A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality)

Jean Tirole

The Review of Economic Studies, Vol. 63, No. 1 (Jan., 1996), 1-22.

Stable URL:
http://links.jstor.org/sici?sici=0034-6527%28199601%2963%3A1%3C1%3AATOCR%28%3E2.0.CO%3B2-C

The Review of Economic Studies is currently published by The Review of Economic Studies Ltd..

Review of Economic Studies (1996) 63, 1–22 0034-6527/96/00010001$02.00

A Theory of Collective Reputations (with applications to the persistence of corruption and to firm quality)

JEAN TIROLE
Institut d'Economie Industrielle and GREMAQ, Toulouse and CERAS, Paris

First version received June 1993; *final version accepted August* 1995 (*Eds.*)

The paper is a first attempt at modelling the idea of group reputation as an aggregate of individual reputations. A member's current incentives are affected by his past behaviour and, because his track record is observed only with noise, by the group's past behaviour as well. The paper thus studies the joint dynamics of individual and collective reputations and derives the existence of stereotypes from history dependence rather than from a multiplicity of equilibria or from the existence of a common trait as is usually done in the literature. It shows that new members of an organization may suffer from an original sin of their elders long after the latter are gone, and it derives necessary and sufficient conditions under which group reputations can be rebuilt. Last, the paper applies the theory to analyse when a large firm can maintain a reputation for quality.

1. COLLECTIVE REPUTATIONS

Collective reputations play an important role in economics and the social sciences. Countries, ethnic, racial or religious groups are known to be hard-working, honest, corrupt, hospitable or belligerent. Some firms enjoy substantial rents from their reputations for producing high-quality products. Some departments are reported to treat their faculty or students fairly. The paper is a first attempt at modelling the idea of group reputation as an aggregate of individual reputations. A member's current incentives are affected by her past behaviour and, because her track record is observed only with noise, by the group's past behaviour as well. The paper studies the joint dynamics of individual and collective reputations in a model in which current generations are progressively replaced by new ones, and derives the existence of stereotypes from history dependence rather than from a multiplicity of equilibria or from the existence of a common trait as is usually done in the literature (see Section 2 for a detailed comparison with the literature). It shows that new members of an organization may suffer from an original sin of their elders long after the latter are gone, and it derives necessary and sufficient conditions under which group reputations can be rebuilt. Last, the paper applies the theory to analyse when a large firm can maintain a reputation for quality.

Let us spell out the building blocks of our theory in more detail:

(a) *A group's reputation is only as good as that of its members.* Each member is characterized by individual traits such as talent, diligence or honesty. Past individual behaviour conveys information about these traits and generates individual reputations.

(b) By contrast with group belonging, *individual past behaviour is imperfectly observed.* If past individual behaviour was fully unobserved, members of the group would have no incentive to sustain their own reputation and therefore the group would always be expected to behave badly. Conversely, the collective reputation would play no role if individual behaviours were perfectly observed. Imperfect observability of individual behaviour thus underlies the phenomenon of collective reputation.

(c) *The past behaviour of the member's group conditions the group's current behaviour and therefore can be used to predict the member's individual behaviour.* Each member's welfare and incentives are thus affected by the group's reputation.

(d) If we further assume that the age in the group, or the frequency of interactions with the group, or the number of past opportunities for cheating are imperfectly observed, cohorts in a self-regenerating group are partly pooled, and therefore, *the behaviour of new members of a group depends on the past behaviour of their elders.*

We offer two variants of the same model. In the first, a member's individual reputation is imperfectly observed by his potential "trading partner" (who may or may not belong to the group). The incentive to sustain an individual reputation stems from the member's fear of *direct exclusion* by the trading partner. By this we mean that the individual reputation may induce the trading partner to behave in a way undesirable for the member (e.g. by not trading), while the belonging to the group is not affected. We apply the direct exclusion variant to the issue of corruption[1] to explain why corruption is a societal phenomenon and why it tends to persist.

This direct exclusion variant does not seem appropriate when the trading partner has a low probability of knowing the member's past behaviour. The buyer of a car does not even know the names of the worker, foreman and engineer who built the car. Yet brand image is an important factor in the car market. The reason why the car manufacturer's employee has an incentive to maintain quality is the fear of *delegated or internal exclusion*: It may be in the interest of the firm to fire employees who have demonstrated undesirable traits.

In this case, the worker is not concerned by the possibility that his supplying poor quality will have a significant impact on buyers' demand for his work, but rather by the possibility of being fired. So, in the delegated exclusion variant the trading partner (the buyer) reacts to the collective reputation and the group (the firm) excludes on the basis of individual reputation, while in the direct exclusion variant the trading partner reacts to both collective and individual reputations and the group does not necessarily control membership. Yet the two variants are formally very similar because imperfect observability of individual behaviour plays the same central role[2].

Industrial organization modelling has depicted the firm as a black box in order to study its quality choices,[3] and has ignored the question of why workers have individual

1. The topic of many articles and books (see, e.g., Gould (1980), Hager (1973), Klitgaard (1986, 1988, 1991), Myrdal (1970), Lui (1986), Noonan (1984), Rose-Ackerman (1978), and Theobald (1990)), corruption has not yet attracted much attention from economic theorists. Notable exceptions include Acemoglu (1994), Andvig-Moene (1990), Banerjee (1994), Cadot (1987), Sah (1991), Strand (1990), and Shleifer–Vishny (1993). These papers focus on issues rather different from the ones considered here.

2. If workers' individual behaviour were not observed within the firm, there would be no incentive to sustain individual reputations, and firms could not build reputations for high quality. If workers' individual behaviour were perfectly observed within the firm, workers would have no incentive to sabotage the firm's reputation, at least under the classic conditions under which individual reputation is sustainable.

3. See Klein–Leffler (1981) and Shapiro (1983), as well as the incomplete information literature following Kreps *et al.* (1982).

incentives to defend the firm's collective reputation[4]. Our work opens the black box. Besides building foundations for the notion of firm quality, it also offers some interesting insights. It shows that a firm's reputation may be hard to rebuild once shattered, and that an increase in product market competition may make it difficult for firms to sustain their reputation.

The paper is organized as follows. Section 3 studies the case of direct exclusion by the trading partner, using corruption as a motivation. Section 3.1 sets up the model. Section 3.2 analyses potential steady states, and identifies a high- and a low-corruption steady state. Section 3.3 studies the issue of persistence of corruption by analysing the sensitivity of equilibrium to initial conditions. In the benchmark, the economy is in a steady state, with a low level of corruption. We then slightly perturb the economy by assuming that at the initial date (date 0), there is a one-shot increase in the gain to being corrupt (or a relaxation in the enforcement of anticorruption laws). Most agents alive at date 0 engage in the corrupt activity at that date. The economy is otherwise unchanged at dates $1, 2, \ldots$. We then ask whether the temporary increase in corruption necessarily has lasting effects, or whether the economy is able to go back to the low steady-state level of corruption. Interestingly, we find that under some conditions (in the unique continuation equilibrium) the economy must remain corrupt not only in the short run, but also in the long run.

Our analysis unveils two effects: First, the agents who were alive at date 0 have smeared their reputation. In our model, they have more incentives to engage in corrupt activities than if they had always behaved honestly. They are thus locked into corruption. This idea explains the short-run persistance of corruption: Shortly after date 0, there are lots of agents locked into corruption. This first effect however does not explain why the long-run steady state is affected by this one-shot increase in corruption, since we assume that agents are progressively replaced by new ones (that is, our model is one of overlapping generations). In particular, why do the agents who arrive with an unsmeared (individual) reputation also necessarily engage in corrupt activities? Why do the young inherit the corrupt practices of their elders?

The answer is that in the early periods after date 0, and because of imperfect observations of track records, the large number of agents who have been corrupt at date 0 and therefore remain corrupt raises a general suspicion. This suspicion affects new agents if their "age" (or more realistically, their age in the group or whether they had opportunities to get corrupt earlier) is imperfectly observed. Agents who arrive at date 1 are victims of this suspicion for at least a number T of periods and, if T is large enough (that is, if agents are not replaced very quickly), have no incentives to remain honest. This implies that the number of agents with a smeared record does not decrease. In turn, agents who arrive at date 2 are victims of this suspicion for at least T periods, and decide to become corrupt. And so forth. We therefore obtain a vicious circle of corruption, where the new generations suffer from the original sin of their elders long after the latter are gone. In this model, corruption ratchets up and not down, in the sense that a one-shot reduction in corruption due, say, to tough enforcement of anti-corruption laws has no lasting effect. It takes a

4. The works of Crémer (1986) and Kreps (1990) are exceptions to this rule. Among other things, our work departs from theirs in that the behaviour of a firm's employees is truly history-dependent. Crémer and Kreps develop repeated-game models of organizations with overlapping generations of workers in which future generations may punish current ones if these do not behave well. Kotlikoff *et al.* (1988), in a similar spirit, show that in an overlapping-generations framework, the young may refrain from taxing the old's capital for fear that the next generation would tax their capital. In these models, the set of equilibria at each point of time is history-independent. And there exist equilibria in which investments in reputation never pay off.

minimum number of periods without corruption to return to a path leading to the low-corruption steady state.

