induced in an equilibrium.

Conjecture. [11] *For some specification of this model, where p_t and y_t each have nonatomic distribution, we can get a fully separating equilibrium.*

Recall that in Crawford and Sobel, the reason for the limit on the size of the partition is, essentially, that the receiver and sender had different interests. In a fully separating equilibrium, the receiver would know the sender's type, but then the sender would wish to dissemble at least to some extent. That reasoning

[10] We haven't written out a complete proof, so this is only a fairly strong conjecture at this stage.

[11] If section III.5 to follow is correct, we assign this conjecture probability .8 of being true. But we are not quite there yet.

doesn't apply here because in this case, the "signalling" is somewhat based on endogenous waste; i.e., the delay in return to potential trust, once breach has occurred.

(III.4) *Efficiency?* Crawford and Sobel also find that the finest of their equilibria is the most efficient, ex ante. We conjecture [12] that this will fail to be the case here or, rather, parameterizations can be found for which the most efficient equilibrium (sum of the expected payoffs) involves an intermediate (finite) number of signals — even if complete separation is possible, it will involve too much wasteful separation by β. [13]

(III.5) *Construction?* In the equilibria of sections I and II, we were able to write down in essentially closed form what they must be. We don't rule out the possibility that this can be done for the equilibria described here — at one point we were convinced that we couldn't write down the equilibria in part II in closed form, so our intuition on these things is not so good — but as we write this, we can't do it.

Nonetheless, we believe we have an algorithm that computes equilibria of this sort, assuming they exist. (The algorithm will tell you when a particular "equilibrium" cannot be constructed.) Begin by partitioning the support of p_t. (We'll assume that p_t has nonatomic distribution, although handling atoms is not too big a problem. Also, We'll proceed as if we are aiming for a finite partition equilibrium) For each interval, compute the value of p^k and find the probability that the B will deal with this interval in the equilibrium. Decide whether the "no trust" message m^0 is to be sent by the bottom partition.

Conjecture $v = 1/(1-\alpha)$. (This is obviously too high.) Take the interval

[12] Probability = .55.

[13] On the other hand, if signalling could be done via liquidated damage amounts, it seems likely that among equilibria of this sort, perfect separation would be relatively the most efficient. But we're really shooting from the hip at this point.

sending the message m^1, and find β^1 just low enough so that (given v as above) these types are induced to honor trust when $x_t = \overline{x}$. Then use the right-hand endpoint of the interval and the probability of trade given by the next interval to find the value of β^2 that just keeps this endpoint value of p_t indifferent between m^1 and m^2. And proceed up the ladder. Note that we are computing a "most efficient" equilibrium for the prespecified partition of the support of p_t. This can terminate one of two ways: Either we "run out of room" in the sense that β^k goes through zero before the last interval in the partition is considered, or we get through to the last group. In the first case, we must give up. In the second, compute v using these values of β^k. Clearly the value of v that you will get will be less than $1/(1-\alpha)$; use the new value of v to redo the process. We believe (and, we stress, we believe) that the lower is the guessed at v, the smaller will be each β^k in turn, hence the smaller will be v that is derived at the end. If in any iteration you run out of room in the β s, no equilibrium with this partition can be constructed. Otherwise, this will converge monotonically to the most efficient equilibrium for this partition.

IV. Discussion: Promises, transfer payments, liquidated damages

From the perspective of a mathematical development, it seems reasonably clear that there is *potentially* enough in the forgoing analysis to, say, get a publication in a middle-level theory journal. To make it into a top-level theory journal, we will probably need to sharpen up our characterizations, and even figure out how to extend Abreu-Pearce-Stacchetti fully to these contexts, so we aren't in the somewhat *infra dig* business of spending time and pages analyzing a class of nicely structured but probably inefficient equilibria.

Notwithstanding this, we believe the forgoing analysis falls on its face along a different dimension. We have already noted that, in the case where $y < -1$ and x_t is known to A, we would get the first best outcome by paying A for the making of a promise (which would then, of course, put A's reputation in jeopardy). It is easy to imagine that when we relax these strong assumptions on y and x_t, we would still promote efficiency by paying A to make a promise; the idea is clear: A's decision to promise or not depends on her own evaluation of the chances she will be punished as against the surplus she forgoes if she doesn't make the promise (and get the business this period). But from a social point of view, she ought to give some consideration to the surplus she will (potentially) provide B. A payment for making a promise (with a corresponding increase in the penalty for breach) would push in this direction.

