Promises, promises, ... *

Juan D. Carrillo
University of Southern California and CEPR

Mathias Dewatripont
ECARES, Université Libre de Bruxelles and CEPR

Revised version: March 2008

Abstract

We consider a hyperbolic discounting individual who has the ability to make promises which are costly to break. We first identify conditions under which promises made are kept, and conditions under which they are (partially) broken. Second, we provide micro-economic foundations for the effectiveness of contractual promises. Specifically, we show how the cost of breaking promises can be reinterpreted in terms of either a reputation loss in the presence of incomplete information or a financial loss under monitoring and explicit contracting. The results imply that strategic interactions between hyperbolic discounting individuals may serve as a commitment mechanism against intrapersonal conflicts.

Keywords: Hyperbolic Discounting, Promises, Commitment, Reputation.

JEL Classification codes: C72, D86, L14.

*The authors are grateful to Isabelle Brocas, an editor, two anonymous referees and participants at the Harvard/MIT theory seminar, Toulouse, ULB, Lisbon and Columbia for comments. Correspondence address: Juan D. Carrillo, Department of Economics, University of Southern California, 3620 S. Vermont Ave., Los Angeles, CA 90089, USA (e-mail: juandc@usc.edu), or Mathias Dewatripont, ECARES, Université Libre de Bruxelles, C.P.114, 50 av. F. Roosevelt, 1050 - Brussels, Belgium (e-mail: mdewat@ulb.ac.be).

1 Introduction

In a seminal paper on time-inconsistent preferences, Strotz (1956) analyzes the behaviour of an individual who overemphasizes instant gratification relative to distant payoffs. Under these preferences, the individual has a tendency to underprovide effort in unpleasant tasks with delayed rewards. The problem is especially appealing given that experiments conducted first by psychologists and more recently by economists suggest that individuals often exhibit this "salience for the present".[1]

The problem of time-inconsistent preferences has generated a fair amount of interest in recent years. Two issues that have received attention are the use of *commitment devices* to alleviate intrapersonal problems and the effect of *strategic interactions* on the behaviour of hyperbolic discounting agents. The literature has explored different commitment technologies that may help the individual achieve his current goals. Caillaud et al. (1996) study the strategy followed by a time-inconsistent individual who can "self-restrain" his future choices. The paper proposes a new equilibrium concept, where the set of deviations is restricted to those strategies in which the individual will not have a further incentive to deviate. In the unique equilibrium of this one-person game, the individual optimally succeeds at each period in moderating his consumption. Laibson (1997) shows that investments in illiquid assets can prevent individuals from incurring inefficiently high levels of consumption. This provides a rationale for the existence of Christmas Clubs and other assets characterized by both high illiquidity and low rates of return. In Carrillo and Mariotti (2000), self-commitment is achieved through strategic ignorance. The paper shows that a researcher with meager but encouraging information on the prospects of a difficult project may optimally stay away from costless information and undertake it. Indeed, extra knowledge may cast some doubts about the quality of the

[1] See Ainslie (1992) and Loewenstein and Prelec (1992) for empirical and theoretical comparisons of exponential vs. hyperbolic discounting, respectively. See also Frederick et al. (2002) for a review of empirical estimates of discount rates and Caillaud and Jullien (2000) for a review of different ways to model time-inconsistencies.

project and, because of his time-varying preferences, lead the agent to beliefs involving inefficient procrastination.

Strategic interactions have also been studied from different angles. Brocas and Carrillo (2001) show that direct competition alleviates the tendency of hyperbolic discounting agents to rush into pleasant but unreasonable activities and to procrastinate in valuable activities whose benefits are delayed. Battaglini et al. (2005) study indirect (informational) spillovers. The paper shows that individuals with enough confidence on their willpower deal better with their own impulses if they can observe and learn from the behaviour of their peers.[2] In both cases, strategic interactions may be welfare enhancing. Time-inconsistent individuals can also be linked through markets. For example, Nocke and Peitz (2003) argue that secondary markets for durable goods can be detrimental for hyperbolic discounting consumers: they may induce individuals to engage in an inefficient behaviour where goods are purchased, enjoyed only one period and resold for a fraction of their value. Brocas and Carrillo (2004) show that the strategic ignorance mechanism developed by Carrillo and Mariotti (2000) can also have market consequences. If cash constrained entrepreneurs (rationally) avoid learning the value of their project before applying for a loan, there will be no self-selection among applicants. As a result, the market interest rate for loans will increase, making each potential entrepreneur worse-off.

The present research combines *commitment devices* with *strategic interactions*. It focuses on *promises towards third-parties*, another quite natural commitment technology against individual time-inconsistency that requires interpersonal relations.[3] It is indeed intuitively plausible to try and get around one's time-inconsistency by "making promises" (to one's family, friends, colleagues, etc.) to work hard, to be on time, to start a diet or

[2] O'Donoghue and Rabin (1999) study interactions between a time-consistent principal and a time-inconsistent, boundedly rational agent. The paper shows the optimality of a deadline-type contract.

[3] The approach to the issue of promises is very different from the one in Holmström and Kreps (1995). They focus on time-consistent individuals and assume away costs of breaking promises. In their setup, a promise is "cheap talk". It can be useful as a way to transmit information about a player's type. We also refer to Ellingsen and Johannesson (2004) for experimental evidence on the value of cheap-talk promises and threats.

to quit smoking. Such promises can alleviate self-control problems only if not fulfilling them results in some loss. This loss can be random, through a probability of being caught shirking, or deterministic, if failing to meet the promise means finishing the job late. It can represent a financial loss, if there is a penalty for shirking, or a reputation loss, if a late job affects the individual's future reliability. In this context, our goal is both to study the optimal use of promises (section 2) and to provide a microeconomic foundation for the effectiveness of such contractual promises (section 3). In the conclusion (section 4), we argue that the above results remain valid when we replace time-inconsistent preferences by limits to contracting as the source of the commitment problem of the individual.

Our main results are summarized as follows. First of all, we investigate in section 2 when promises will be made and which form they will take. We are interested in the extent to which promises are kept, and show that the answer to this question depends on the functional form of the detection probability or reputation loss of shirking (from now on we will simply refer to this function as the cost of breaking a promise). When the marginal cost is increasing in the size of the failure to meet the promise, then an equilibrium effort slightly lower than the one promised is relatively inexpensive, while big departures are the expensive ones. In this case, the individual makes promises he knows in advance he will not keep but that, at least, will force him to increase the effort relative to his future desired level. By contrast, if the marginal cost is decreasing in the size of the failure to meet the promise, individuals only announce promises that will be kept.

In a second step, in section 3, we provide two microeconomic foundations for the effectiveness of promises. After all, in a rational expectations world, any deviation from the promise should be perfectly anticipated by every individual. This could make the promise non-credible, and therefore useless as a tool for influencing future conduct.

In section 3.1, we show that it is possible to get around this argument if there is incomplete information about the cost of exerting effort. Think of the hyperbolic discounting individual as a seller who, through his effort level, chooses the quality of an input that

has to be later used by a buyer. The seller can early on make a promise concerning this quality, but this promise is cheap talk, and the buyer knows it. When it is time for the seller to exert effort, the buyer simultaneously has to choose a technology which will be the best "fit" for the quality of the input that will be provided by the seller. We assume that, ceteris paribus, the seller incurs an ex-post loss whenever the buyer makes a technology choice that "counts" on a higher quality input than what the seller has decided to provide. In the absence of incomplete information about the seller's cost, the two individuals play a Nash equilibrium in effort and quality, and it can easily be shown that the earlier promise does not change the outcome. Time-inconsistency means that the seller would like to commit to higher effort than what he will in the end exert. However, the buyer will understand the seller's incentives to exert effort and will appropriately "scale down" the technology choice. Promises are thus ineffective. In the presence of private information about the seller's effort cost, the situation is different because the buyer does not know a priori which technology is appropriate. If the seller is time-inconsistent, a pooling equilibrium may exist whereby high-cost sellers can benefit from pooling in their promise with low-cost sellers. Indeed, this can induce the buyer to make a more "ambitious" technology choice. The promise then serves as a commitment device for high-cost sellers to exert more effort and produce higher quality, even if in equilibrium they may choose to fall short of the promised quality and pay a certain cost for that.

