EIEF WORKING PAPER SERIES

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 24/04

March 2024

When is Trust Robust?

By

Luca Anderlini
(Georgetown University & University of Naples Federico II)

Larry Samuelson
(Yale University)

Daniele Terlizzese
(EIEF)

When is Trust Robust?*

Luca Anderlini
Georgetown University and University of Naples Federico II

Larry Samuelson
Yale University

Daniele Terlizzese
EIEF

March 2024

Abstract. We examine an economy in which interactions are more productive if agents can trust others to refrain from cheating. Some agents are scoundrels, who always cheat, while others cheat only if the cost of cheating, a decreasing function of the proportion of cheaters, is sufficiently low. The economy exhibits multiple equilibria. As the proportion of scoundrels in the economy declines, the high-trust equilibrium can be disrupted by arbitrarily small perturbations or infusions of low-trust agents, while the low-trust equilibrium becomes impervious to perturbations and infusions of high-trust agents. The resilience of trust may thus hinge upon the prevalence of scoundrels.

JEL Classification: C72, C79, D02, D80.
Keywords: Trust, Robustness, Fragility, Assimilation, Disruption.
Correspondence: Larry Samuelson — Larry.Samuelson@yale.edu

*Part of this research was done while Luca Anderlini and Larry Samuelson were visiting the EIEF in Rome. They are both grateful to EIEF for its hospitality.

1. Introduction

Trust is important. For example, Levitsky and Ziblatt (2019) argue that democracies require two ingredients to function effectively, namely that competing parties accept one another as legitimate rivals and that they trust one another to exercise restraint in exploiting their institutional advantages. Trust can also be fragile—it can be laborious to build, easy to destroy, and difficult to rebuild (Slovic (1993, 1999)). As Mr. Darcy explains to Elisabeth Bennett in Jane Austen's *Pride and Prejudice*, "My good opinion once lost is lost forever."

This paper examines the conditions under which trust is more or less robust. We say that an equilibrium characterized by a high level of trust is fragile if a relatively small shock, in the form of an infusion of agents acclimated to an alternative, low-trust equilibrium, disrupts the high-trust equilibrium and prompts convergence to the low-trust equilibrium. We say the high-trust equilibrium is robust if it is capable of withstanding such shocks, assimilating the new arrivals and converging back to high trust.

Specifically, we study an economy in which interactions are more productive if agents on one side trust those on the other side to refrain from cheating, and agents on the other side indeed do not cheat. Some agents are "scoundrels", who always cheat, while the remaining agents are responsive, cheating if and only if the combination of a private and social cost of cheating is sufficiently low. The social cost of cheating results from opprobrium heaped on a cheater who is believed to be a scoundrel. Bayesian updating then implies that the social cost of cheating is a convex and decreasing function of the prevalence of cheating by responsive agents in the economy.

Because the social cost of cheating depends on the fraction of people who cheat, multiple equilibria can arise. If the fraction of scoundrels is sufficiently large, there is a unique equilibrium in which no responsive agents cheat and trust is relatively high. If the fraction of scoundrels is smaller than a certain threshold, a low-trust and a high-trust equilibrium coexist (along with an unstable equilibrium exhibiting an intermediate level of trust).[1] The basins of attraction of the two stable equilibria depend on the fraction of scoundrels. The smaller is this fraction, the more convex is the social cost of cheating. Hence, when scoundrels are scarce and cheating is low, a small increase in cheating sharply reduces its social cost, validating the increase and potentially catapulting a high-trust, low-cheating economy outside the basin of attraction of the high-trust equilibrium. The high-trust equilibrium thus

[1] We will often term the high-trust equilibrium as good and the low-trust one as bad.

exhibits less cheating when there are fewer scoundrels, but sits more precariously within a smaller basin of attraction. In contrast, fewer scoundrels *increase* the prevalence of cheating in the low-trust equilibrium and expand its basin of attraction.