The minimum number of periods of anti-corruption campaigns required to allow a (long-run) return to the low-corruption steady state increases with the level of trust required by the principal, and decreases with the rate of renewal of generations, the probability of detection of corruption and with the young generations' eagerness to build a reputation. It is also shown that an amnesty, which is always detrimental in a steady state, yields a Pareto improvement.

Section 3.4 offers a more general analysis of off-steady-state behaviour (in the absence of anti-corruption campaigns or amnesties). It shows that even when the equilibrium is not unique, there always exists a Pareto-dominant equilibrium. This equilibrium either is a high-corruption equilibrium (if corruption has been widespread in the past), or, after a length of time over which suspicion is phased out, returns progressively toward the low-corruption steady state.

A variant of the corruption model is explored in Appendix 2. There the members of the group want to build or maintain a reputation for being corrupt because that enables them to extract bribes from their trading partners. In this extortion model, the members of the group attempt to develop a reputation for socially harmful rather than beneficial behaviour. Despite this sharp difference in interpretation, the extortion model is mathematically identical with the basic corruption model.

Section 4 extends the ideas to the case of exclusion by the group and applies them to formalize a firm's reputation for quality, and Section 5 concludes.

2. RELATIONSHIP TO THE LITERATURE

It is useful to explain how our history-dependence approach differs from two complementary approaches to the notion of stereotype.

2.1. *The theory of conventions (coordination in a situation of multiple equilibria)*

A *convention*[5] refers to the coordination on a particular Nash equilibrium in a situation of multiple Nash equilibria. Many models in economics have multiple equilibria, for example coordination games, repeated games, macroeconomic models with aggregate demand externalities or models of racial and sexual discrimination (which we discuss shortly). The interpretation of a convention as the selection of a particular equilibrium is stressed for example in Cole *et al.* (1992), Kandori (1992), Seabright (1992), and Young (1993). Kreps (1990) compares corporate culture to a convention, in that corporate culture in a firm is meant to communicate to its employees the (focal) behaviour that they are expected to follow.

One prominent model of convention is Arrow's (1973) statistical theory of discrimination of minorities by employers.[6] Arrow looks at a one-shot employment decision and

5. We use the sociological definition of convention (see, e.g., Elster (1989), Sugden (1989) and Ullman-Margalit (1977)). Economists often seem to call a norm what a sociologist would label a convention. The sociological notion of a *norm*, unlike that of a convention, stresses psychological factors at the expense of methodological individualism. That individuals are eager to be approved by others generates norms of etiquette or consumption norms. (Such norms might also be rationalized by the economic theory of wasteful signalling, but the emphasis is rather on the eagerness to be approved.) Unlike convention behaviours, norm behaviours need not be in one's self-interest (at least narrowly defined). Alternatively, individuals want to be approved by themselves ("I will not litter in the park even if no-one will see me").

6. See also Akerlof (1976), Coate-Loury (1991), Kremer (1993), Lundberg-Startz (1983), Milgrom-Oster (1987), Phelps (1972), and Rosen (1993) for related ideas.

assumes that workers first (secretly) invest in skills and then the employers run an imperfect test of the resulting ability. Because the test is imperfect, the employer uses the prior beliefs about whether the worker has invested in assessing the workers' true ability. If a higher prior belief that the worker has invested also makes it more profitable for the worker to invest, there is scope for multiple equilibria. The literature has interpreted the multiplicity of equilibria as the possibility of a differential treatment of workers based on their race, sex or other observable characteristics.

Our theory differs from the statistical discrimination theory (or, for the most part, the theory of conventions) in three important respects. First, there is a sense in which the statistical discrimination theory is not about societal behaviour; for, the multiplicity of equilibria in the discrimination model is independent of whether there are other employers or workers besides the employer and the worker in question. Because that theory applies equally well to a situation in which there is a single agent, there is nothing intrinsically linked to group behaviour. Second, the statistical discrimination theory, which is a static theory, is not about collective reputations, an intrinsically dynamic phenomenon. Third, the statistical discrimination theory shows that group stereotypes *may* emerge while they necessarily emerge in a history-dependence approach.

2.2. *Common trait*

While the theory of conventions relies on bootstrapping, a second approach attributes stereotypes to an intrinsic and unknown characteristic of the agents. The observation of one agent's behaviour then reveals information about this common trait and induces outsiders to update their beliefs about the likely behaviour of other agents in the same group (see Bénabou–Gertner (1993) for an application of this idea to a problem of search when sellers are affected by unobserved aggregate and idiosyncratic shocks, and Meyer–Vickers (1994) for an application to relative performance evaluation and career concerns). Interestingly, reputation then becomes a public good (Besley–Kandori (1992)). Suppose that a U.S. university considers admitting a PhD student from a foreign university, that it has limited information about the foreign university or its grading, and that it has admitted another student from that same university in the past. Then, the probability of admission of the latter student is likely to depend on the performance in the PhD programme of the former student.

The common trait model has a number of useful applications to professions (How well trained are the members of the profession?), franchised outlets (Are they being monitored carefully by the company? Is the recipe for Big Macs a good one?), or systemic risk in banking (Are certain types of balance- and off-balance positions currently risky?).

The common-trait and the history-dependence approaches depict different situations and differ in a number of respects. While the common trait approach has much to say about situations in which some common and unobserved training, supervision or shock generates informational externalities among members of the group, we are more reluctant to introduce a common trait to explain why some populations or groups are particularly corrupt, or why some firms' corporate culture generates high-quality products. Second, while the externalities generated by a common trait disappear as this common trait is learned, the reputational externalities in our model are long lasting. Third, the common-trait model is consistent with a transient (one-shot) belonging of the members to the group. By contrast, a key to our modelling is that members stay with the group for some extent of time and that their current behaviour is affected by their own past behaviour (through individual reputations).

3. INDIVIDUAL AND SOCIAL STIGMAS: THE CASE OF TRUST

3.1. *The model*

This section develops a simple model in which the efficient organization of economic activity requires a minimum level of trust between contracting parties. More precisely, a principal (the buyer of a service) will contract with an agent (the supplier of the service) only if she is sufficiently confident that the agent will not engage in corrupt activities. The principal has some, albeit imperfect, information about the agent's track record, namely about whether the agent has engaged in corrupt activities in the past.

Matching. We consider a large stationary economy ($-\infty < t < \infty$) in which agents alive at date t remain in the economy up to (at least) date $t+1$ with probability $\lambda \in (0, 1)$. With this "Poisson death process", we assume that each quit is offset by the arrival of a new agent, so that the population of agents is constant. The model is a matching model. Agents have no chance to meet the same principal twice. At each date t, each (alive) agent is matched with a new principal[7]. The principal decides whether to offer task 1 or task 2 to the agent. Task 1 is the efficient task. Task 2 is a less efficient task, but, for the principal, it is less sensitive to the agent's choosing to be corrupt. (In a slightly different version of the model, task 2 corresponds to the absence of a hire.) We will make an assumption guaranteeing that it is always optimal for the principal to at least offer task 2 to the agent rather than not hiring him. Once hired, the agent chooses whether to engage in the corrupt activity, that is whether "to cheat" (behave dishonestly). The principal's payoff from task 1 in the period is H if the agent behaves honestly and D if he cheats. Similarly her payoffs from task 2 are h and d. That task 1 is more sensitive to corruption than task 2 (given that the principal faces a non-trivial choice) means that

$$H > h \geqq d > D.$$

We also assume that $d \geqq 0$ so that it is optimal to hire the agent.