In section 3.2, we offer another microfoundation for the effectiveness of promises. This one is based on explicit contracting. The problem we consider is the following. A hyperbolic discounting agent (the producer) has to exert effort at a future date and asks another hyperbolic discounting agent (the monitor) to check this effort. Both agents agree on a transfer payment from the producer to the monitor whenever the former has been caught shirking. A good monitoring scheme is one where the contractual probability of being monitored is high and is also credible, i.e., renegotiation-proof. We assume that both monitoring and renegotiation require a physical meeting between the parties and that,

due to limited time availability, meetings have to be arranged in advance. This gives the individuals an ex-ante commitment power on the frequency of meetings. However, when they do get together, they cannot commit not to renegotiate the prescribed effort level. This assumption formalizes the idea that interpersonal renegotiation is more difficult than intrapersonal renegotiation: interpersonal renegotiation requires coordinating a meeting, which is time-consuming especially since it means freeing up time collectively. In our model, the optimal frequency of meetings is the result of the following two effects. First, lowering this frequency is costly because it reduces the opportunities of monitoring. Deciding never to meet provides full commitment against renegotiation but also destroys all the discipline provided by monitoring. Second, meeting all the time also destroys monitoring, this time through systematic renegotiation of the prescribed current effort level. In our setup, the optimal frequency of meetings is well-defined and it allows individuals to partially get around their time-inconsistency.

An interesting general implication of our study of promises is that interpersonal relations partially help solving individual incentive problems. While under time-consistent preferences interpersonal relations are at best neutral in terms of incentives and typically a source of problems, in our framework collective relations can improve individual outcomes.[4] Thus, our work complements Brocas and Carrillo (2001) and Battaglini et al. (2005) in that it provides another positive role for social interactions in a world of hyperbolic discounting individuals.

2 Time-inconsistent preferences and promises

We analyze the behaviour of an individual with time-inconsistent preferences, in the sense of Strotz (1956). This short-run impatience implies that current payoffs are *overweighed* (or *salient* in the words of Akerlof (1991)) relative to future payoffs. Using the quasi-

[4]Naturally, we must be careful with what is considered an "improvement": a promise *increases* intertemporal welfare from the perspective of the individual who makes it, but it is *detrimental* for the future incarnation who is constrained to behave suboptimally from his own viewpoint.

hyperbolic notation introduced by Phelps and Pollak (1968), we posit that, from the perspective of the individual at date t, period $t+s$ ($s \geq 1$) is discounted at a rate $\beta\delta^s$ where $\beta \leq 1$. Naturally, $\beta = 1$ is the standard case of exponential discounting, and therefore time-consistent preferences. Without loss of generality, we will assume that $\delta = 1$, and we will call self-t the incarnation of the agent at date t. Furthermore, we will also assume that the agent is "sophisticated" (that is, aware at every period of his self-control problem).

At date 1, the individual is required to put some effort e in order to complete a task. The cost of this effort is immediate and equal to $\psi(e)$, where $\psi'(e) > 0$ and $\psi''(e) > 0$. The benefit comes one period later, at date 2, and has a value e (output is deterministic and equal to effort). Given the individual discounting, the surplus of self-0 and self-1 are respectively given by:

$$\beta\Big[e - \psi(e)\Big] \qquad \text{and} \qquad \beta\, e - \psi(e). \tag{1}$$

As one can easily see, the optimal effort that self-0 would like to exert at date 1 is strictly higher than the level of effort that self-1 is effectively willing to exert when date 1 arrives. Short-run impatience (captured with the parameter β) is the source of an intrapersonal incentive problem.

Assume now that self-0 can "announce" an effort e^* to be exerted at date 1, and that failure to reach this effort level leads to a cost $f(e^* - e)$ for any effort e exerted at date 1. This effort level e^* can be thought of as a "resolution" made to oneself or as a "promise" made to other individuals (boss, friend, corporate partner, etc.). In a single agent context, the cost of breaking the resolution can be for example a decrease in self-reputation over one's willpower (Bénabou and Tirole, 2004). In this paper, however, we will focus on the multi-agent situation, where promises are made to other people. In this social context, we will provide two alternative interpretations of the cost $f(e^* - e)$ of breaking a promise: a reputation loss towards third parties from not being reliable, or a probability of being detected shirking by an external monitor.

It is natural to assume that making a promise is costless as long as it is fulfilled or exceeded. When promises are not met, the cost is positive and increasing in the difference between the promise and the effort realized. This is summarized as follows.

Assumption 1 *$f(e^* - e) \equiv 0$ for all $e \geq e^*$ and $f'(e^* - e) > 0$ for all $e < e^*$.*

Note that $f(\cdot)$ convex (respectively, concave) means that larger departures from the effort promised are relatively more (respectively, less) costly than smaller departures. We could also assume the function $f(\cdot)$ to be increasing in $|e^* - e|$, which is to say that exceeding the promise is also costly. This would leave most of our results unaffected. For the time being, we take the cost $f(\cdot)$ of breaking a promise as exogenously given. However, a main objective of the paper is to develop interpersonal games that provide foundations for this cost structure. This is developed in section 3. We will then be able to give an economic interpretation to the shape of this cost function which, for the time being, is rather abstract. As mentioned in the introduction, the interpretation will be based either on reputation loss vis-à-vis other agents or on a probability of detection by an external source. The timing of the game is summarized in Figure 1 (note that it does not matter when the cost of breaking a promise is paid so, for simplicity, we assume that it occurs at the date in which output is produced).

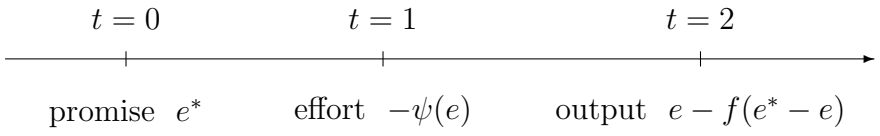


Fig. 1. Timing.

In our setting, self-0's intertemporal utility function is:

$$W(e, e^*) = \beta\Big[e - f(e^* - e) - \psi(e)\Big]. \tag{2}$$

Because of the dynamic inconsistency of preferences, when the date of exerting effort

arrives self-1's intertemporal utility function becomes:

$$V(e, e^*) = \beta\Big[e - f(e^* - e)\Big] - \psi(e). \tag{3}$$

From (2) and (3) we can notice that, absent a possibility of promises ($e^* = 0$), self-1 ends up exerting too little effort from self-0's viewpoint. Formally, denote $e_0(e^*) = \arg\max_e W(e, e^*)$ and $e_1(e^*) = \arg\max_e V(e, e^*)$. We get:

$$e_1(0) < e_0(0)$$

with $\psi'(e_1(0)) = \beta$ and $\psi'(e_0(0)) = 1$.

Assume the only instrument that individuals have at date 0 for "forcing themselves" to exert a high level of effort is a promise towards a third party. Naturally, this promise will be ex-post costly, whenever it is not fulfilled. Our first concern is to provide a full characterisation of the optimal promise e^* given the functional form $f(\cdot)$ for the cost of breaking it.

What we are considering is technically equivalent to a moral hazard problem, with self-0 setting the promise e^* as an incentive scheme for self-1. As is well-known, moral hazard problems are easily plagued by non-concavity of the overall maximand (see e.g., Grossman and Hart (1983)). In order to avoid that, we rely on the following technical assumption:

Assumption 2 $\beta\, f''(e^* - e) > -\psi''(e)$, $f'''(e^* - e) \leq 0$ *and* $\psi'''(e) \geq 0$ $\forall\, e^*$ *and* $e \leq e^*$.

The first part says that $f(\cdot)$ is never "too concave" relative to $-\psi(\cdot)$. It guarantees that $V(e, e^*)$ is concave in the effort e exerted by self-1. The second and third parts say that the rates of concavity of $f(\cdot)$ and $-\psi(\cdot)$ are non-decreasing in their arguments. They guarantee that $W(e_1(e^*), e^*)$ is concave in the promise e^* made by self-0, even when $f'' > 0$.