Moving beyond the analysis of the stability of the equilibria, we examine the implications of introducing into an economy, characterized by either the high-trust or low-trust equilibrium, a small mass of agents with behavior characteristic of the other equilibrium. If the fraction of scoundrels is sufficiently small, then an arbitrarily small infusion of agents accustomed to the low-trust equilibrium can disrupt the high-trust equilibrium, while a large infusion of agents accustomed to the high-trust equilibrium is required to disrupt the low-trust equilibrium. The intuition for this asymmetry again lies in the convexity of the social cost of cheating. If the fraction of scoundrels is small enough, agents in the high-trust equilibrium face a very steep portion of the cost function; when they observe more cheating than expected (due to the infusion of agents acclimated to the low-trust equilibrium), the perceived cost falls sharply, inducing the formally high-trust agents to cheat more, eventually pushing the society to the low-trust equilibrium. Conversely, agents in the low-trust equilibrium are in a much flatter portion of the cost curve. Hence, observing less cheating than expected, their perceived social cost increases very little and their cheating changes very little. This allows the low-trust equilibrium to survive.

A similar intuition explains why, even when the basins of attraction of the two equilibria are identical, the low-trust equilibrium is more robust than the high-trust one: the fraction of agents accustomed to the low-trust equilibrium required to disrupt the high-trust equilibrium is considerably smaller than the fraction of accustomed to the high-trust equilibrium required to disrupt the low-trust one.

There is thus a sense in which, other things equal, scoundrels serve a useful purpose. A society in which scoundrels are rare, and in which therefore the social cost of cheating is sharply convex in the equilibrium fraction of cheaters, is one in which a good social norm can be easily disrupted, while a bad social norm is much more resilient. The best outcome is to have few scoundrels and coordinate on the good equilibrium, but this is fragile and risky. Tolerating some scoundrels may be a price worth paying for rendering the good equilibrium more robust.

There is a growing recognition of the importance of social capital in shaping the performance of an economy. Arrow (1974) argued that even the simplest of economic transactions

calls for a foundation of trust. A large literature, energized by Putnam (2000), with Jackson (2020) providing a recent point of entry, now explores the link between social capital and economic development. We view the trust in our model as a stylized representation of social capital, and view the fragility results as a cautionary tale that social capital can be not only hard to build but easy to dissipate.

Section 2 presents the model. Section 3 characterizes the stability of the various equilibria and explains how this depends on the proportion of scoundrels. Section 4 examines the robustness of the the good and bad equilibria to infusions of agents accustomed in each case to the other equilibrium. Section 5 interprets and discusses the results. To streamline, all proofs are gathered in an Appendix. Any item with a number prefixed by "A" is to be found in the Appendix.

2. The Model

2.1. The Game

The game is adapted from Anderlini and Terlizzese (2017). We view the game as capturing the spirit of the trust game of Berg, Dickhaut, and McCabe (1995), with the minimum modification required to ensure that an endogenously determined, positive level of equilibrium trust can emerge.

In each period, the members of a continuum of agents are matched into pairs to play a game. Each time they are drawn to play the game, each agent is equiprobably assigned to be either a proposer or a responder. The proposer in the game first chooses a quantity $x \in \mathbb{R}_+$. The responder then chooses either to cheat or not cheat. If the responder does *not* cheat, then the proposer and responder each receive a payoff of x. If the responder cheats, then the proposer receives 0 and the responder receives $2x$ minus the cost of cheating. We can interpret x as a proposed scale at which to operate a joint project. As the scale increases, so do the payoffs of both agents if they indeed share the proceedings, but so does the payoff to the responder from cheating and thereby appropriating the entire payoff (minus the cost of cheating).

Proportion q of the responders are *scoundrels*, who cheat at every opportunity, and proportion $1 - q$ are *responsive*. In each interaction, a responsive responder has a type z drawn from the uniform distribution on $[0, 1]$. The payoff of a responsive responder who cheats is

given by

$$2x - c(z, s),$$

where s is the proportion of responsive responders who cheat and $c(z, s)$ is the cost of cheating.[2]

We model the cost of cheating as the sum of two components. The first component is simply z, the agent's type. We interpret z as the intrinsic cost of cheating, with a low value of z identifying a person who suffers little intrinsic cost from cheating. We will in turn typically interpret a low intrinsic cost as reflecting a high need to cheat or high benefit from cheating. Suppose, for example, that cheating takes the form of cutting ahead of others in traffic. We might think of a low z as identifying a person who is on the way to the hospital, and so has an urgent need for haste. A medium value of z might identify a person who is late for work, and so has a moderate need for haste. A high value of z is a person not pressed for time.