Agents' preferences. There are three types of agents: "honest", in proportion α, "dishonest", in proportion β, and "opportunistic", in proportion γ, where $\alpha + \beta + \gamma = 1$. The proportions are the same for each cohort and therefore for the entire population. Honest agents have a strong distaste for and never engage in corrupt activities (alternatively, if corruption has a probability of being exposed and directly punished, "honest" agents might be ones for whom being punished is very costly). Dishonest agents always cheat, for instance because they derive a high benefit from it (alternatively, in a slightly different model, they might be transient agents who do not care about their reputation). Because honest and dishonest agents behave mechanistically (never and always cheat, respectively), the focus of our analysis is on opportunists.[8] These have no aversion to being corrupt, but trade-off the current benefit from corruption and the loss in reputation. Their benefits from being hired in tasks 1 and 2 and not cheating are B and b, respectively, where

$$B > b \geqq 0.$$

They enjoy an additional short-run gain $G > 0$ from being corrupt in either task. That G is the same in both tasks simplifies the formal analysis. Note also that we do not model

7. Principals can be either short-lived or long-lived.

8. This formulation of preferences is standard in reputation models, see, e.g., Diamond (1991). The presence of honest agents creates an incentive for opportunists to build a reputation. The presence of dishonest agents eliminates a spurious multiplicity of equilibria associated with an indeterminacy of off-the-equilibrium path beliefs.

explicitly anti-corruption campaigns.[9] The simplest, albeit extreme interpretation of the model is that there is no hard evidence that could lead to the indictment of a corrupt agent. Alternatively, G could be an expected gain from being corrupt, which would allow a probability of confronting legal sanctions. Last, the agents' discount factor is δ_0. We will let $\delta \equiv \delta_0 \lambda \leqq 1$ denote the "relevant discount factor".

Information. Agents know their own preferences (that is, their types). Principals know the proportions α, β, γ and imperfectly observe the track record of the agent they are matched with. There are several ways of formalizing the imperfect observability of the track record. We choose a simple one in order to illustrate easily the main ideas. The principal has probability x_k of finding out that the agent has engaged in the past at least once in a corrupt activity when the agent has in fact cheated k times[10]. So the *observed* track record, that is the information of the principal the agent is matched with, is binary. The principal knows that the agent has been corrupt at least once, or has no such knowledge.[11] The assumption that the principal does not know the agent's age is important for the effect unveiled in Section 3.3 and giving rise to everlasting effects of a one-time shock in corruption. Of course this assumption should not be taken too literally. It is a metaphor for the idea that the principal may not be fully informed about the number of times the agent had an opportunity to be corrupt in the past or about the length of the member's relationship with the group.

Assumption 1.

$$x_0 = 0 < x_1 < x_2 < x_3 < \cdots < 1 \quad \text{and} \quad x_{k+1} - x_k < x_k - x_{k-1} \quad \text{for all } k.$$

Assumption 1 says that the leakage of information about corruption becomes more likely when the agent has cheated more in the past; and that this increase occurs at a decreasing rate. The second part of this assumption simplifies the analysis by guaranteeing that an individual is locked into corruption after having been corrupt a certain number of times.

3.2. *Possible steady states*

We first analyse steady states of the model developed in Section 3.1. There are either one or three steady states. In the latter case, two involve pure strategies and for conciseness

9. Carrillo (1995a, b) develops a different, hierarchical model of corruption and analyzes the impact of wages, monitoring and promotion policies on the extent of corruption.

10. It would be interesting to extend the analysis to alternative information technologies. In particular it would seem reasonable to allow for forgetfulness (witnesses or evidence disappear over time). Our insights ought to carry over to such specifications, but new insights (such as the possibility of an individual's resuming an honest behaviour after being corrupt) would arise.

We have performed a different check of robustness by assuming that once an individual is exposed a public file exposes him for the rest of his life. The expressions of Y and Z below are slightly altered, but the analysis goes through under the same Assumptions 1 and 4. It is then very similar to that of Section 4, in which, once exposed, the agent is excluded for the rest of his life.

11. The analysis can be extended to the more general case in which the number of times the agent has cheated in the past is observed with noise by the principal, but at the cost of substantial complications. Note also that it does not matter whether the agent knows that the current principal knows his record, as the principal moves first.

will be the focus of this section, and one is in mixed strategies.[12] It is worth emphasizing that the possible multiplicity of steady states, while interesting in its own right, is not the primary focus of the paper, and that, for given initial conditions, there may be a unique equilibrium even when there are several steady states, as we will see. Furthermore, we will later select the Pareto-dominant equilibrium when there are multiple equilibria.

(*a*) *Low-corruption steady state.* Suppose that all opportunists always behave honestly. A principal offers task 2 to an agent who she knows has been corrupt in the past, since the agent is necessarily a dishonest agent and since $d > D$. In contrast, when the principal has no such information, the agent may be honest or opportunistic, or else be a dishonest agent with a deceivingly clean observed track record. The proportion of honest and opportunistic agents in the economy is $(\alpha+\gamma)$. The proportion of dishonest agents with a clean track record is βY where Y is the average probability that past corruption activities go unnoticed[13]:

$$Y=(1-\lambda)[1+\lambda(1-x_1)+\lambda^2(1-x_2)+\cdots+\lambda^k(1-x_k)+\cdots].$$

The probability that the agent will not cheat given a clean observed record is $(\alpha+\gamma)/(\alpha+\gamma+\beta Y)$. The principal offers task 1 if and only if the following assumption holds:[14]

Assumption 2.

$$\frac{\alpha+\gamma}{\alpha+\gamma+\beta Y}(H-h)+\frac{\beta Y}{\alpha+\gamma+\beta Y}(D-d)>0.$$

Do opportunists have an incentive not to become corrupt? By never being corrupt, they keep a clean (real and observed) record and are always offered task 1. Their payoff is therefore $B+\delta B+\delta^2 B+\cdots=B/(1-\delta)$. Suppose that they instead cheat today and keep cheating in the future. Their expected payoff is then

$$(B+G)+\delta(B+G)\left[\frac{1}{1-\delta}-Z\right]+\delta(b+G)Z,$$

where

$$Z=x_1+\delta x_2+\delta^2 x_3+\cdots$$

is the present discounted probability of being found out in the future given that one has cheated once and will continue cheating. So, a necessary condition for a low-corruption

12. The reader will check that when the two pure-strategy equilibria co-exist, the mixed-strategy equilibrium has the following characteristics. When faced with an agent with a spotless record, principals offer task 2 with probability θ and task 1 with probability $1-\theta$. An agent with a spotty record is offered task 2. The parameter θ satisfies

$$G=(1-\delta)\delta(B-b)\theta Z$$

(see below for the definition of Z). Agents who have cheated in the past cheat. Agents who have not yet cheated randomize between cheating and being honest, in such a way that the overall probability of honest behaviour, v, satisfies

$$v(H-h)+(1-v)(D-d)=0.$$

13. The proportion of "newborns" (who therefore have not yet cheated) is $(1-\lambda)$, the proportion of "one-period old" (who have cheated once) is $(1-\lambda)\lambda$, and so forth.

14. If Assumption 2 is violated, it can be shown that the only *equilibrium* (and not only steady state) starting from any initial condition has the principals always offering task 2, and the opportunists always cheating.

steady state is:

Assumption 3.

$$\frac{G}{1-\delta} \leqq \delta(B-b)Z.$$

Appendix 1 shows that *the low-corruption steady state indeed exists under Assumptions 1 through 3*. The intuition is that from Assumption 1, the agent has more incentive to be corrupt, the more he has been corrupt in the past. In this sense, *agents are locked into corruption once they start being corrupt.*

Note also that a low-corruption steady state exists only if the principals are not poorly informed[15]. Agents must have enough incentives to maintain their reputation for honesty.

(*b*) *High-corruption steady state.* Suppose now that opportunists are always corrupt and principals always offer task 2. Because keeping a clean slate has no value, it is indeed optimal for opportunists to be always corrupt. Is is optimal for a principal to offer task 2 to an agent with a clean slate? Such an agent is honest with probabliity $\alpha/[\alpha+(\beta+\gamma)Y]$ and either opportunistic or dishonest with probability $(\beta+\gamma)Y/[\alpha+(\beta+\gamma)Y]$. We thus make

Assumption 4.

$$\frac{\alpha}{\alpha+(\beta+\gamma)Y}(H-h)+\frac{(\beta+\gamma)Y}{\alpha+(\beta+\gamma)Y}(D-d)<0.$$

The high-corruption steady state exists if and only if Assumption 4 holds. Note that Assumption 4 holds when there are enough opportunistic and dishonest agents and when the principals' information is not very precise. The role of imperfect observability is highlighted by the facts that Assumption 3 is violated if the principals' information is very bad and that Assumption 4 is violated if the principals' information is very good and λ is close to 1. The reader can also check that an increase in the population's renewal rate $1/\lambda$ makes it harder for Assumptions 2 and 3 to be satisfied (and so for the low-corruption steady state to exist) and easier for Assumption 4 to hold (and so for the high-corruption steady state to exist).