Denote by $\bar{e}$ the optimal level of effort exerted by self-1 *conditional* on the promise e^* being fulfilled (i.e., given $e^* = \bar{e}$). Naturally, this effort will depend on the marginal cost of a departure from the full promise $f'(0)$. Formally,

$$\left.\frac{\partial V(e,\bar{e})}{\partial e}\right|_{e=\bar{e}} = 0 \quad \Leftrightarrow \quad \psi'(\bar{e}) = \beta\Big[1 + f'(0)\Big].$$

Using this definition, we are in a position to state our first result.

Proposition 1 *Consider the game where self-0 makes a promise and self-1 exerts effort. If the intrapersonal conflict is sufficiently small or if a departure from the promise is sufficiently costly, then self-0's optimal effort level is exerted by self-1 and promises are fulfilled (case (i)). Otherwise, self-0's optimal effort level is not reached. In that case, promises will remain unfulfilled if the marginal cost is sufficiently increasing in the difference between promise and effort (case (iii)) and they will be fulfilled if it is not (case (ii)). Formally,*[5]

(i) *If* $f'(0) > \frac{1-\beta}{\beta}$, *then* $e^* = e_0(0)$ *and* $e_1(e^*) = e^*$.

(ii) *If* $f'(0) < \frac{1-\beta}{\beta}$ *and* $\beta f''(0) < f'(0)\frac{\psi''(\bar{e})}{1-\psi'(\bar{e})}$, *then* $e^* = \bar{e} < e_0(0)$ *and* $e_1(e^*) = e^*$.

(iii) *If* $f'(0) < \frac{1-\beta}{\beta}$ *and* $\beta f''(0) > f'(0)\frac{\psi''(\bar{e})}{1-\psi'(\bar{e})}$, *then* $\bar{e} < e_1(e^*) < e_0(0)$ *and* $e_1(e^*) < e^*$.

Proof. See Appendix A1. □

The idea of the proposition is the following. A promise makes sense only if it requires an effort higher than $e_1(0)$, self-1's preferred level, and it will never induce an effort higher than $e_0(0)$, the optimal effort from self-0's viewpoint. If the marginal cost of any deviation from the prescribed effort is sufficiently high (i.e., if $f'(0)$ is high enough so that, for example, even small amounts of shirking are detected and punished with high probability), then failing to meet the promise is too costly, and therefore self-0 only imposes targets that can be fulfilled. These targets may not necessarily imply that self-0's optimal effort level is reached. Indeed, when the intrapersonal conflict is too important,

[5] A sufficient condition for the second inequality in case (ii) to hold is $f'' < 0$, and a sufficient condition for case (iii) to hold is $f'(0) = 0$ and $f''(0) > 0$.

an excessively demanding promise does not act as a commitment device for higher future effort. It is then more interesting to set mild promises that are fully honored. Now, suppose that failing to meet the target "by a little" is not too costly (i.e. $f'(0)$ small) but this marginal cost increases with the difference between effort promised and effort realized ($f'' > 0$). In this case, by setting higher and higher targets, the individual is committing to exert more and more effort, even though these promises are never fulfilled. Targets are then raised by self-0 until the (constant) gains of a higher commitment to effort are offset by the (increasing) costs of unfulfilled promises. In equilibrium, self-1 loses part of his reputation or is detected with some probability.

Let us now illustrate this result with two examples.

Example 1: Unfulfilled promises. Consider the quadratic case, where $f(e^* - e) = (e^* - e)^2/2$ and $\psi(e) = e^2/2$. In this case, $e_0(0) = 1$ and $e_1(0) = \beta$. Note that $f'(0) = 0$ and $f'' > 0$, so we are in case (iii) of Proposition 1. Maximizing the payoff at date 1 given a promise $e^* \geq \beta$ yields:

$$e_1(e^*) = \frac{\beta}{1+\beta}(1 + e^*).$$

Note that $e_1(e^*) < e^*$ for all $e^* > \beta$ and $\frac{\partial e_1}{\partial e^*} = \frac{\beta}{1+\beta} \in (0, 1)$. In words, whenever the promise exceeds self-1's optimal effort $e_1(0)$, the individual chooses not to meet the promise to its full extent. At the same time, raising the promise still raises future effort. Consequently, the individual makes promises that remain unfulfilled. Specifically, the optimisation at date 0 yields:

$$e^* = \frac{2\beta}{1+\beta^2} \quad \text{and} \quad e_1(e^*) = \frac{\beta + \beta^2}{1+\beta^2}$$

so that we have $e_1(0) < e_1(e^*) < e^* < e_0(0)$.

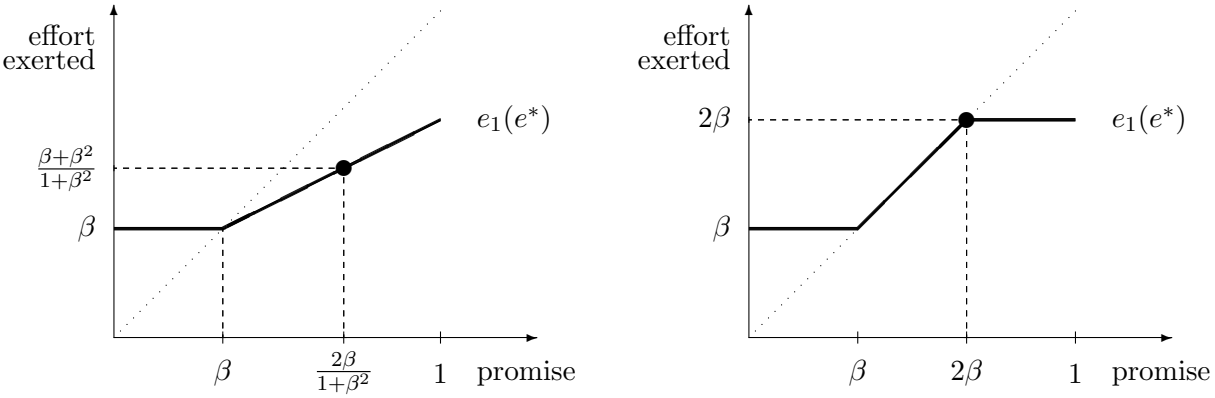
Example 2: Fulfilled promises. Consider now the case of linear promise $f(e^* - e) = (e^* - e)$ and quadratic cost of effort $\psi(e) = e^2/2$. We have $f'(0) = 1$ and $f''(0) = 0$ so, according to Proposition 1, we are either in case (i) or case (ii). We then get:

- If $\beta > 1/2$ (i.e. $f'(0) > (1-\beta)/\beta$), then $e_1(e^*) = e^* = e_0(0) = 1$.

• If $\beta < 1/2$ (i.e. $f'(0) < (1-\beta)/\beta$), then $e_1(e^*) = e^* = \bar{e} = 2\beta < e_0(0) = 1$.

The individual always fulfills his promises. Indeed, effort is set to match the promise until a certain level, and then stays constant. Setting the promise beyond this threshold thus makes no sense: it raises the probability of detection or reputation loss while failing to increase effort. As seen from the equation above, if the intrapersonal conflict is sufficiently weak ($\beta > 1/2$), first-best effort is achieved $e_1(e^*) = e_0(0) = 1$. By contrast, if the conflict is strong enough ($\beta < 1/2$), only a second-best effort can be obtained $e_1(e^*) < e_0(0)$, but the promise is still a useful commitment device $e_1(e^*) = 2\beta > e_1(0) = \beta$.

Examples 1 are 2 are graphically represented in Figure 2.



Example 1. Unfulfilled promises

Example 2. Fulfilled promises

Fig. 2. Some examples.

The results presented in Proposition 1 can be generalized in a number of directions. First, one might think of the promise as the commitment of self-0 to reach a certain target in terms of quantity produced or date of delivery. If such output is used as an input by another agent, meeting the target might be essential. This would amount to a highly concave (or even step) function for the cost of breaking promises. Still, there are circumstances in which the total amount of product or the delivery date may partly depend on factors beyond the control of the individual. Formally, effort may affect *stochastically* production or time of delivery. One can show that this uncertain world with rigid targets

can be formally translated into a cost of breaking the promise function which is first convex and then concave.[6] Naturally, Proposition 1 holds in this new framework. This means that, as long as there is some uncertainty, unfulfilled promises may occur even if targets are rigid.