The second component, the social cost of cheating, takes the form of public disapproval, ostracism, or other forms of sanction. We believe that people dislike cheating that is simply without a reason. We capture this by assuming that the social sanctioning becomes more severe the more likely it is that the cheater is a scoundrel. Hence, the traffic menace is sanctioned more gently the more likely it is that she is an imminently expectant mother headed for the hospital, and is sanctioned more severely the more likely it is that he is a scoundrel who routinely flaunts traffic conventions. Recalling that s is the proportion of responsive responders who cheat, the posterior probability that someone observed cheating is a scoundrel is

$$\frac{q}{q + (1 - q)s}.$$

We then take the social cost of cheating to be

$$f(s) = \theta \frac{q}{q + (1 - q)s}, \tag{1}$$

where θ is a parameter that allows us to tune the relative importance of the idiosyncratic

[2]The wastefulness of cheating is captured directly by the cost of cheating, and indirectly by its effect on the equilibrium value of x. We could introduce an an additional shrinking factor, so that the payoff to a cheating responder is less than $2x - c(z, s)$, without relevant changes in the results of the analysis.

and social components of the cost of cheating. The total cost of cheating is then

$$c(z, s) = z + f(s).$$

If s is small, then a cheater is likely to be a scoundrel, and cheating will be punished heavily. If s is large, then it is relatively unlikely that a cheater is a scoundrel, and cheating will be only lightly punished. We can thus think of s as representing a social norm, determining whether cheating is rare and heavily sanctioned or whether cheating is common and tolerated.

We note that the function $f : [0, 1] \to \mathbb{R}_+$ is a decreasing, convex function with $f(0) = \theta$ and $f(1) = \theta q$. Therefore the social component of the cost of cheating is maximal when no responsive responder cheats, falls quickly as soon as a few of them cheat, and keeps falling, but at a decreasing rate, as more and more of them cheat. The fewer the scoundrels, i.e. the smaller is q, the greater is the convexity of f. In particular, the steeper is f near 0.[3] Figure 1 illustrates.