Moreover, and worse, in the models or, more precisely, in the equilibria above, breach is penalized by a withdrawal of trade for some length of time. This is manifestly inefficient, relative to a system of penalty transfers from A to B in the event of breach — something like a system of liquidated damages. Both A and B suffer by a suspension of trade (assuming we have an equilibrium where both obtain surplus from the outset). Why this sort of penalty, when/if direct compensatory damages can be paid?

N.B., this problem is one that goes far beyond the current context of a theory of promises. In many applications of the folk theorem, we have to wonder why the punishment will be carried out, if it will hurt both sides. The obvious gratuitous reference to the sizeable literature on renegotiation is in order, but we think it goes beyond this: In some applications of the folk theorem — i.e., to implicit cartels — compensatory sidepayments are illegal. Depending on what is observable by the players, in other contexts they may be hard to implement. In standard agency problems, risk aversion may prove a problem. (References both to Fudenberg-Holmstrom-Milgrom and to Polinsky (JLS, 1983) come here.) But we feel fairly comfortable in saying that in some contexts, direct compensatory damage payments are too quickly dismissed. The current context is particularly favorable for thinking about why/how they can be dismissed on more solid grounds.

(IV.1) The discussion should begin with a proposition. We'll dignify it with that name, although it isn't in state to be proved, just yet.

Proposition. *In any of the models above, if A and B can engage in sidepayments contingent on the observables (whether A makes a promise to B and, in the multi-strength case, what is the strength of the promise given; whether B trusts A or not; whether A honors B's trust or breaches), if A and B's payoffs are measured in those sidepayments (so, in particular, A and B are risk neutral in the sidepayments), if neither A nor B face immediate liquidity constraints, and if there is no frictional loss in making sidepayments, then self-enforcing contracts calling for contingent sidepayments can do no worse than "promises." Moreover, insofar as the promises-equilibrium involves positive probability of breach and (hence) positive probability of some punishment of A by the withholding of trade, contingent sidepayments can do strictly better.*

This needs to be made more precise, but we think it has content, lent to it by the list of assumptions and by the term self-enforcing. To take the latter part first, we have in mind a case in which the "contracts" are not enforced by the courts but instead by self-interested behavior by A and B. In particular, when A is called upon to pay compensation to B, then A will do so, under the threat of a withdrawal of trade from B. The point, and the content, is that whenever (without sidepayments) A is to suffer a loss in payoff from a withdrawal of trade by B, it is better for B and no worse for A to take a direct payment in that amount, where if A reneges, B can always resort to the withdrawal of trade. Indeed, insofar as this increases the efficiency of the $A-B$ relationship and those gains are shared between A and B, it enhances the value of the ongoing relationship and so makes the (never-to-be-used) second level of punishment-by-withdrawal all the more to be avoided.

Obviously, frictional losses in making sidepayments will compromise this result. If A and/or B faces short-run liquidity constraints, such that they can't borrow (at rate α) against the value of their ongoing relationship, punishments stretched out over time may be more powerful. (See Boot, Greenbaum, and Thakor, 1993)

As for the assumption about risk neutrality, we leave it there in case it is needed (although we don't quite see the need) — if we want to extend this result to say that, with these contingent sidepayments we can reach the first-best efficient outcome, then risk neutrality will certainly play a role: cf. Fudenberg-Holmstrom-Milgrom and Polinsky, for example.

(IV.2) Yet when X asks Y whether she would give a seminar at some date six months hence, Y promises to do so without any discussion of the damages that would be paid if Y fails to show up (or if, after the seminar, X announces that he has been had). Although X is a forgiving soul, Y may well imagine that had she phoned in sick the morning of the seminar, X would have considered this the breaking of a promise, and Y would have suffered a withdrawal of "trade" with X for some period of time. Why don't we have a simple schedule of damage payments — phone in sick, and send X a check for whatever is appropropriate?

To put the matter fairly strongly, all those earlier functional equations aren't much of a theory of promises unless we can explain this.