Second, the above one-shot problem can also be extended to a dynamic setting. Specifically, suppose that, at the beginning of each date t, the individual inherits a stock e_t^* of past unfulfilled promises. He exerts an effort level e_t ($\leq e_t^*$) that may or may not account for all these promises, and makes a promise e_{t+1}^* ($\geq e_t^* - e_t$) for the next period. Under certain conditions about the cost of effort and the cost of breaking promises, it may be optimal for the individual to exert at each date a positive level of effort which, nevertheless, falls short of the promise. Overall, we may obtain the paradoxical result that the agent never fulfills his promises ($e_t < e_t^*$), and yet he still finds it optimal to always commit to higher future levels of effort ($e_{t+1}^* > e_t^* - e_t$).

3 Foundations for promises

In the previous section we have assumed the existence of an *exogenous cost* whenever a strategy announced by self-0 is "renegotiated" by self-1. We have then studied the optimal use of this tool in shaping future behaviour. In particular, we have shown that commitment through promises is a useful way to avoid procrastination. At the same time, it is a costly mechanism: excessively high targets which are not respected ex-post may (optimally) be announced in equilibrium.

In this section, we *endogenize the function* $f(\cdot)$ that captures the cost of breaking a promise. There is, indeed, a tension between perfect foresight by all players in the game and the cost borne by the agent who breaks a promise. After all, if individuals are rational, a deviation by self-1 from the strategy announced by self-0 will be perfectly anticipated by

[6] Intuitively, small departures are not very costly, since the risk to end up with insufficient production or to exceed the deadline is small. As the risk starts growing, the expected cost increases. Finally, the cost stabilizes once it become almost sure that the promise will never be kept.

third parties. This will make the promise non-credible, and therefore useless as a tool for influencing future conduct. To be more precise, suppose that a supplier announces that its product will be ready in two days as a costly commitment to have it ready in three days. A rational retailer should expect to receive it in three days (not in two), dissipating the cost of the announcement, and therefore making the commitment ineffective in the first place. In this example, a three-day delivery can only occur in equilibrium if the retailer is fooled by the announcement and expects a two-day delivery.

We argue, however, that the basic model can be extended in a way to circumvent this reasoning. By explicitly modelling the interaction of our hyperbolic discounting agent with the outside world, we can endogenize the cost of breaking promises. Sticking to the time-inconsistency paradigm, we offer two different foundations for the cost function $f(\cdot)$. In both cases, they lead to situations where individuals make promises anticipating that they will be broken with positive probability –or even with probability one– in equilibrium.[7]

We first study an incomplete information setting where the individual makes promises in order to entertain a reputation for exerting high effort. We show that this creates a credible *endogenous reputation cost* associated to the reneging on promises which usefully influences ex-post behaviour. The reputation cost comes from the fact that the outside world is not sure whether the promise will be fulfilled: the type of individual who plans to renege on his promise (the one whose cost of exerting effort is high) pools with another type of individual that will honor it (the one whose cost of exerting effort is low). Consequently, when a promise is broken ex-post, this was not fully anticipated ex-ante.[8]

Our second rationale for promises builds on interpersonal monitoring. We show that, by starting a contractual relation, an individual can ex-ante create an *endogenous financial*

[7]The paper still uses hyperbolic discounting as a premise. However, the key reason why there is a need for promises is the existence of a commitment problem. In the conclusion, we argue that the results would hold if we replaced time-inconsistent preferences with limits to contracting as the source of the commitment problem.

[8]Bénabou and Tirole (2004) derive a similar cost function. The paper focuses on a single agent with incomplete information about his own willpower and imperfect recall of past motivations. The cost is a loss in "self-reputation" incurred by an individual who succumbs to his vices.

cost associated to the reneging on promises. The cost is credible: the individual pays a penalty whenever a deviation from the contractual arrangement is detected. The cost is also efficient in affecting ex-post behaviour: because of the penalty, the individual has diminished (but still some) incentives to shirk.

Note that, for both types of foundations, it is possible to provide testable predictions about whether promises will be kept or broken in equilibrium. For this, we need to be able to quantify (i) the decrease in the joint surplus as a function of the difference between expected effort and realized effort (for the reputation case) or (ii) the probability of detection as a function of the difference between effort exerted and effort contracted (for the contracting case).

3.1 Reputation cost of breaking a promise

Consider the following stylized buyer/seller adaptation of our basic model.[9] A hyperbolic discounting seller (he) can, at date $t = 0$, promise to deliver a good at $t = 2$. To produce the good by that date, he has to exert some effort e at date $t = 1$ with an immediate cost $\psi(e)$. This effort can be thought of as determining the *quality* of the good, with higher effort implying higher quality.

At $t = 0$, the seller can promise an effort/quality level that we shall call e^p (and not e^*, for reasons that will become clear shortly). This cheap-talk promise is observed by the buyer (she), but the effort e actually exerted is not. The buyer must also take an action. More precisely, she must choose at $t = 1$ the technology that will be used to transform the good purchased from the seller into a final product. Since the buyer is only active in two consecutive periods (1 and 2), it is irrelevant whether her preferences are time-inconsistent or not. Let us call the action of the buyer e^*. One can think of this

[9] Technically speaking, the model in this section is not a "classical" reputation model (in game-theoretic parlance), but instead a cheap-talk model whereby a promise can signal private information. We use the term reputation nonetheless because game-theoretic reputation models work de facto through a signaling mechanism. It would be interesting to extend our setup to multiple periods, in which case repeated promises would result in a changing reputation of the individual.

choice as the buyer trying to adopt the technology that is "most compatible" with the quality of the good produced by the seller. For simplicity, there is no cost associated to the selection of a specific technology. However, the total ex-post surplus of the trade will depend on both the buyer's technology and the seller's product quality. Formally, it is given by:

$$h(e, e^* - e).$$

where the *total* derivative of $h(\cdot)$ is increasing in e (a higher quality input is always valuable for the ex-post surplus) and, at the same time, its partial derivative is decreasing in $|e^* - e|$ (the best technology is the one which fits most closely with the quality of the input).

We do not explicitly model how this surplus is split between buyer and seller. Instead, we suppose that the seller's date-2 benefit of production is, independently of his (cheap-talk) promise e^p, equal to $e - f(e^* - e)$ if $e < e^*$ and to e if $e \geq e^*$. Therefore, the buyer's ex-post surplus is given by:

$$g(e, e^* - e) = \begin{cases} h(e, e^* - e) - \Big(e - f(e^* - e)\Big) & \text{if } \ e < e^* \\ h(e, e^* - e) - e & \text{if } \ e \geq e^* \end{cases}$$

where, as for $h(\cdot)$, we assume that the total derivative of $g(\cdot)$ is increasing in e and its partial derivative is decreasing in $|e^* - e|$.

The above assumptions are meant to match exactly the formalisation of section 2. They are also natural: the seller has a payoff which increases in the quality of the input he produces, but he loses some of the surplus if this quality is inferior to the one the buyer "has counted on". In other words, a difference between e^* and e will thus involve a loss for the buyer as well as for the seller. It is important to notice that these losses do not come directly from a difference between the seller's promise e^p and the seller's effort e, but from a difference between the buyer's choice of technology e^* and the seller's effort e. It is only when the buyer is unable to infer the effort of the seller that a loss will be incurred. The timing is summarized in Figure 3.

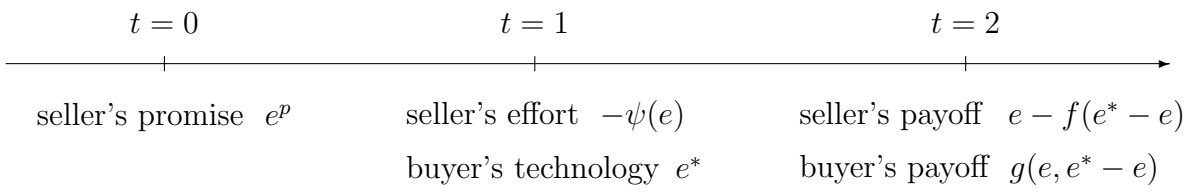


Fig. 3. Timing of the reputation game.