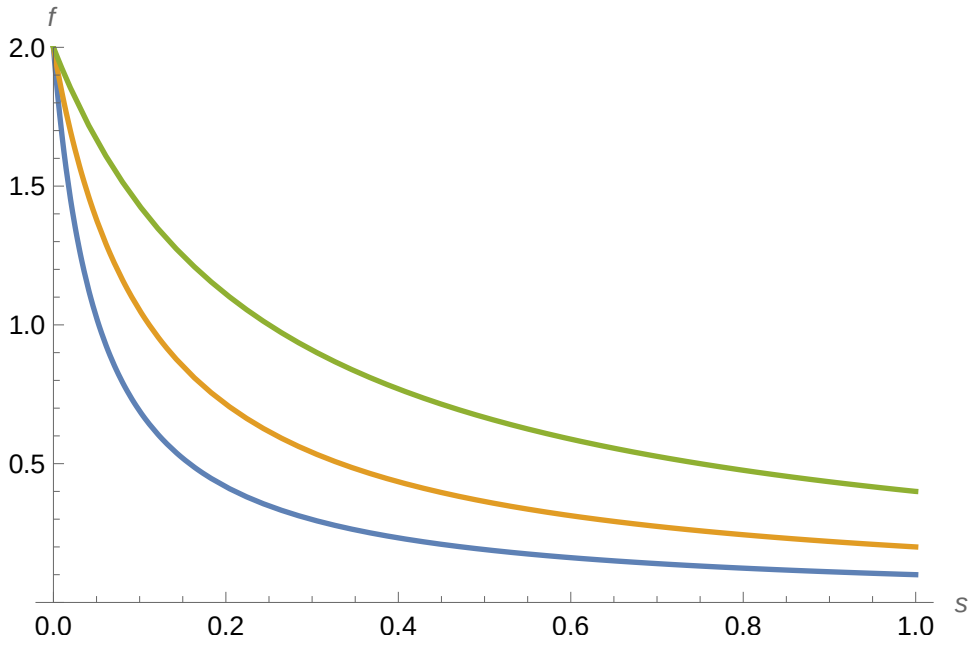

Figure 1: Illustration of the social cost of cheating $f(s)$ as a function of the proportion s of responsive responders who cheat, for $\theta = 2$ and the proportion of scoundrels (top to bottom) $q = 0.2, 0.1, 0.05$.

[3]For instance, when s increases from 0 to $\frac{q}{1-q}$, the posterior probability that a cheater is a scoundrel falls from 1 to $\frac{1}{2}$. Hence, the increase in s that generate such a decrease becomes smaller as q decreases.

As q approaches zero, the function f converges (but not uniformly) to

$$\begin{aligned} f(0) &= \theta \\ f(s) &= 0 \quad \text{for} \quad s > 0. \end{aligned}$$

We will be especially interested in cases in which q is small, so that the proportion of scoundrels is low.

2.2. Equilibrium

A responsive responder takes the proportion s of responsive responders who cheat as given, and when facing an offer x, will cheat if her cost of cheating z falls short of a cutoff $\zeta(x, s)$ and will not cheat if $z \geq \zeta(x, s)$. The cutoff $\zeta(x, s)$ equalizes the payoffs of cheating and not cheating, and hence when interior solves

$$2x - [\zeta(x, s) + f(s))] = x.$$

In general, we have

$$\zeta(x, s) = \max\{0, x - f(s)\}. \tag{2}$$

The maximum reflects the possibility of a corner solution in which the responsive responder does not cheat even if $z = 0$. In principle we could also have a corner solution in which the responsive responders cheat even if $z = 1$. In the next two paragraphs we will see that this does not arise.

A proposer takes the proportion s of responsive responders who cheat as given and chooses a value x to maximize the payoff

$$(1 - ((1 - q)\zeta(x, s) + q))x,$$

where $1 - ((1 - q)\zeta(x, s) + q)$ is the (overall) probability that the current responder does *not* cheat. Using (2), we can write the maximization problem of a proposer as

$$\max_{x \geq 0}(1 - ((1 - q)\max\{0, x - f(s)\} + q))x.$$

The proposer will never set $x \geq f(s) + 1$. Doing so would induce all responsive responders

to cheat and hence would yield a proposer payoff of 0, while the proposer can ensure a positive payoff by setting $x < f(s)+1$. Equivalently, we will never have a corner solution in which all responsive responders cheat. The proposer will similarly never set $x < f(s)$, since doing so would ensure that no responsive responders would cheat, *and* that the proposer could increase the offer without inducing additional cheating.

We can thus restrict attention to offers $x \in [f(s), f(s)+1)$. The proposer's objective is then to solve

$$\max_{x\in[f(s),f(s)+1)} (1-(1-q)(x-f(s))-q)x.$$

The first-order condition if $x > f(s)$ is

$$1+f(s)-2x=0 \;\Leftrightarrow\; x=\frac{1}{2}+\frac{1}{2}f(s).$$

This is the relevant solution as long as $x > f(s)$, i.e. as long as $\frac{1}{2}+\frac{1}{2}f(s) > f(s)$, or, equivalently, as long as $f(s) < 1$. Let s^* be such that $f(s^*) = 1$. If $s < s^*$, the proposer will choose the highest value of x consistent with no responsive responders cheating, namely $x = f(s)$.

We thus have

$$x=\begin{cases} \frac{1}{2} + \frac{1}{2}f(s) & s \geq s^* \\ \\ f(s) & s \leq s^*, \end{cases}$$

Using (1), we can solve $f(s^*) = 1$ to obtain

$$s^* = \frac{q(\theta-1)}{1-q}. \tag{3}$$

When $s \leq s^*$, cheating is sufficiently costly that the proposer chooses the highest offer consistent with no responsive responders cheating. For $s > s^*$, the proposer will choose an interior solution in which some responsive responders cheat.