(IV.3) The first thing to say is that if Y breaks her promise, there are ways she can (presumably) accelerate X's forgiveness and end the inefficient punishment cycle. That is, a check in compensation may not be called for, but that doesn't mean that compensation can't be (and isn't) given when a promise is broken. (Just to be clear, We are not concerned here with promises that are broken for reasons that can be credibly explained. We want to stick, in these ramblings, to a case where, effectively, x_t is and remains private information to the promise giver.)

(IV.4) In a related vein, observe that the efficiency gains that might accrue if X could be paid for making a promise may actually be present as well. Y asks X to give a talk, and it was clear to both parties that this would be somewhat costly to X. By promising to give the talk, X hopes to cause Y to entertain warm and friendly feelings, so that if at some future date when X asks Y for a favor, Y would compensate with a promise of his own. Putting this more strongly than is (perhaps) appropriate, we are probably less concerned about payments for giving promises when the parties involved have a mutually beneficial relationship, in which each does things for the other. In such cases, the compensation needed to get us closer to first-best may be there in somewhat "hidden" form. (In terms of a formal model, imagine in the simple-promises setting that at alternate dates X asks favors from Y which put X in jeopardy, and think of constructing an equilibrium in which X and Y are more inclined to give a promise (consent to do a favor) the more the other has done so; hence (looking forward) each is more inclined to give a promise in order to build up credit.)

(IV.5) The previous two paragraphs essentially argue that promises may not be as inefficient as our simple formal models have suggested, but they do not change the basic observation that direct sidepayments would be at least as efficient and perhaps even more so. So we return to the questions, Why doesn't X give Y a check when she promises to give a talk, and why don't they set (or at least understand) that Y can, for some predetermined amount, gracefully bow out.

One argument might be that the cost of negotiating these amounts would soon swamp the two, particularly insofar as in most cases the promise will be kept and negotiations over compensatory damages will be wasted. But this isn't convincing; since the equilibria we have described involve both parties understanding how long a withdrawal from trade is required in the event of breach. There is no reason we can think of why one computation would be easier than the other.

A second argument is that contract law gets in the way. We would want to set compensatory damages in a way that depends on Y's incentives and not on the damage done to X, while courts award damages based on damages done (or so we understand). But if we set the damage level in the contract a priori, as a level of liquidated damages, the courts will not intervene. (Can party Y, having signed a contract with a large level of liquidated damages called for in the event of breach, go to the courts and ask that the contract be set aside because the level of damages does not correspond to the damage done? Any legal scholars reading/hearing this?) Moreover, court enforcement shouldn't be an issue, insofar as we are looking for *self-enforcing* compensatory damage payments.

(IV.6) While we do not like either of the previous two arguments, we can offer three and one-half possible explanations, all of which are easy to model in an economic model, but only if you assume the conclusion.

The first involves incomplete information about "type" and reputation. Suppose there are saints out there who keep promises even when it is personally costly to do so. Such saints avoid relationships that are tied up with negotiations about liquidated damages on the like, because they find such things to be too crass and materialistic. Even if the percentage of saints in the population is small, in the usual way we can find a larger percentage of members of a society acting saintly, in order to leave open the possibility that they are really saints. And their desire to protect this reputation gives their trading partners a measure of assurance. Of course, to act saintly (we are assuming, and it is nothing more than that) is to forgo liquidated damage negotiations and payments. Hence Y wouldn't dream of sending X a check for canceling; to do so would reveal her as being the sort... who would do so.

(IV.7) Explanation number 2 concerns pure expectations. In both the promises and the compensatory-damage award "equilibria," A's motivation to keep her promises/pay called-for compensation is based on the expectation that B will continue to trade with her if she does so (and will call a halt to trade, at least temporarily, if not). Y is motivated to keep her promise to give a paper because she puts a value on her future relationship with X, based on an expectation that he will be around to trade with in the future. This is purely expectational, and it is entirely possible that her expectations concerning her future relationship with X are colored by the character of that relationship. That is, as long as X and Y engage in trade based on promises, small personal gifts (Y buys the coffee today, X does tomorrow), and the like, then we both expect that we will continue to have this sort of relationship. But if, when X asks Y to give a seminar, he also begins to specify damages to be paid to him if Y breaches, Y might have constructed this as an indication that X was thinking of early retirement. Accordingly, Y might have given the promise, fully intending to breach if the costs of fulfilling were moderately high, and expecting as well that when/if X came for his liquidated damages, she would tell him to take her to court. X, understanding all this, would not trust a promise backed by negotiations over liquidated damages.