Assume first that payoffs are common knowledge, and that the seller is time-consistent ($\beta = 1$). In this case, things are simple. The two parties play a Nash equilibrium at $t = 1$: the buyer optimally selects the technology that coincides with the anticipated effort of the seller ($e^* = e$), and the seller chooses his optimal effort level $e = e_0(0)$, where $\psi'(e_0(0)) = 1$. Promises at date 0 then play no role.

Consider now a second case, where payoffs are still common knowledge but the seller is time-inconsistent ($\beta < 1$). Things are also simple. In any pure-strategy equilibrium, the buyer will again select the technology that coincides with the anticipated effort of the seller at date 1 ($e^* = e$), and the seller will choose an effort level $e = e_1(0)$, where $\psi'(e_1(0)) = \beta$. In this case, as of date 0, the seller would like to commit his future self to a higher effort level, but he is unable to: whatever the date-0 promise e^p, the buyer will correctly understand the incentives of the seller when he takes his decision at date 1, so the promise will have no effect whatsoever.

We now introduce private information. Assume that the seller's cost of effort is:

$$\gamma\psi(e),$$

where $\gamma \in \{\gamma_L, \gamma_H\}$ and $\gamma_H > \gamma_L > 0$. The seller privately knows his value of γ while the buyer has only a prior $p \equiv \Pr(\gamma = \gamma_H)$. If the seller is time-consistent ($\beta = 1$), the date-0 promise can serve as a costless separating device. Call type i (with $i \in \{L, H\}$) the seller whose cost of effort is $\gamma_i \cdot \psi(e)$ and denote by $\hat{e}_i$ his optimal effort, that is, the one that

satisfies:

$$\gamma_i \psi'(\hat{e}_i) = 1. \tag{4}$$

It is in the interest of this seller to promise an effort level $e_i^p = \hat{e}_i$. Besides, since type i seller can be counted on to keep his promise ($e_i = e_i^p = \hat{e}_i$), the buyer will choose also a technology $e^* = \hat{e}_i$. Overall, promises *inform* the buyer about the seller's cost of effort, and therefore they are useful *separating devices*. However, the role of promises here is different from the one in section 2, since they do not serve as *commitment devices* for future decisions.[10]

The most interesting situation arises when we simultaneously have private information about the cost of effort and time-inconsistency. Denote by $\tilde{e}_i$ the optimal effort of type i agent from his self-1 perspective. Formally, this effort satisfies:

$$\gamma_i \psi'(\tilde{e}_i) = \beta. \tag{5}$$

Recall that the optimal effort from his self-0 viewpoint corresponds to $\hat{e}_i$ ($> \tilde{e}_i$). Therefore, as in section 2, the seller at date 0 would like to commit his self-1 incarnation to a higher effort level ($e_i^p > \tilde{e}_i$). While there is nothing type L can hope to achieve, there may be room for type H to pool with type L in order to induce a higher technology choice e^* from the buyer than in the separating case. This will induce self-1 of type H to work harder, but it may carry the cost of a broken promise. Considering these two motives, we obtain the following result.

Proposition 2 *Under private information and time-inconsistency, there are two types of equilibria with promises:*

(i) A separating equilibrium (promises as information devices), where type i's promise is $e_i^p = \tilde{e}_i$, his effort exerted is $e_i = \tilde{e}_i$, and the buyer chooses technology $e^* = \tilde{e}_i$.

[10] In fact, promises are cheap talk and serve only as separating devices. It is therefore irrelevant which promises are announced as long as each type of seller announces a different promise and the buyer is able to know which promise corresponds to each type.

(ii) A pooling equilibrium (promises as commitment devices), where both types of sellers announce a promise $e_i^p = \tilde{e}_L$ and type L exerts effort $e_L = \tilde{e}_L$. Type H may either fulfill his promise ($e_H = \tilde{e}_L$) in which case the buyer also chooses $e^* = \tilde{e}_L$, or fail short of it ($e_H \in (\tilde{e}_H, \tilde{e}_L)$), in which case the buyer chooses $e^* \in (e_H, \tilde{e}_L)$.

A sufficient condition for pooling being optimal is $\beta\,\gamma_H < \gamma_L$.

Proof. See Appendix A2. □

The idea of an equilibrium with promises as a commitment device is the following. By pooling on the promise, the high-cost seller prevents the buyer from learning which type of individual will deliver the good at date 2. If the optimal effort of the type L seller from his self-1 viewpoint is sufficiently close to the optimal effort of the type H seller from his self-0 viewpoint, then the type H seller will mimic the type L seller. This comes at no cost since the technology chosen by the buyer will correspond exactly to this effort level. More interestingly, the type H seller can also decide to underprovide effort compared to his low-cost peer. The buyer anticipates the departure but she is unsure about which seller is going to deliver the good, since both have announced the same target. She then chooses a technology strictly in-between the two levels, with a corresponding cost for the agent who underperforms, but no cost for the one who overperforms.

As p increases, the buyer is more confident that she will face a high-cost seller. Then, the technology she chooses if she anticipates different efforts becomes closer to e_H. This reduces the cost of breaking promises but, at the same time, it also decreases the value of the promise as a commitment device (in the extreme case $p = 1$, we are back to the situation with full information in which promises are useless). Note also that pooling can be excessively costly only when it induces the type H seller to exert effort above his self-0 first-best level ($\tilde{e}_L > \hat{e}_H$). This is never the case if the intrapersonal conflict is sufficiently important (β low) and the two costs (γ_L and γ_H) are sufficiently close to one another.

Proposition 2 thus provides foundations for the results of section 2, showing that

time-inconsistent individuals will make promises in order to commit future selves to alter their effort level. An equilibrium with promises is sustainable even though every agent understands that future selves will (partially) renege on these promises if the cost of doing so is not too high. The cost of breaking promises here comes from one's *reputation*: just as in reputation models, the promise is made in order to keep the buyer uncertain about the seller's type, i.e., to keep her believing that the seller might have a low cost of effort. This leads her to revise upwards her technological decision, which is what even self-0 of the high-cost seller may prefer, despite the fact that a cost of breaking the promise may be incurred later on. Naturally, the primitives of the asymmetric information model (probability of being a high cost type and cost differential between types) as well as the production function that combines the effort of the seller with the technology of the buyer will determine whether the same promises are made by both types and, if so, whether they are fulfilled in equilibrium.

Note that, in our model, only the high-cost seller may choose to perform below the announced level, and therefore bear the cost of breaking a promise. This is mainly due to the simple informational and technological structures adopted in the paper. For example, with more than two types, all but the lowest cost seller may decide to break their promise. More interestingly, even with only two types, we may obtain deviations by both sellers if we assume that exceeding the quality target is also costly for the seller ($f(e^* - e) > 0$ for $e^* < e$).

We finally illustrate our result with the following example.

Example 3: pooling and separating reputation equilibria. Consider the case where $f(e^* - e) = (e^* - e)$, $\psi(e) = e^2/2$, $g(e, e^* - e) = e - (e^* - e)^2/2$ and $e \in [0, 1]$. Besides, $\gamma_L = 1$ and $\gamma_H = 1/\alpha$ with $\alpha < 1$.

From (4) and (5), we have: $\hat{e}_H = \alpha$, $\hat{e}_L = 1$, $\tilde{e}_H = \alpha\beta$ and $\tilde{e}_L = \beta$. If a separating

equilibrium exists, the welfare of type H from his self-0 perspective is:

$$\beta\Big[\tilde{e}_H - \gamma_H\left(\tilde{e}_H\right)^2/2\Big] = \alpha\beta^2\left(1-\beta/2\right).$$

A sufficient condition for pooling being optimal is then $\tilde{e}_L < \hat{e}_H \Rightarrow \beta < \alpha$.