The equilibrium condition is that the proportion s of cheating by responsive responders must induce a proposer offer x that in turn causes the cutoff $\zeta(x, s)$ to match s. We thus

have three conditions which jointly determine the equilibrium values of s, ζ and x:

$$s = \zeta(x, s) \tag{4}$$

$$\zeta(x, s) = \max\{0, x - f(s)\} \tag{5}$$

$$x = \begin{cases} \frac{1}{2} + \frac{1}{2} f(s) & s \geq s^* \\ f(s) & s \leq s^* \end{cases} \tag{6}$$

The final condition (6) can be rewritten as

$$x = \max\left\{ f(s), \frac{1}{2} + \frac{1}{2} f(s) \right\}.$$

2.3. Equilibrium Cheating

Our first result is that if the social cost of cheating is sufficiently low, then there is a unique equilibrium, which exhibits some cheating. The proof, contained Section A.1, is a straightforward calculation. The left panel of Figure 2 below illustrates this case.

Proposition 1*: If $\theta < 1$, there exists a unique equilibrium. In equilibrium, some responsive responders cheat.*

We are interested in the case of multiple equilibria. Accordingly, from this point on we assume that the social component of the cost of cheating is sufficiently important.

Assumption 1*: $\theta > 1$*

In this case, one corner-solution equilibrium configuration is

$$s = \zeta = 0, \qquad x = f(0) = \theta.$$

This is a high-trust, no cheating equilibrium, featuring a relatively large offer x and no cheating on the part of responsive responders. Given the assumption that $\theta > 1$, this equilibrium always exists. We refer to this as the good equilibrium, and denote the proportion of responsive responders who cheat in this equilibrium by $s_g = 0$.

If the social cost of cheating $f(s)$ decreases sufficiently rapidly in s, then we have two additional, interior solutions. Each of these must satisfy $s \geq s^*$, and hence must satisfy the interior versions of (4)–(6), or $\zeta(x,s) = x - f(s)$ and $x = \frac{1}{2} + \frac{1}{2}f(s)$. We can reduce (4)–(6) to a single equation in s, given by

$$\frac{1}{2} + \frac{1}{2}f(s) = s + f(s),$$

which in turn can be rearranged to read

$$f(s) = 1 - 2s. \tag{7}$$

Given the specification of $f(s)$ as in (1) this is a quadratic equation, whose solutions are

$$s_b = \frac{1 - 3q + \sqrt{(q+1)^2 - 8\theta q(1-q)}}{4(1-q)} \tag{8}$$

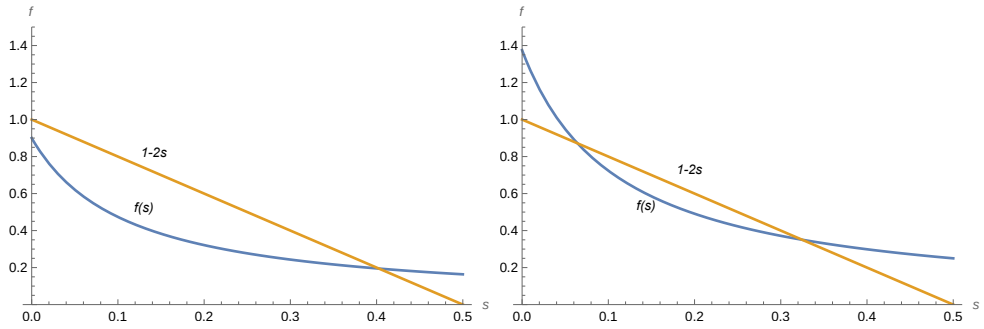

Figure 2: Illustration of equilibria. In the left figure, $\theta = 0.9$ (the social cost of cheating is relatively low), and there is a single, interior equilibrium. In the right picture, $\theta = 1.375$ and there are relatively few scoundrels ($q = 0.1$ in both panels), giving rise to three equilibria. The high-trust equilibrium $s = 0$ corresponds to the intersection of $f(s)$ with the vertical axis, while the intermediate and low-trust equilibria are determined by the two interior intersections. As the proportion of scoundrels increases, the function f shifts upward, pushing the intermediate and low-trust equilibria closer together, until a point is reached at which $q = \hat{q}(\theta)$ and these equilibria first coincide and then disappear, leaving only the high-trust equilibrium.