Proposition 1. (*Steady states*). *The economy has three steady states when Assumptions 2, 3 and 4 are satisfied (and one otherwise). The multiplicity of steady states stems from the dynamic complementarity between past and future reputations. Due to the (endogenous) hysteresis in individual behaviours, good collective behaviour in the past increases trust in the future and thus raises individual incentives to maintain a reputation, resulting in collective reputation being maintained at a high level.*

3.3. *Persistence of corruption: an example*

We now investigate the effect of a one-time shock in corruption on the equilibrium. To keep the analysis simple, we specialize the model further in this section by making

Assumption 5. $x_1=x_2=\cdots=x\in(0,1)$.

15. If the x's are close to zero, Z is close to zero and Assumption 3 is violated.

That is, the probability of exposure to corrupt activities is independent of the number of past corrupt acts. Assumption 5, which is a limit case of Assumption 1, implies in particular that an opportunist remains corrupt once he has started; it also implies that $Y=1-\lambda x$ and $Z=x/(1-\delta)$.

We make Assumptions 2 through 4. Let the economy start in the low-corruption steady state (see Section 3.4 for a justification). Suppose now that the economy faces a temporary shock at date 0. The gain from being corrupt at that date is very large, and so all opportunistic agents alive at date 0 get corrupt. The parameters of the model (including the gain G from cheating) are unchanged at dates $1, 2, \ldots$. We show that under an additional assumption, the economy cannot go back to the low-corruption steady state. Indeed, the *unique* continuation equilibrium exhibits a high level of corruption forever.

Let us perform the following thought experiment. Suppose that the opportunistic agents born at date 1 through t behave honestly before and at date t. This presumption gives the best chance to the existence of trust at date t. The probability of honest behaviour at date t given an observed clean record and given that opportunists born at or before date 0 are locked into corruption is

$$p(t)\equiv\frac{\alpha+\gamma(1-\lambda)(1+\lambda+\cdots+\lambda^{t-1})}{[\alpha+\gamma(1-\lambda)(1+\lambda+\cdots+\lambda^{t-1})]+[\beta Y+\gamma(1-x)(1-\lambda)(\lambda^t+\lambda^{t+1}+\cdots)]}$$

$$=\frac{\alpha+\gamma(1-\lambda^t)}{[\alpha+\gamma(1-\lambda^t)]+[\beta Y+\gamma(1-x)\lambda^t]}=\frac{1-\beta-\gamma\lambda^t}{1-\beta\lambda x-\gamma x\lambda^t}.$$

Recall that $p(1)(H-h)+(1-p(1))(D-d)<0$ (this is Assumption 4) and that $p(\infty)(H-h)+(1-p(\infty))(D-d)>0$ (this is Assumption 2). Noting that p is an increasing function, we let T denote the largest t such that

$$p(T)(H-h)+(1-p(T))(D-d)<0. \tag{1}$$

That is, under the most optimistic assumption, principals still do not trust agents with observed clean records at date T; thus $(T+1)$ is the minimum length of time for suspicion to be phased out. Suppose now that

Assumption 6.

$$G(1+\delta+\cdots+\delta^{T-1})>\frac{x\delta^T(B-b)}{1-\delta}\Leftrightarrow G(1-\delta^T)>x\delta^T(B-b).$$

Assumption 6 states that it is a dominant strategy for an agent born at date 1 to cheat at date 1 (and therefore forever) given that the agent will not be trusted before (at best) date $(T+1)$. The left-hand side of Assumption 6 is the gain from cheating from date 1 through date T (discounted at date 1), and the right-hand side is an upper bound on the cost of not being offered task 1 after date $(T+1)$. Note that Assumption 6 requires T not to be too small, since with x_k constant for $k\geqq 1$, Assumption 3 is equivalent to $G\leqq x\delta(B-b)$.

Consider now the generation born at date 2. All its elders have been corrupt in the past, and Assumption 6 ensures similarly that cheating at date 2 and thereafter is a dominant strategy. By induction, the same is true for all generations. *Corruption has ratcheted up and does not subside even after the generations that have committed the original sin have by and large disappeared.* Conversely, if Assumption 6 is violated, then the

economy can return to the low-corruption steady state in the long-run: Just specify that all opportunistic agents joining the group after date 0 behave honestly and that principals offer task 2 from date 1 through date T, and offer task 1 (for a spotless record) from date $T+1$ on. This equilibrium converges to the low-corruption steady state as $t \to \infty$.

This simple model also illustrates the possible failure of a short-run anti-corruption campaign. Suppose that at date 1 (or, equivalently at any later date) the government runs a tough anti-corruption campaign that lasts one period and makes it unprofitable for opportunists to engage in corruption at that date. Suppose further that the following strengthening of Assumption 6 holds:

$$G(1+\cdots+\delta^{T-2}) \geqq x\delta^{T-1}(B-b)/(1-\delta).$$

Then it is a dominant strategy for generations born at dates 1 and 2 to cheat at date 2, and corruption prevails at all dates after date 1. *The anti-corruption campaign only implies a decrease in corruption during the campaign and has no effect thereafter. Corruption does not ratchet down.*

Comparative statics. Condition (1) defines the minimum length of time for suspicion to phase out. Let $T = T(\lambda, x, \rho)$ where $\rho \equiv (d-D)/[(H-D)-(h-d)]$ is the smallest probability of honest behaviour that makes a principal choose task 1. T is increasing in λ and ρ and decreasing in x: The suspicion takes longer to phase out, the slower the rate of renewal of generations, the higher the level of trust required by principals, and the smaller the probability of detection of past corrupt behaviour.

Thus, perpetual corruption (that is, an equilibrium coinciding with the high-corruption steady state from date 1 on) is the unique continuation equilibrium if (rewriting Assumption 6):

$$T(\lambda, x, \rho) > \frac{\log\left(1+\dfrac{x(B-b)}{G}\right)}{\log\left(\dfrac{1}{\delta}\right)}.$$

We see that the increase in the probability x of detection not only lowers T, but also raises the young generations' incentive to be honest, so both effects reinforce each other. As for the effect of λ, we must keep $\delta = \lambda\delta_0$ constant to obtain an unambiguous conclusion; for, if we keep the interest rate (δ_0) constant, an increase in λ has two opposite effects: It makes young generations more eager to maintain a good reputation, but it also slows down the renewal of generations.

We can alternatively state our results in terms of the smallest number τ of periods of anti-corruption campaigns needed to allow a (long-run) return to the low-corruption steady state. (An anti-corruption campaign at date t is defined as one in which the probability of being caught and the resulting penalties are high enough so that being corrupt at date t is a dominated strategy for an opportunist. Dishonest agents still cheat during the campaign.) A prolonged anti-corruption campaign from date 1 through date τ allows a (long-run) return to the low-corruption steady state if and only if opportunists with a spotless record at $\tau+1$ find it optimal to be honest conditionally on trust being restored from date $T+1$ on. Thus the minimum number of periods of anti-corruption campaigns

is the smallest integer τ such that:

$$T(\lambda, x, \rho) - \tau \leqq \frac{\log\left(1 + \frac{x(B-b)}{G}\right)}{\log\left(\frac{1}{\delta}\right)}.$$

Amnesty. Consider now a policy of amnesty for past corrupt behaviour (if feasible). One may for example have in mind a strict libel law that severely punishes those who communicate to a principal evidence about an agent's past corrupt behaviour (alternatively, an amnesty could exempt the corrupt agent from a jail sentence in the variant in the spirit of Section 4, in which punishments are inflicted by the group itself.)

Amnesty is irrelevant in a high-corruption steady state, in which there is no trust anyway. And it can only be detrimental if the low-corruption steady state prevails, because it destroys the information principals use to give task 2 to some of the corrupt agents, and (for a sustained amnesty) because it destroys opportunistic agents' incentives to behave honestly.

By contrast, an amnesty may be welfare enhancing out of steady state. An amnesty limited to corrupt acts performed at date 0 enables the group to return immediately to the low-corruption steady state. (Of course, the issue is whether such an amnesty is feasible. It certainly is possible to exempt corrupt acts at date 0 from jail sentences or fines, but it may be harder for a government to dispel mistrust among citizens through a governmental measure such as a libel law.)