If a pooling equilibrium exists then, conditional on the effort e_H exerted by type H, the optimal technology adopted by the buyer solves:

$$\max_{e^*} \quad p\left[e_H - (e^* - e_H)^2/2\right] + (1-p)\Big[\beta - (e^* - \beta)^2/2\Big].$$

Therefore, for all $e_H \leq \beta$, $e^* = p\, e_H + (1-p)\beta$ and $(e^* - e_H) = (1-p)(\beta - e_H)$. In a pooling equilibrium, the optimal date-1 effort of type H (if $e_H < \beta$) solves:

$$\max_{e_H} \quad \beta[e_H - (e^* - e_H)] - e_H^2/2\alpha.$$

Therefore, $e_H = \alpha\beta(2-p) \in (\tilde{e}_H, \tilde{e}_L)$ if $\alpha(2-p) < 1$ (unfulfilled promises) and $e_H = \beta = \tilde{e}_L$ if $\alpha(2-p) \geq 1$ (fulfilled promises). Last, suppose that $\alpha(2-p) \geq 1$. In a pooling equilibrium with fulfilled promises, the welfare of a type H seller from his self-0 viewpoint is:

$$\beta\Big[\tilde{e}_L - \gamma_H\left(\tilde{e}_L\right)^2/2\Big] = \beta^2\left(1-\beta/2\alpha\right).$$

Hence, the necessary and sufficient condition for pooling being optimal is $\beta^2\left(1-\beta/2\alpha\right) > \alpha\beta^2\left(1-\beta/2\right) \Rightarrow \beta < 2\alpha/(1+\alpha)$, which is a weaker condition than $\beta < \alpha$.

3.2 Financial cost of breaking a promise

Consider now an alternative extension of the model presented in section 2. A hyperbolic discounting individual (the producer, P) will be required to manufacture a good once between dates 1 and n. Production requires an immediate cost $\psi(e)$ and provides a one-period delayed benefit e. Given his time-inconsistent preferences, in the absence of a commitment technology, the individual at the date of producing $t \in \{1, ..., n\}$ will underprovide effort relative to the optimal level from his perspective at date 0 ($e_1(0)$

rather than $e_0(0)$). In particular, since $\delta = 1$, the surplus of production for self-0 is the same independently of the date of production.

Suppose now that, at date 0, P may enter a contractual relation with another individual (the monitor, M) who has the same dynamically inconsistent rate of time-preferences. The contract between P and M can specify an effort e^* to be exerted by P whenever production is necessary as well as the dates at which P and M meet. Upon such meetings, M costlessly checks the levels of previous efforts by P provided these were exerted no more than y (≥ 1) dates before. Formally, we assume that if P exerts an effort $e < e^*$ at date t, then the probability that M detects P shirking in period $t + \tau$ is:[11]

$$\begin{cases} q(e^* - e) & \text{if} \quad \tau \leq y \\ 0 & \text{if} \quad \tau \geq y + 1. \end{cases}$$

The time at which effort has to be exerted is unknown at date 0 (for simplicity, it is uniformly distributed between 1 and n). It is only learned by P at the beginning of the period in which production takes place and, more specifically, before meeting with M (if a meeting is scheduled for that period). Therefore, upon a meeting, parties can use this opportunity not only to check past effort but also to *renegotiate* prescribed effort levels for that date or for any future date.

Given that both individuals are time-inconsistent, if effort is required at date t, the *joint surplus* of P and M (including any frictionless interpersonal transfer) from their self-0 and self-t perspective are respectively:

$$\beta\Big[e - \psi(e)\Big] \qquad \text{and} \qquad \beta\, e - \psi(e), \tag{6}$$

which is exactly the same as in (1). It is important to notice that an interpersonal contractual relation is no miracle cure to an intrapersonal incentive problem: given that both individuals have identical salience for the present and that interpersonal transfers

[11] Instead of a constant probability of detection that drops to zero after some periods, one could assume a function q that smoothly decreases in τ. The results would not change significantly under this alternative formalisation.

are costless, renegotiation of the effort to be exerted takes exactly the same form as in the exogenous cost of breaking promise case studied in section 2. There is still a crucial aspect of the monitoring game that was not captured in the basic model: for interpersonal relations to happen, some degree of *coordination* is necessary. This difference is summarized in the following assumption.

Assumption 3 *Interpersonal meetings cannot be organized on the spot, but have to be arranged in advance.*

In an intrapersonal game, "meetings with oneself" are not only possible at any moment but even unavoidable. Therefore, self-checking of previous effort and self-renegotiation of current effort can and will be conducted at every date.[12] By contrast, when we carefully model promises to (or, equivalently, monitoring by) a third party, it is not unreasonable to assume that coordination problems combined with limited time availability reduce the ability of individuals to schedule meetings on the spot. In that spirit, note for example that in the literature on monitoring in hierarchies (Calvo and Wellisz, 1978) having to monitor potentially many individuals reduces one's ability to monitor each one of them. In the same vein, Aghion and Tirole (1997) argue that having many agents can serve as a commitment device for a principal who wants to commit not to be interventionist.

In this setting, a date-0 contract between the producer and the monitor consists of three elements. First, a prescribed effort level e^* for the producer. Second, a transfer C from the producer to the monitor for being caught exerting less effort than the prescribed level. This penalty will depend on the agents' attitude towards risk. In the case of risk-neutrality and limited liability, individuals will agree to set the highest possible penalty whenever the first-best cannot be achieved. Third, a set of dates at which P and M meet to check previous effort levels. Given assumption 3, meetings have both costs and

[12]The recent neuroscience literature suggests that the brain is divided into systems with conflicting objectives. If we adopt this approach, one could defend assumption 3 even in an intrapersonal game. We will limit our attention to a more traditional view, in which the individual does not have any conflict of goals *at a given date*.

benefits from the individuals' self-0 perspective. On the one hand, the expectation of a future meeting keeps P on his toes, because he fears being caught at that time and therefore having to pay the penalty. On the other hand, a meeting can also be used to renegotiate away previously set effort levels. Given these considerations, we have the following result.

Proposition 3 *For any probability of detection $q(e^* - e)$, interpersonal interactions alleviate intrapersonal incentive problems. If we restrict attention to deterministic meeting patterns, these should optimally be organized every $y+1$ dates.*

Proof. See Appendix A3. □

The intuition behind this result is the following. A meeting date is a period "lost" in terms of commitment not to shirk: any effort level in excess of $e_1(0)$ to be exerted *at that date* will be renegotiated away. From this point of view, meetings should be organized as infrequently as possible. On the other hand, prescribing high effort levels can have an impact in the producer's decision only if he expects to be controlled within y periods. From this point of view, meetings should be organized so as to minimize the opportunity for P to "get away" with low effort. This means that, for each period without a meeting, the next meeting should take place no more than y periods later, thereby leading to our result. Under this optimal frequency of communication and with a uniform ex-ante distribution for the period at which effort has to be expended, ex-post control (and therefore, high effort) occurs with probability $y/(y+1)$, whereas renegotiation (and therefore, low effort) occurs with probability $1/(y+1)$. The intertemporal joint welfare from the agents' perspective at the date of production t is therefore:

$$V^t = \frac{y}{y+1}\Big(\beta\Big[e_1(e^*) - q(e^* - e_1(e^*))\,C\Big] - \psi(e_1(e^*))\Big) + \frac{1}{y+1}\Big(\beta\, e_1(0) - \psi(e_1(0))\Big).$$

Compared to section 2, e^* is the result of the same maximisation problem $\mathcal{P}$ by self-0 and $e_1(0)$ is the same optimal effort of self-1 in the absence of promises. Finally, $e_1(e^*)$ is the

same optimal effort exerted by self-1 conditional on the promise e^* made by self-0, except that the cost function $f(e^* - e)$ is now replaced by $q(e^* - e)\,C$.