and

$$s_u = \frac{1 - 3q - \sqrt{(q+1)^2 - 8\theta q(1-q)}}{4(1-q)}. \tag{9}$$

We thus have a low-trust equilibrium in which proportion s_b of responsive responders cheat, and an intermediate equilibrium in which proportion s_u of responsive responders cheat. We refer to these as the bad equilibrium and the unstable (for reasons made clear in Section 3) equilibrium, respectively.

The bad and unstable equilibria exist if the expression under the square root in (8)—(9) is positive. This is true if there are not too many scoundrels, with the upper threshold on the fraction of scoundrels given by

$$\hat{q}(\theta) = \frac{4\theta - 1 - 4\sqrt{\theta(\theta-1)}}{1 + 8\theta}. \tag{10}$$

If $q < \hat{q}(\theta)$, we have $0 = s_g < s_u < s_b$. If $q > \hat{q}(\theta)$, then no cheating is the only solution.[4]

We summarize with the following proposition, illustrated in Figure 2:

Proposition 2: *[2.1] The good equilibrium is the unique equilibrium if $q > \hat{q}(\theta)$, where the function $\hat{q}(\theta) : [1, \infty) \to [0, 1]$ is decreasing, and*

$$\hat{q}(1) = \frac{1}{3} \qquad \text{and} \qquad \lim_{\theta \to \infty} \hat{q}(\theta) = 0.$$

[2.2] If $q < \hat{q}(\theta)$, then in addition to the high-trust, no cheating (good) equilibrium, there is a low trust, high cheating (bad) equilibrium in which a proportion s_b of responsive responders cheat, and an intermediate (unstable) equilibrium in which a proportion s_u of responsive responders cheat.[5]

[2.3] The offers made by proposers are the highest and cheating is the lowest in the good equilibrium, while offers are the lowest and cheating the most prevalent in the bad equilibrium.

[4]The discriminant would also be positive if q were larger than the larger solution of the quadratic. In this case, however, both s_b and s_u would be negative. Therefore, only the smaller solution of the quadratic is relevant. For the boundary case of $q = \hat{q}$, the solutions s_b and s_u exist and coincide.

[5]There are two boundary cases. When $q = \hat{q}(\theta)$, there exist only two equilibria, a stable equilibrium $s_g = 0$ and another equilibrium (intuitively, $s_u = s_b$) that is stable from above but not from below. When $q = 0$, there exist only two equilibria, a stable equilibrium s_b and an another, unstable equilibrium (intuitively, $s_g = 0 = s_u$).

3. Stability

We now investigate the resilience of the high trust equilibrium, in two steps. The first, examined in this section, asks about the stability of the good equilibrium in response to small perturbations. We consider a society that is in one of the three equilibria, and suppose that the beliefs of all the society's members are perturbed. Will the original equilibrium survive, or will the result be convergence to a different equilibrium?

We need to assume an adjustment dynamic. The one we postulate is explicitly belief-based, as it is geared to a *perceived* level of cheating, which may or may not be an equilibrium level of cheating. Given this perceived level, the best-response behavior of the responsive responders generates an *actual* level of cheating. The perceived level of cheating then adjusts, moving towards the actual level. When perceived and actual cheating coincide, the system is at a resting point which coincides with one of the equilibria characterized in Proposition 2. We think these belief-based dynamics are well suited to the task at hand if trust is viewed as a social phenomenon primarily concerning beliefs.