We summarize the results of this section in:

Proposition 2. *Consider a one-time shock in corruption (all opportunistic agents are corrupt at date 0). Under Assumptions 2 through 6:*

(a) in the absence of anti-corruption campaigns or amnesty, the high-corruption steady state (perpetual corruption) is the unique continuation equilibrium;

(b) the minimum number of periods of anti-corruption campaigns (discouraging opportunists from being corrupt during the period) required to allow a (long-run) return to the low-corruption steady state increases with the level of trust ρ required by principals, and decreases with the rate of renewal of generations (keeping the agents' discount factor constant), with the probability of detection of corruption, and with the agents' "eagerness" $(B-b)/G$ to maintain a reputation;

(c) an amnesty limited to corrupt acts performed at date 0 enables the group to return immediately to the low-corruption steady state, and yields a Pareto improvement relative to the unique continuation equilibrium described in (a).

3.4. *Basins of attraction of the steady states: General results*

We now return to the model of Sections 3.1 and 3.2 (that is, we no longer make Assumptions 5 and 6), but we allow the history of corruption at date 1 to be arbitrary. That is, at dates $t \leqq 0$, some cohorts of opportunists may have been corrupt and some honest (a given cohort of opportunists may even have had both behaviours at some date).

Let us define strategies from $t \geqq 1$ on a bit more formally. Let μ_t denote the probability that at date t a principal (or principals "on average") chooses task 1 for an agent with a

spotless record. Let θ_{kt} denote the expected (over all cohorts) probability that an opportunistic agent who has cheated k times before date t behaves honestly at date t.[16] To simplify the analysis, we make an assumption that guarantees that it is a dominant strategy for a principal to offer task 2 when confronted with an agent with a spotty record:

Assumption 7.

$$\frac{\gamma}{\beta+\gamma}(H-h)+\frac{\beta}{\beta+\gamma}(D-d)<0.$$

(Note that for any history, $\beta/(\beta+\gamma)$ is a lower bound on the probability that the agent with a spotty record is dishonest and therefore on the probability that he will be corrupt.)

The game exhibits strong strategic complementarities. *Heuristically*, for all t, a weak increase in θ_{kt} for some k weakly increases the reaction correspondence μ_t, and has no impact on the other reactions $\{\mu_\tau\}_{\tau<t}$ (and has no "cross-effects" on $\{\theta_{k't'}\}_{(k't'\neq kt)}$). Conversely, an increase in μ_t weakly increases the reaction correspondences $\theta_{kt'}$ for all k and $t'<t$, and has no impact on others. Because of these strategic complementarities, it turns out that one equilibrium is preferred by the principals and by all agents (regardless of their types and histories). (We have not quite demonstrated that the game is supermodular. We have not shown that an increase in θ_{kt} for some k weakly increases the reactions $\{\mu_\tau\}_{\tau>t}$. A sufficient condition for this would be that $\theta_{\cdot t'}$ be weakly decreasing for all $t'\in\{t+1,\ldots,\tau\}$. We have not proved this generalization of the steady-state property demonstrated in Appendix 1. Yet we conjecture that the game can be made supermodular. If this conjecture turns out to be correct then part (a) of Proposition 3—the existence of a Pareto-dominant equilibrium—is simply Topkis' theorem (1979).)

When there are multiple equilibria (which need not be the case: see Section 3.3), we thus select, we feel in a reasonable way, a unique equilibrium. This illustrates again the sharp contrast in approach with the theory of conventions (including the statistical theory of discrimination) which makes the multiplicity of equilibria the focus of the analysis and assigns different equilibria to different groups. We now ask whether the equilibrium converges to one of the three steady states and we characterize its dynamics:

Proposition 3. *Under Assumptions 1, 2, 3, 4 and 7, for any history of corruption before date 1, starting at date 1:*

(a) there exists a Pareto-dominant equilibrium,

(b) either the equilibrium is unique and coincides from date 1 on with the high-corruption steady state, or there exists $T\geqq 0$ such that in the Pareto-dominant equilibrium trust prevails from date $T+1$ on (that is, $\mu_t=0$ for $t\leqq T$ and $\mu_t=1$ for $t\geqq T+1$), and this equilibrium converges to the low-corruption steady state as t goes to infinity.

Proof. Either there is is a unique equilibrium, which satisfies $\mu_t=0$ for all $t\geqq 1$, and then this equilibrium necessarily coincides with the high-corruption steady state, or else there exists an equilibrium and a date t such that $\mu_t>0$. To prove the proposition, we

16. Note that we impose that this probability is the same whether the agent is given task 1 or 2. This is a reasonable restriction as the agent's von Neumann–Morgenstern payoffs when choosing whether to be corrupt at t are independent of the task allocation at that date (the benefit from being assigned to task 1 rather than to task 2 at date t is additive).

show that this equilibrium is weakly dominated by another equilibrium satisfying $\mu_t = \mu_{t+1} = \mu_{t+2} = \cdots = 1$ (if the high-corruption equilibrium also exists, it is clear that it is then dominated).

Let

$$L_k \equiv \frac{\delta}{1-\delta}(x_{k+1} - x_k)(B-b).$$

The sequence L_k is strictly decreasing and converges to 0 from assumption 1. Let

$$\mathscr{L}_k \equiv [\textstyle\sum_{n \geq 1} \delta^n (x_{k+n} - x_k)](B-b)$$

$$= \textstyle\sum_{n \geq 0} \delta^n L_{k+n} < \dfrac{L_k}{1-\delta}.$$

The sequence $\mathscr{L}_k$ is also strictly decreasing and converging to 0. Last, let k^* be the smallest k such that

$$\mathscr{L}_k < \frac{G}{1-\delta}.$$

k^* is strictly positive from Assumption 3. Note that, in an hypothetical situation in which the principals always trust the agents with apparently spotless records, an opportunistic agent who has cheated k times in the past prefers cheating forever rather than never cheating again if and only if $\mathscr{L}_k < G/(1-\delta)$.

We claim that cheating is a dominant strategy at any date for an opportunist who has cheated at least k^* times in the past. Note first that L_k is an upper bound on the future cost of cheating today for an opportunist who has cheated k times in the past (cheating today increases the probability of being caught in any future date by at most $x_{k+1} - x_k$, which creates a maximum loss of $B-b$ in that period). Hence for k large enough, $L_k < G$, and it is indeed a dominant strategy to cheat forever.

Second, suppose that "always cheat" is a dominant strategy for an opportunist who has cheated at least $k+1$ times. Consider an opportunist who has cheated k times in the past and compare the strategy of behaving honestly for N periods and cheating afterwards (which is then optimal from the induction hypothesis), and the strategy of not waiting and cheating right away. The difference in continuation payoffs between the two strategies, Δ, can be bounded above by assuming that $\mu_{t+1} = \cdots = 1$ (even if this is not the case). So,

$$\begin{aligned}\Delta \leqq & [\delta(x_{k+1} - x_k) + \cdots + \delta^N (x_{k+N} - x_k) + \delta^{N+1}(x_{k+N+1} - x_{k+1}) \\ & + \delta^{N+2}(x_{k+N+2} - x_{k+2}) + \cdots](B-b) - G(1 + \delta + \cdots + \delta^{N-1}) \\ = & (1-\delta^N)\left(\mathscr{L}_k - \frac{G}{1-\delta}\right) < 0,\end{aligned}$$

as long as $k \geqq k^*$. The induction hypothesis is thus confirmed.

So, at any date t, all opportunistic agents who have been corrupt at least k^* times cheat. Conversely, *if* $\mu_{t+1} = \mu_{t+2} = \cdots = 1$, then from the definition of k^* it is optimal for an opportunistic agent who has cheated fewer than k^* times in the past not to cheat at date t. Furthermore, *if* all such agents do not cheat at date t, then principals at date t *a fortiori* are willing to trust the agents (which they already were willing to do, since $\mu_t > 0$), and we can set $\mu_t = 1$ as well. We thus conclude that $\{\mu_t = \mu_{t+1} = \mu_{t+2} = \cdots = 1$, and $\theta_{kt} = \theta_{k,t+1} = \cdots = 1$ for $k < k^*$, and $= 0$ for $k \geqq k^*\}$ is a continuation equilibrium from

*t*on, which weakly dominates the presumed equilibrium. To complete the proof, we can then work by backward induction: The substitution of the continuation equilibrium from date t on weakly improves the best equilibrium $\theta_{k,t-1}$ for all k, which in turn weakly improves the highest equilibrium μ_{t-1}, and so forth. So, Proposition 3 holds. ‖

Proposition 3 also vindicates our focus on perpetual corruption vs. return to the low-corruption steady-state for the more special history considered in Section 3.3.