To sum up, the key assumption of our analysis is that *interpersonal* relations (monitoring and renegotiation) require some amount of coordination. Consequently, when an individual learns that he has to exert effort and that no meeting is scheduled, he is happy to have this commitment device. Besides, at that time, it is not possible to renegotiate the commitment away anymore. Therefore, the promise imposes some discipline on the producer through the probability of being caught shirking. Contracting with another agent is necessary because the constraint in organizing a meeting is not present in the case of an intrapersonal game: "self-meeting" (and therefore "self-renegotiation") is possible any time. Note also that avoiding renegotiation does not necessarily imply that all the effort specified in the contract is exerted: as shown in Proposition 1, self-0 may commit to a certain effort anticipating that this level will never be attained ($e_1(e^*) < e^*$). In that case, detection and interpersonal transfers occur in equilibrium with strictly positive probability. The primitives of the monitoring function (here, the probability of detection $q(\cdot)$ and the maximal punishment C) are crucial in determining which type of promises are optimal. In other words, if we are able to determine the detection technology and the extent of the punishment, it is then possible to have testable predictions on whether individuals announce promises that will be fulfilled or not.

We have not considered costly side-transfers and/or contracts between agents with different rates of time-preferences because we wanted the joint surplus to be as close as possible to the case discussed in section 2 (see equations (1) and (6)). This facilitates renegotiation at the date in which effort has to be exerted, and therefore makes it the most difficult case for solving the intrapersonal problem. Furthermore, it allows us to sidestep peripheral issues (who makes the offer, what is the relative bargaining power of parties, what is the outside option of each player, etc.), which would only divert attention from the core question of the paper. The assumption of equal time-inconsistency of both

players seems adequate in social interactions, where the monitor is an individual who "helps" a colleague start a diet, quit smoking, keep his appointments, study for an exam, etc. By contrast, when we consider a pure moral hazard situation, the monitor is likely to be a time-consistent outside organization. Since renegotiation is harder when one of the parties does not exhibit a salience for the present, Proposition 3 provides, for that case, a lower bound on the ex-ante joint benefits of interpersonal contracts.

We conclude section 3 with a general remark. In standard models with time-consistent individuals, interpersonal relations are at best neutral in terms of incentive effects (e.g., under perfect contracting) and otherwise detrimental. Under time-inconsistency, it is possible to give a new, "positive" role to interpersonal relations between self-interested individuals: they can partially solve intrapersonal incentive problems. As mentioned earlier, this was first pointed out by Brocas and Carrillo (2001) in a full information framework with direct externalities and studied in more detail by Battaglini et al. (2005) in an incomplete information setting with indirect (informational) externalities. Our paper discusses another benefit of social interactions.

4 Conclusion

This paper has identified conditions under which promises, made by a time-inconsistent individual and which lead to a financial or reputation loss if broken, are fulfilled and conditions under which they are partially broken. Two different foundations for the cost of broken promises have been considered. First, a reputation loss in the presence of incomplete information. Second, an endogenous financial loss arising from interpersonal monitoring and explicit contracting. The results imply that strategic interactions by hyperbolic discounting individuals may serve as a commitment mechanism that alleviates intrapersonal conflicts.

Further exploring the generality of our results would be an interesting avenue for future research. We have mentioned the possibility of extending the model to a multiperiod

environment. In this more comprehensive setting, we could analyze the evolution of reputation and its effect on the desirability and effectiveness of promises. Multitask extensions would also be natural topics for further study. Which tasks would be chosen for extending promises? How would the possibility of making promises affect the portfolio of tasks pursued by individuals?

Finally, our model is relevant beyond hyperbolic discounting. Indeed, our results essentially rely on the fact that the individual starts with a *commitment problem*, something which can arise in the absence of time-inconsistent preferences: In strategic situations, tying one's hands in advance can be helpful as a way to influence the behaviour of others. The results of this paper would then remain valid if we replaced hyperbolic discounting by limits to contracting as the source of the commitment problem, and this whether we take the reputational or the monitoring foundations for the effectiveness of promises. Beyond stressing the generality of the usefulness of promises made even when one knows ex ante that, in equilibrium, they will be partially broken, this point argues for more mutual learning between the literatures on time-inconsistent preferences and on limits to contracting.

Appendix

A1. Proof of Proposition 1

Consider first the optimisation at $t = 1$. First, if $e^* \leq e_1(0)$, then by (3) the promise is not binding, so $e_1(e^*) = e_1(0)$. Second, recall that the optimal effort from self-0's perspective is $e_0(0)$. Now, given a promise e^*, if there is an interior solution to (2), it must satisfy:[13]

$$V_1(e_1, e^*) = 0 \;\Rightarrow\; \beta\Big[1 + f'(e^* - e_1)\Big] = \psi'(e_1). \tag{7}$$

Given Assumption 2, $V_{11}(e_1, e^*) < 0$ so the second-order condition of our maximisation problem is satisfied. From (7), $V_1(e^*, e^*) = \beta[1+f'(0)]-\psi'(e^*) = \psi'(\bar{e})-\psi'(e^*)$. Therefore, if $e^* \in (e_1(0), \bar{e})$, then $V(e, e^*)$ is increasing in e for all $e \in [0, e^*]$, and the promise will be fulfilled.

Third, taking the derivative of the first-order condition with respect to the promise yields:

$$V_{11}(e_1, e^*)\frac{\partial e_1}{\partial e^*} + V_{12}(e_1, e^*) = 0,$$

which implies:

$$\frac{\partial e_1}{\partial e^*} = \frac{\beta f''(e^* - e_1)}{\beta f''(e^* - e_1) + \psi''(e_1)}.$$

So, if $e^* > \bar{e}$, then $\dfrac{\partial e_1}{\partial e^*} < 0$ if $f'' < 0$ and $\dfrac{\partial e_1}{\partial e^*} \in (0, 1)$ if $f'' > 0$.

We can now turn to the optimisation at date $t = 0$. Obviously, the promise becomes binding only when $e^* \geq e_1(0)$. Then, self-0's optimisation problem $\mathcal{P}$ amounts to:

$$\mathcal{P}: \quad \begin{cases} \max\limits_{e^*} \quad e^* - \psi(e^*) & \text{if} \quad e^* \in [e_1(0), \bar{e}] \\ \max\limits_{e^*} \quad e_1(e^*) - f(e^* - e_1(e^*)) - \psi(e_1(e^*)) & \text{if} \quad e^* > \bar{e}. \\ \qquad \text{s.t.} \quad \dfrac{\partial e_1}{\partial e^*} = \dfrac{\beta f''(e^* - e_1)}{\beta f''(e^* - e_1) + \psi''(e_1)}. \end{cases}$$

These two cases thus concern respectively the second and third possibilities above.

[13] Subscript l in $V(\cdot)$ means partial derivative with respect to the lth argument.

By definition of $\bar{e}$, we have

$$\bar{e} > e_0(0) \;\Leftrightarrow\; f'(0) > \frac{1-\beta}{\beta},$$

so we can now conclude. First, case (i) of the Proposition is obvious. For cases (ii) and (iii), the first-order condition of the maximisation problem with respect to e^* yields:

$$\frac{\partial e_1}{\partial e^*}\Big(1-\psi'(e_1)\Big) - f'(e^*-e_1)\left(1-\frac{\partial e_1}{\partial e^*}\right) = 0.$$

Assumption 2 ensures that the second-order condition is satisfied.[14] Note that, for $e^* = \bar{e}$, the promise will be fulfilled and the derivative of the maximand at this point with respect to e^* is:

$$\frac{\beta f''(0)}{\beta f''(0)+\psi''(\bar{e})}\Big(1-\psi'(\bar{e})\Big) - f'(0)\frac{\psi''(\bar{e})}{\beta f''(0)+\psi''(\bar{e})}. \tag{8}$$

Assumption 2 ensures that the denominator is positive and that the derivative is decreasing in the promise from this point on. Then, if (8) is negative, we shall have a fulfilled promise $e^* = \bar{e}$ (case (ii)). Otherwise, we shall have a higher promise and a higher effort that however fails short of meeting the promise (case (iii)).