This is something like the previous explanation, except the previous argument is based on a "seed"; a small percentage of saints within the population. This argument is completely self-starting; based on the idea that all these equilibria depend to some extent on individuals' expectations of the future; and the social context in which a relationship is conducted can affect those expectations. Of course, there isn't much of an economic theory to this, at least unless and until we take this into some sort of adaptive learning story, since the story turns on the idea that form affects expectations, and thus content, in a particular way.

(IV.8) The half-story builds on number 2. We have considered models in which Y deals with a single X, and Y's incentive to keep her promises are that X will withdraw his trade. Imagine that X engages in the relationship only once, to be replaced by a new X next time. Or imagine that X's trade doesn't matter all that much to Y; X alone doesn't have the power to materially harm Y. Even so, we know that we can construct reputational equilibria in which Y is held to her promises by the threat that all the X's in the world will punish her for a while (at least) if she mistreats any one of them. The standard models of this phenomena are where Y deals sequentially with the X's, but models of simultaneous exchange also exist and are not very much more complex.

Of course, the key to these constructions is that X_{1023} will refrain from trade with Y if Y abuses X_{321}. If Y suspects that other X's won't pay much mind to Y's dealings with X_{321}, Y loses the incentive to honor her promise. And X_{321}, fearing this thought will enter Y's mind, will not find Y's promise credible.

As in the previous stories, the key is Y's (and the Xs') expectations about the actions of other X's. Y may not fear the wrath of X_{202}, but if it extends to the wrath of $X_{203}, X_{204},$ and X_{205} Y might be more mindful of the consequences of stiffing X_{202}.

Now imagine two cases: First, X_{203} hears Y promise X_{202} to give a talk and, on the appointed day, Y fails to show up. Second, Y promises X_{202} she will give this talk, with a payment of $150 if she fails to do so, and then on the appointed day, both Y and the check were missing. Our expectations (which, if they match the expectations of X_{202} are what matters here, together with Y's expectations of X_{202}'s expectations) are that in the first case, X_{203} will think Y is a jerk and treat her accordingly. In the second case, however, X_{203} might consider that X_{202} entered into a commercial transaction with Y, and it is up to X_{202} to collect his damages; no concern of X_{203}. Of course the problem is that X_{202}'s ability to collect his $150 depends either on court enforcement (which is costly, and which requires detailed negotiation of a contract), or on X_{203}, etc. threatening Y that they will treat her as a jerk if the check doesn't arrive.

The point is that insofar as a compensatory-damages clause affects how third parties react to the contract (a good-faith handshake promise between X_{202} and Y should be kept, but a commercial contract between them is none of X_{203}'s business), the good-faith handshake may be backed with greater social force than the contract. In theory, there is no reason why expectations/behavior should be this way. But we believe that they are.

(IV.9) Finally, X's welfare to some extent may enter Y's utility function. If Y doesn't show up, X will be adversely affected, and to some extent, Y may internalize this. Y may do so because she and X have dealt with each other over a period of years without the need for formal negotiations, damage clauses, and the like. At times Y may have to disappoint X (this may be one), and X will have withdrawn somewhat as a result. But their bond — *which is reflected in Y's utility function* — may be stronger as a result.

Again, there is a strong element of assuming the answer here, at least as far as "economic" reasoning goes. We would be happier deriving this than assuming it. But it seems to us likely to be true, and insofar as it provides a reason for the sort of promise-based constructions of earlier sections, it gives us somewhat more motivation for studying those functional equations.

(IV.10) *Noncalculative trust?* To finish up, a word about Williamson's recent work on calculative vs. noncalculative bases for trust is in order. The latter two and one-half stories have the ring of his noncalculative trust, but we imagine he would say that they fall short, since people are still doing calculations. Taking a descriptive view of rational choice would seem to square the two accounts to some extent; if X and Y don't really sit down and do their sums unless and until they get into the business of computed liquidated damages and the like — but because their actions are nearly rational, they act nearly as if they did — then we have a bit of a basis for the stories; those who don't do explicit calculations but act as if they did are "kinder" to each other, either on the basis of what seems like more generous expectations or what seems like a greater concern for the welfare of others.