Last, it is worth pointing out that one can extend the analysis of Section 3.3 to general histories at date 1: From Proposition 3, either corruption is long lasting (case a) or one can find some date T such that the equilibrium evolves as in case b. This date T (if it exists) is given by a fixed point: Given T, the opportunists' behaviour is simply to always cheat if k (their history of corruption at date 1) exceeds some cutoff $k^*(T)$ where $k^*(T) \leqq k^*$ is weakly decreasing in T, and to always be honest if $k < k^*(T)$. Conversely, this behaviour, together with the distribution over k at date 1, generates some minimal length of time for suspicion to phase out, in a manner similar to equation (1).

4. EXCLUSION FROM THE GROUP: THE CASE OF A FIRM'S REPUTATION FOR QUALITY

When the trading partner hardly observes the past individual behaviour of the member, the latter's incentive to behave well can only come from the threat of retaliation by the group itself. We now assume that belonging to the group generates a rent but is no longer a *fait accompli.* While we develop the analysis in the context of a firm's reputation for quality, *it applies* equally well *to any organization or group that co-opts its members and can freely exclude them.*

We consider a stylized model of a firm as a workers' cooperative. Each period the workers share and consume the firm's profit. (We abstract from issues such as unequal treatment, hierarchies and delayed compensation in order to focus better on that of collective reputation. The key to the results is that incentive problems are not perfectly solved by alternative methods and that the workers enjoy a higher rent in a higher reputation firm.) Demand is inelastic and constant over time; the (large) number of workers is constant and, by normalization, equal to one worker per unit of good produced in a period. Workers who either quit (which occurs, as earlier, with Poisson probability $(1-\lambda)$) or are fired are immediately replaced. No screening among workers at the hiring stage is feasible in our model. We also assume for the moment that there is no cost for the firm of firing a worker and hiring a new one.

The consumers in each period observe the *firm's* track record, namely the average quality of items produced in each past period, but not the quality of the item they purchase. An item's quality is equal to H (high) or L (low), with $H > L > 0$. The consumers' (common) reservation price at date t is equal to the expected quality produced by the firm conditional on the firm's track record. Consumers do not observe the individual track record of the worker who has produced the particular item they buy. So, if ν_t is the consumers' posterior probability of buying a high-quality item given the firm's track record, the firm charges price $p_t = \nu_t H + (1-\nu_t)L$.

Producing a low-quality item costs nothing to the worker in charge of the item, while producing a high-quality unit involves a disutility of effort. There are three types of workers. Using the same notation and terminology as earlier, a worker is "honest" with probability α, meaning that this worker has no disutility of effort and always produces a high-quality item. With probability β, the worker is "dishonest"; he then has a very high

disutility of effort for producing high quality and always produces a low-quality item. Last, with probability γ, the worker has disutility of effort G of producing high quality and behaves opportunitistically.

Keeping with the notation of the paper, let x_k denote the probability that the firm, i.e., the co-workers, find out that a worker has produced at least one low-quality item in the past when the agent has in fact produced k low-quality items in the past. The firm may find out either directly through word of mouth or observation of past decisions of the worker, or indirectly through consumers' complaints about the durability of the product. The sequence $\{x_k\}$ satisfies Assumption 1. For consistency with Section 3, we assume that the firm does not know the "age" of its workers (or, more realistically, the number of times they had an opportunity to shirk); however, the analysis would not be affected if the firm knew the workers' age, because, in the derivations below, the firm fires a worker anytime it has evidence of a low-quality production. As before, we also assume that, in each period, conditionally on their not being fired, workers stay in the firm with probability λ (the survival rate), and we let δ denote the relevant discount factor, namely λ times the workers' discount factor.

We look at steady states and define a high- (low-) reputation firm as a firm in which opportunists always produce high- (low-) quality items. We follow Section 3 in proving the possibility of existence of a good and a bad steady states. In both cases, the firm keeps workers for whom it has no evidence of wrongdoing and fires the others.

(*a*) *High-reputation firm.* The expected present discounted tenure in the firm of a worker who produces low quality in each period is:

$$\tilde{Y}=(1-\lambda)\sum_{t=0}^{\infty}\lambda^t(1-x_0)\cdots(1-x_t),$$

(where $x_0=0$): λ^t is the probability of not quitting the firm before age t, and $(1-x_0)\cdots(1-x_t)$ is the probability of not being caught. Note that $\tilde{Y}$ differs from $Y=(1-\lambda)\sum_{t=0}^{\infty}\lambda^t(1-x_t)$ only to the extent that exclusion when it happens is permanent and not temporary. Assuming that new workers are drawn in a pool with proportions (α, β, γ) of honest, dishonest and opportunistic workers[17], the steady-state proportions in a high-reputation firm are $(\alpha, \beta\tilde{Y}, \gamma)$. Consumers therefore pay per unit:

$$p_H=\frac{\alpha+\gamma}{\alpha+\gamma+\beta\tilde{Y}}H+\frac{\beta\tilde{Y}}{\alpha+\gamma+\beta\tilde{Y}}L.$$

A necessary condition for an opportunist to produce high quality is that he prefers to always produce high quality rather than always producing low quality:

$$\frac{p_H-G}{1-\delta}\geqq p_H\sum_{t=0}^{\infty}\delta^t(1-x_0)\cdots(1-x_t)\equiv p_H\left[\frac{1}{1-\delta}-\delta\tilde{Z}\right],$$

where $\delta\tilde{Z}$ is the expected present discounted reduction in tenure due to producing low quality. Again, $\tilde{Z}$ differs from Z only to the extent that exclusion is permanent. We thus make:

Assumption 3'. $G/(1-\delta)\leqq\delta p_H\tilde{Z}$.

17. In a general equilibrium context, this assumption means that, once fired by a firm, workers are not re-hired by another firm; otherwise dishonest and (depending on the equilibrium) opportunistic workers would be over-represented in the labour market compared to the prior distribution. We could alternatively have conducted the analysis under the assumption that individual types are firm-dependent, so as to avoid the "over-representation problem". (On the other hand, it is not difficult to allow over-representation in the model.)

Following the reasoning in Appendix 1 shows that Assumption 3′ (together with Assumption 1) is also sufficient for the existence of a high-reputation equilibrium.

(*b*) *Low-reputation firm.* In a low-reputation firm only the honest workers produce high quality. In particular, because opportunists don't produce high quality, there is more firing and therefore *more turnover in a low-reputation firm than a high-reputation firm.* Consumers pay per unit:

$$p_L=\frac{\alpha}{\alpha+(\beta+\gamma)\tilde{Y}}H+\frac{(\beta+\gamma)\tilde{Y}}{\alpha+(\beta+\gamma)\tilde{Y}}L<p_H.$$

A necessary and sufficient condition for the existence of a low-reputation-firm equilibrium is:

Assumption 6′. $G/(1-\delta)\geqq\delta p_L\tilde{Z}$.

Assumption 6′ states the rent attached to working in a low-reputation firm is too small to dissuade the worker from shirking. Because Assumptions 3′ and 6′ are not mutually inconsistent, there may exist multiple steady states.

Hysteresis. The analysis of Section 3.3 suggests that reputation is a very valuable asset in the sense that it may be impossible for the firm to rebuild its reputation after having lost it. Things however depend on the existence of costs of firing the whole labour force. Suppose that a firm goes through a period of lax management in which the opportunists produce low quality and that, as in Section 3.3, these are locked into producing low quality in the future. If the cost of mass firing is large, so that the firm relies on quits and firings based on evidence to renew its labour force, we know from Section 3.3 that the firm will never be able to (re)build a reputation for high quality even long after the period of negligent management. Mass firing (implying firing without evidence and therefore firing even the honest workers) is the firm's only chance to recover, if this can be done at a reasonable cost.

Impact of increased competition. Introduce now an imperfect substitute, that yields net surplus $v>0$ to the consumers. So, competition can be conveniently parametrized by v. One can apply the previous analysis as long as p_H and p_L are replaced by p_H-v and p_L-v, respectively. Assumption 3′ then becomes

$$\frac{G}{1-\delta}\leqq\delta(p_H-v)\tilde{Z},$$

and is less likely to be satisfied, the more intense the competition. As competition heats up, the workers' rent shrinks and individual incentives are reduced. Firm reputation may be damaged.