A2. Proof of Proposition 2

First, note that it is always optimal for type L to announce $e_L^p = \tilde{e}_L$ and exert effort $e_L = \tilde{e}_L$.[15] If a separating equilibrium exists, then the buyer can infer the type of the agent from his promise. Sellers exert the optimal effort from their self-1 perspective $\tilde{e}_i$, which also correspond to the buyer's selected technology. The payoff of type H from his self-0 perspective is then:

$$\beta\Big[\tilde{e}_H - \gamma_H\psi(\tilde{e}_H)\Big]. \tag{9}$$

[14] The S.O.C. is: $\frac{\partial^2 e_1}{\partial(e^*)^2}\big(1-\psi'(e_1)+f'(e^*-e_1)\big) - \psi''(e_1)\left(\frac{\partial e_1}{\partial e^*}\right)^2 - \left(1-\frac{\partial e_1}{\partial e^*}\right)^2 f''(e^*-e_1) < 0$. By Assumption 2, $\frac{\partial^2 e_1}{\partial(e^*)^2} < 0$ which guarantees that the second-order condition is satisfied.

[15] His problem is simple because we have assumed a cost for *underperforming* ($e < e^*$) but no cost for *overperforming* ($e > e^*$) relative to the buyer's selected technology. Consequently, the low-cost seller makes the same choices and obtains the same payoff in the separating and pooling equilibrium. This simplification is not crucial for the result that hyperbolic discounting and private information provide a foundation for promises that are announced and broken.

Instead, in a pooling equilibrium, type H makes the same promise as type L, that is $e_H^p = \tilde{e}_L$. The buyer then chooses a technology e^* which solves:

$$\max_{e^*} \quad p\ g(e_H, e^* - e_H) + (1-p)\, g(\tilde{e}_L, e^* - \tilde{e}_L) \tag{10}$$

where e_H is correctly anticipated. Simultaneously, type H solves:

$$\max_{e_H} \quad \beta\Big[e_H - f(e^* - e_H)\Big] - \psi(e_H).$$

From (10), the seller knows that any downward departure from $\tilde{e}_L$ will induce the buyer to choose a technology strictly between the two efforts, $e^* \in (e_H, \tilde{e}_L)$. Two cases are then possible.

First, if the cost $f(e^*(e_H) - e_H)$ of underperforming is steep enough at 0 and above, it is optimal to fulfill promises in equilibrium: $e_H = \tilde{e}_L$ and $e^* = \tilde{e}_L$. The payoff of the high-cost seller from his self-0's perspective is then:

$$\beta\Big[\tilde{e}_L - \gamma_H \psi(\tilde{e}_L)\Big]. \tag{11}$$

Second, if the cost $f(e^*(e_H) - e_H)$ of underperforming is not that steep, then in equilibrium $e_H < \tilde{e}_L$ and, by (10), $e_H < e^* < \tilde{e}_L$. Type H fails to fulfill his promise (for which he pays a cost) and his payoff from the perspective of date 0 is:

$$\beta\left[e_H - f(e^*(e_H) - e_H) - \gamma_H \psi(e_H)\right]. \tag{12}$$

Given (9), pooling necessarily implies that $e_H > \tilde{e}_H$. From the analysis above, it is clear that p, $f(\cdot)$ and $g(\cdot)$ will determine whether promises can serve as a commitment device and, if they do, whether they will be broken in equilibrium. However, recall that self-0 of type H would ideally want to implement an effort $\hat{e}_H > \tilde{e}_H$. Obviously $\tilde{e}_L > \tilde{e}_H$. Therefore, from (9) and (11), a sufficient condition for the seller to prefer a pooling rather than a separating equilibrium is $\tilde{e}_L < \hat{e}_H$. Given (4) and (5) this occurs when $\beta/\gamma_L < 1/\gamma_H$.

A3. Proof of Proposition 3

Suppose that meetings are organized every x periods. Recall that M observes with probability $q(e^* - e)$ a deviation from the prescribed effort level of P in the past y periods. Two cases must be analyzed separately.

If $x \leq y$, all past efforts can be observed by M. Renegotiation takes place with ex-ante probability $1/x$ (whenever the effort has to be exerted in the current period). Optimal effort is enforced with probability $(x-1)/x$, since P anticipates that otherwise he will be detected. Conditional on $x \leq y$, P optimally minimizes the probability of renegotiation and sets $x = y$.

If $x \geq y + 1$, shirking can go unnoticed. Effort can be enforced only if there is a meeting during one of the next y periods. This occurs with ex-ante probability y/x. With probability $1/x$ there is renegotiation. Last, with probability $(x - y - 1)/x$ there is no need to comply with the effort prescribed as there is no meeting in the following y periods. Conditional on $x \geq y + 1$, P optimally maximizes the probability of a meeting and sets $x = y + 1$.

The proof is completed by noting that $x = y + 1$ yields a higher utility than $x = y$: in both cases there is either high effort or renegotiation, and in the former the probability of renegotiation $(1/(y + 1))$ is lower than in the latter $(1/y)$.

References

1. Aghion, P. and Tirole, J. (1997). 'Formal and Real Authority in Organizations', *Journal of Political Economy*, vol. 105(1) (February), pp. 1-29.

2. Ainslie, G. (1992). *Picoeconomics*, Cambridge: Cambridge University Press.

3. Akerlof, G.A. (1991). 'Procrastination and Obedience', *American Economic Review*, vol. 81(2) (May), pp. 1-19.

4. Battaglini, M., Bénabou, R. and Tirole, J. (2005). 'Self-Control in Peer Groups', *Journal of Economic Theory*, vol. 123(2) (August), pp. 105-134.

5. Bénabou, R. and Tirole, J. (2004). 'Willpower and Personal Rules', *Journal of Political Economy*, vol. 112(4) (August), pp. 848-887.

6. Brocas, I. and Carrillo, J.D. (2001). 'Rush and Procrastination under Interdependent Activities', *Journal of Risk and Uncertainty*, vol. 22(2) (March), pp. 141-164.

7. Brocas, I. and Carrillo, J.D. (2004). 'Entrepreneurial Boldness and Excessive Investment', *Journal of Economics & Management Strategy*, vol. 13(2) (June), pp. 321-50.

8. Caillaud, B. and Jullien, B. (2000). 'Modeling Time-Inconsistent Preferences', *European Economic Review*, vol. 44(4-6) (May), pp. 1116-1124.

9. Caillaud, B., Cohen, D. and Jullien, B. (1996). 'Towards a Theory of Self-Restraint', *mimeo*, CEPREMAP, Paris.

10. Calvo, G. and Wellisz, S. (1978). 'Supervision, Loss of Control and the Optimal Size of the Firm', *Journal of Political Economy*, vol. 86(5) (October), pp. 943-952.

11. Carrillo, J.D. and Mariotti. T. (2000). 'Strategic Ignorance as a Self-Disciplining Device', *Review of Economic Studies*, vol. 67(3) (July), pp. 529-544.

12. Ellingsen, T. and Johannesson, M. (2004). 'Is there a Hold-up Problem?', *Scandinavian Journal of Economics*, vol. 106(3) (September), pp. 475-494.

13. Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). 'Time Discounting and Time Preference: A Critical Review', *Journal of Economic Literature*, 40(2) (June), pp. 351-401.

14. Grossman, S. and Hart, O. (1983). 'An Analysis of the Principal-Agent Problem', *Econometrica*, vol. 51(1) (January), pp. 7-45.

15. Holmström, B. and Kreps, D. (1995). 'Notes on a Theory of Promises', *mimeo*, MIT.

16. Laibson, D.I. (1997). 'Golden Eggs and Hyperbolic Discounting', *Quarterly Journal of Economics*, vol. 112(2) (May), pp. 443-477.

17. Loewenstein, G. and Prelec, D. (1992). 'Anomalies in Intertemporal Choice: Evidence and an Interpretation', *Quarterly Journal of Economics*, 107(2) (May), pp. 573-597.

18. Nocke, V. and Peitz, M. (2003). 'Hyperbolic Discounting and Secondary Markets', *Games and Economic Behavior*, 44(1) (July), pp. 77-97.

19. O'Donoghue, T. and Rabin, M. (1999). 'Incentives for Procrastinators', *Quarterly Journal of Economics*, 114(3) (August), pp. 769-816.

20. Phelps, E.S. and Pollak, R.A. (1968). 'On Second Best National Saving and Game-Theoretic Growth', *Review of Economic Studies*, 35(2) (April), pp. 185-199.

21. Strotz, R.H. (1956). 'Myopia and Inconsistency in Dynamic Utility Maximisation', *Review of Economic Studies*, vol. 23(3) (July), pp. 166-180.