Labour market externalities. As we already noted, we were able to take the proportions (α, β, γ) in the population of unemployed workers as given for our partial equilibrium analysis. It would be interesting to extend the analysis to study the labour market equilibrium. Indeed the proportions in the pool of unemployed workers depend on how many firms have a high reputation, and therefore fire only dishonest workers.

Proposition 4. (*a*) *The analysis can be extended to situations in which members can be excluded by the group, rather than by the trading partners.*

(*b*) *Increased product market competition may jeopardize the firm's provision of quality.*

5. CONCLUSION

We all belong to organizations, cultures, and racial groups. Our welfare and our incentives depend not only on our own reputation but also on that of the groups we are associated with. This paper has shown that individual reputations are determined by collective reputations, and vice versa. A member's incentive to maintain an individual reputation is stronger, the better the group's reputation. When discipline is sustained by the threat of exclusion from the group, low rents attached to being in a low-reputation group create low individual incentives to remain in that group and therefore perpetuate the group's bad reputation. When belonging to a group is an unalterable trait, poor collective behaviour in the past may make current good behaviour a low-yield individual investment and thus generate poor collective behaviour in the future.

Even more fascinating is the history-dependence of collective reputations. In our view, stereotypes are long-lasting because new members of a group at least partially inherit the collective reputation of their elders. We have seen that a one-time, non-recurrent shock on the behaviour of a population can prevent the population from ever returning to a satisfactory state even long after the members affected by the original shock are gone. We have provided a more general study of history-dependence; after episodes of bad behaviour, either the group is stuck in a bad-reputation steady state, or trust takes several periods to re-establish, after which the group's reputation returns progressively to the good-reputation level. In the context of corruption, we have analysed the determinants of the number of campaigns against bad behaviour needed to extract the group from the bad-reputation steady-state in the former case, and we have also seen that amnesties may be Pareto-improving.

The genesis of collective reputations is a complex phenomenon. The modest object of this paper has been to shed light on some of its facets. We hope that the topic will soon receive from economic theorists the attention it deserves.

APPENDIX 1. INCENTIVES TO CHEAT IN A LOW-CORRUPTION STEADY STATE (SECTION 3.2)

Let V_k denote an agent's expected present discounted value of present and future payoffs when the agent has cheated k times in the past. These are "continuation valuations". An agent who has cheated k times in the past will cheat again only if

$$G+\delta V_{k+1}\geqq\delta V_k. \tag{A.1}$$

Suppose that the agent finds it optimal to cheat when he has cheated k times, and not to cheat when he has cheated $(k+1)$ times. Then

$$V_k=(1-x_k)B+x_kb+G+\delta V_{k+1}\geqq(1-x_k)B+x_kb+\delta V_k, \tag{A.2}$$

and

$$V_{k+1}=(1-x_{k+1})B+x_{k+1}b+\delta V_{k+1}. \tag{A.3}$$

(A.2) and (A.3) yield

$$G\geqq\delta(x_{k+1}-x_k)(B-b)/(1-\delta). \tag{A.4}$$

On the other hand, the agent prefers stopping to cheating with record $(k+1)$ to cheating once more and then stopping. So

$$G+\delta\tilde{V}_{k+2}\leqq\delta V_{k+1}, \tag{A.5}$$

where

$$\tilde{V}_{k+2}=(1-x_{k+2})B+x_{k+2}b+\delta\tilde{V}_{k+2} \tag{A.6}$$

$$\tilde{V}_{k+2}\leqq V_{k+2}. \tag{A.7}$$

Equations A.5 and A.6 yield

$$G\leqq\delta(x_{k+2}-x_{k+1})(B-b)/(1-\delta). \tag{A.8}$$

Inequalities A.4 and A.8 are inconsistent with Assumption 1. So if it is optimal to cheat with record k, it is also optimal to cheat with any record $k'>k$.

APPENDIX 2. DIRECT EXCLUSION: THE CASE OF EXTORTION

This appendix tests the robustness of the analysis in Section 3 to the case of extortion.

Suppose a foreign company wants to do business in a country and wonders whether it should bribe low- or high-level government employees to process goods through customs, issue work permits for company personnel or building permits for plants, grant a government contract or provide police protection. It has been well documented by Jacoby *et al.* (1977) and many others that this is unfortunately one of the first questions business persons confront. Leaving aside any moral issue, we ask whether there can exist multiple steady states with different levels of extortion. This is indeed the case. In a non-corrupt steady state, government officials do what they are meant to do even if they are offered no bribe, firms can get away with offering no bribe, and government officials have no incentive to give them trouble given that they will not be offered bribes in the future. In a corrupt steady state, firms attach a low probability of being able to conduct business without giving bribes, and they do offer bribes. Government officials are reluctant to do their job in the absence of a bribe because this might reveal their "softness" to future bribers.

The model shares a number of similarities with the model of trust, and will purposedly share some of its notation. As before, the model is one of matching. In each period, the agent (the government official, the bribee) is matched with a new principal (the firm, the briber). The timing within the period is as follows: First, the firm decides whether or not to offer a bribe to the official. For simplicity, we let B denote the size of the bribe.[18] The firm gains $V>B$ if the agent provides the service. Second, the agent decides whether to provide the service. There are three types of agents: "honest", in proportion α, "corrupt", in proportion β, and "opportunist", in proportion γ, where $\alpha+\beta+\gamma=1$. The proportions are the same for each cohort. Honest officials always provide the service. Corrupt officials never provide the service unless they receive bribe B. Opportunists, when they are offered no bribe, trade off a short-term cost $c>0$ of not providing the service and the long-term loss of reputation for being tough. They provide the service if offered a bribe. One can think of c as coming either from scruples associated with not doing one's job or from a probability of being caught and punished. The probability of survival λ and the relevant discount factor δ are defined as before.

We again posit imperfect information about the agent. The principal has probability x_k of finding out that the agent has been weak at least once in the past, when the agent has in fact been weak k times, where "being weak" or "giving in" means that the agent provides the service to a principal who does not offer the bribe[19]. The x_k sequence satisfies Assumption 1.

(*a*) *No-extortion steady state.* In a no-extortion steady state, the firms never offer a bribe even when they don't know of any occurrence in which the government official gave in. In such a steady state, opportunists always give in, since they will never be offered a bribe in the future. Is it irrational for a firm not to offer a bribe when it does not know whether it faces an honest or opportunistic agent? Let

$$Y\equiv(1-\lambda)[1+\lambda(1-x_1)+\lambda^2(1-x_2)+\cdots]$$

denote the average probability over the population of opportunists and honest agents that an opportunist or honest agent is not observed to have been weak in the past. The firm does not offer a bribe to an official whose

18. See Carrillo (1995a) for an endogeneization of the size of bribes in a career-concern model.

19. It would be worth investigating alternative assumptions on individual reputations. This restrictive, but simple assumption allows us to make direct use of the preceding analysis.

type it does not know if and only if the following assumption holds:

Assumption 8.

$$B > \frac{\beta}{\beta + (\alpha + \gamma)Y} V.$$

Assumption 8 states that the size of the bribe exceeds the conditional probability that the official is corrupt times the value of the service to the firm. Note that the no-extortion equilibrium exists only if the firm is not perfectly informed about the agent's track record (if λ and the x's are close to 1, Y is close to 0 and Assumption 8 is violated).

(*b*) *Extortion steady state.* Suppose now that the firms offer a bribe to those agents who are not known to have given in, and no bribe to those who are known to have given in; and that opportunists do not give in (unless they have already given in at least $k^* > 1$ times, in which case they give in) when offered no bribe.

If the firm knows that the official has given in at least once when offered no bribe, this official must be honest and therefore it is optimal for the firm not to offer a bribe. In contrast, if the firm does not know that the official has given in the past, the firm optimally offers a bribe if the probability that the service will not be provided in the absence of a bribe times the value of the service exceeds the bribe:

Assumption 9.

$$B < \frac{\beta + \gamma}{\beta + \gamma + \alpha Y} V.$$

In an extortion steady state, it must also be the case that when offered no bribe an opportunist does not want to give in. Let z denote the present discounted expected value of future bribes received by an official who gives in every time that he is not offered a bribe[20]. A necessary condition for the existence of the extortion equilibrium is that

Assumption 10.

$$\delta\left(\frac{B}{1-\delta} - z\right) > c.$$

Conversely, the extortion steady state exists if Assumptions 9 and 10 hold (the proof is almost identical to that in Appendix 1.). Note that it can exist only if the principals' information is not too imprecise (if the x's are close to 0, z is close to $B/(1-\delta)$ and Assumption 10 is violated).

We thus conclude that under Assumptions 8, 9 and 10, the extortion and no-extortion steady states co-exist. The formal analysis is almost identical to that of Section 3.2. Yet the economics of trust (Section 3) and extortion (this appendix) differ in a few respects. In the extortion context, individuals want to build a reputation for the behaviour that society tries to eradicate. In the trust context, they want to build a reputation for honesty. This distinction will have implications when adapting the design of anti-corruption policies to the targeted form of corruption. A careful analysis of this conjecture falls outside the scope of this exploratory paper.

Acknowledgements. This paper was prepared under a cooperative agreement between the Institute for Policy Reform (IPR) and Agency for International Development (USAID), Cooperative Agreement No. PDC#0095-A-00-1126-00. Views expressed in this paper are those of the author and not necessarily those of IPR or USAID. The author is grateful to Juan Carrillo, Jacques Crémer, Drew Fudenberg, David Martimort, Patrick Rey, Paul Seabright, Lars Stole, Barry Weingast, and an anonymous referee for helpful comments.

20. z is given by the following recursive equation: Let V_k denote the valuation of an opportunist who has given in k times in the past, and gives in whenever he has not been offered a bribe:

$$V_k = x_k \delta V_{k+1} + (1 - x_k)(B + \delta V_k).$$

Let

$$x = \lim_{k \to \infty} x_k \quad \text{and} \quad V_\infty = (1-x)B/(1-\delta).$$

Then $z \equiv V_1$.

REFERENCES

ACEMOGLU, D. (1994), "A Dynamic Model of Collusion" (mimeo, MIT).
AKERLOF, G. (1976), "The Economics of Caste and of the Rat Race and other Woeful Tales", *Quarterly Journal of Economics*, **91**, 599–617.
ANDVIG, J. C. and MOENE, K.O. (1990), "How Corruption may Corrupt", *Journal of Economic Behavior and Organization*, **13**, 63–76.
ARROW, K. (1973), "The Theory of Discrimination", in O. Ashenfelter and A. Rees. (eds.), *Discrimination in Labor Markets* (Princeton: Princeton University Press).
BANERJEE, A. (1994), "A Theory of Misgovernance" (mimeo, MIT).
BECKER, G. (1968), "Crime and Punishment: An Economic Approach", *Journal of Political Economy*, **76**, 169–217.
BÉNABOU, R. and GERTNER, R. (1993), "Search with Learning from Prices: Does Increased Inflationary Uncertainty Lead to Higher Markups?", *Review of Economic Studies*, **60**, 69–94.
BESLEY, T. and KANDORI, M. (1992), "Reputation as a Public Good" (mimeo, Princeton University).
CADOT, O. (1987), "Corruption as a Gamble", *Journal of Public Economics*, **33**, 223–244.
CARRILLO, J. (1995a), "Grafts, Bribes, and the Practice of Corruption" (mimeo, GREMAQ, Toulouse).
CARRILLO, J. (1995b), "Corruption in Hierarchies" (mimeo, GREMAQ, Toulouse).
COATE, S. and LOURY, G. (1991), "Will Affirmative Action Policies Eliminate Negative Stereotypes?" (mimeo, University of Pennsylvania and Harvard University).
COLE, H., MAILATH, G. and POSTLEWAITE, A. (1992), "Social Norms, Savings Behavior and Growth", *Journal of Political Economy*, **100**, 1092–1125.
CRÉMER, J. (1986), "Cooperation in Ongoing Organizations", *Quarterly Journal of Economics*, **101**, 33–49.
DIAMOND, D. (1991), "Monitoring and Reputation: The Choice between Bank Loans and Directly Placed Debt", *Journal of Political Economy*, **99**, 689–721.
ELSTER, J. (1989) "Social Norms and Economic Theory", *Journal of Economic Perspectives*, **3**, 99–118.
GOULD, D. (1980) *Bureaucratic Corruption and Underdevelopment in the Third World: The Case of Zaire* (New York: Pergamon Press).
HAGER, M. (1973), "Bureaucratic Corruption in India: Legal Control of Maladministration", *Comparative Political Studies*, **6**, 197–219.
JACOBY, N., NEHEMKIS, P. and EELLS, R. (1977) *Bribery and Extortion in World Business: A Study of Corporate Political Payments Abroad* (New York: Macmillan Publishing Co.).
KANDORI, M. (1992), "Social Norms and Community Enforcement", *Review of Economic Studies*, **59**, 61–80.
KLEIN, B. and LEFFLER, K. (1981), "The Role of Market Forces in Assuring Contractual Performance", *Journal of Political Economy*, **81**, 615–641.
KLITGAARD, R. (1986) *Elitism and Meritocracy in Developing Countries: Selection Policies for Higher Education* (Baltimore: Johns Hopkins University Press).
KLITGAARD, R. (1988) *Controlling Corruption* (Berkeley: University of California Press).
KLITGAARD, R. (1991) *Adjusting to Reality* (San Francisco: ICS Press).
KOTLIKOFF, L., PERSSON, T. and SVENSSON, L. (1988), "Social Contracts as Assets: A Possible Solution to the Time-Consistency Problem", *American Economic Review*, **78**, 662–677.
KREMER, M. (1993), "The O-Ring Theory of Economic Development", *Quarterly Journal of Economics*, **108**, 551–575.
KREPS, D. (1990), "Corporate Culture and Economic Theory", in J. Alt and K. Shepsle (eds.), *Perspectives on Positive Political Economy*, (Cambridge: Cambridge University Press), 90–143.
KREPS, D., MILGROM, P., ROBERTS, J. and WILSON, R. (1982) "Reputation and Imperfect Information", *Journal of Economic Theory*, **27**, 253–279.
KRUEGER, A. (1974), "The Political Economy of the Rent-Seeking Society", *American Economic Review*, **64**, 291–303.
LUI, F. (1986), "A Dynamic Model of Corruption Deterrence", *Journal of Public Economics*, **31**, 215–236.
LUNDBERG, S. and STARTZ, R. (1983), "Private Discrimination and Social Intervention in Competitive Labor Markets", *American Economic Review*, **73**, 340–347.
MEYER, M. and VICKERS, J. (1994), "Performance Comparisons and Dynamic Incentives" (mimeo, Oxford University).
MILGROM, P. and OSTER, S. (1987), "Job Discrimination, Market Forces and the Invisibility Hypothesis", *Quarterly Journal of Economics*, **102**, 453–476.
MYRDAL, G. (1970), "Corruption as a Hindrance to Modernization in South Asia", in A.Heidenheimer (ed.), *Political Corruption: Readings in Comparative Analysis* (New York: Holt, Rinehart and Winston).
NOONAN, J. (1984) *Bribes*, (New York: Macmillan).
PHELPS, E. (1972), "The Statistical Theory of Racism and Sexism", *American Economic Review*, **62**, 659–661.
ROSE-AKERMAN, S. (1978) *Corruption: A Study in Political Economy* (New York: Academic Press).
ROSEN, A. (1993), "An Equilibrium Search-Matching Model of Discrimination" (mimeo, Trade Union Institute for Economic Research, Stockholm).
SAH, R. (1991), "Social Osmosis and Patterns of Crime", *Journal of Political Economy*, **99**, 1272–1295.
SHAPIRO, C. (1983), "Premiums for High-Quality Products as Rents to Reputation", *Quarterly Journal of Economics*, **98**, 659–680.

SHLEIFER, A. and VISHNY, R. (1993), "Corruption", *Quarterly Journal of Economics*, **106**, 289–295.
STRAND, J. (1990), "Bureaucratic Corruption in Government Contract Procurement: A Theoretical Model" (mimeo, University of Oslo).
SEABRIGHT, P. (1992), "Is Cooperation Habit-Forming?" (mimeo, Churchill College, Cambridge).
SUGDEN, R. (1989), "Spontaneous Order", *Journal of Economic Perspectives*, **3**, 85–98.
THEOBALD, R. (1990) *Corruption, Development and Underdevelopment* (Durham: Duke University Press).
TOPKIS, D. (1979), "Equilibrium Points in Nonzero-Sum n-Person Submodular Games", *SIAM Journal of Control and Optimization*, **17**, 773–787.
ULLMAN-MARGALIT, E. (1977) *The Emergence of Norms* (Oxford: Oxford University Press).
YOUNG, P. (1993), "The Evolution of Conventions", *Econometrica*, **61**, 57–84.