Suppose that every agent perceives the level of cheating among responsive responders to be s_P. Proposers choose the value of x that would maximize their payoff if s_P was the prevailing proportion of responsive responders who cheat, and responders similarly take s_P to be the prevailing proportion of responsive responders who cheat when making their decisions.

The perception s_P then gives rise to a realized proportion of cheating among responsive responders s that solves the following system of equations, constructed to duplicate the best responses (5) and (6) above,

$$\begin{aligned} s &= \max\{0, x - f(s_P)\} \\ x &= \max\left\{f(s_P), \frac{1}{2} + \frac{1}{2}f(s_P)\right\}, \end{aligned}$$

which we can rearrange to obtain

$$s = \begin{cases} 0 & s_P \leq s^* \\ \frac{1}{2} - \frac{1}{2}f(s_P) & s_P \geq s^* \end{cases}$$

The potentially erroneous perception s_P moves toward the induced realization s. The details

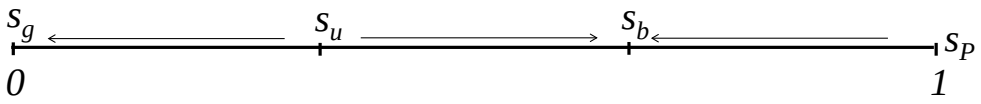

Figure 3: Illustration of the adjustment dynamics and basins of attraction for the candidate equilibrium proportion s_P of responsive responders who cheat.

of this dynamical system are not important, beyond noting that by construction it has three rest points, s_g, s_u and s_b.

As long as $s_P \leq s^*$, the induced realization of s is always 0, so s_P falls towards 0; if $s_P \in (s^*, s_u)$, we have that $s = \frac{1}{2} - \frac{1}{2}f(s_P) < s_P$, hence again s_P falls towards 0; if $s_P \in (s_u, s_b)$ the realized s is larger than s_P, which therefore increases towards s_b; finally, if $s_P > s_b$, the realized s is smaller than s_P, implying that s_P falls back towards s_b.

The good-equilibrium rest point s_g and the bad-equilibrium rest point s_b are therefore stable under this dynamic, while the intermediate rest point s_u is unstable. The latter divides the interval $[0, 1]$ of possible values of s_P into the basin of attraction $[0, s_u)$ of the lower rest point s_g and the basin of attraction $(s_u, 1]$ of the upper rest point s_b. Figure 3 illustrates.

The good equilibrium is "more stable" the larger is its basin of attraction, i.e. the larger is $s_u - s_g$, and the bad equilibrium is similarly "more stable" the larger is $s_b - s_u$. The comparative statics in the following proposition, which is proved formally in Section A.2, are an immediate consequence of (8)–(10),.

Proposition 3*: Assume that $q < \hat{q}(\theta)$, so that all three equilibria exist. As either q or θ increase,*

$$s_u - s_g \text{ increases , and } s_b - s_u \text{ decreases .}$$

Moreover, as q approaches zero, s_u also approaches zero and hence the basin of attraction of the good equilibrium s_g becomes arbitrarily small.

Hence, when q is small, there is relatively little cheating in the good equilibrium (since there are few scoundrels), but the good equilibrium is fragile, in the sense that it has a small basin of attraction, while cheating is relatively prevalent in the bad equilibrium. As q increases, so does the incidence of cheating in the good equilibrium, but the good equilibrium

has a larger basin of attraction, while the incidence of cheating in the bad equilibrium decreases. When q hits $\hat{q}(\theta)$, the unstable and bad equilibria coincide, and for larger values of q only the good equilibrium remains, albeit with more scoundrels. As θ increases, the proportion of scoundrels needed to eradicate the unstable and bad equilibria decreases.

We now see two respects in which it can be "good" to have more scoundrels. First, the more scoundrels there are, the "more likely" is the good equilibrium to be the unique equilibrium (more precisely, the smaller is the value of the social-cost-of-cheating parameter θ required to ensure the good equilibrium is unique). Second, when multiple equilibria exist, the good equilibrium is "more likely" the more scoundrels there are (more precisely, the larger is the basin of attraction of the good equilibrium).

Of course, scoundrels come at a cost—society has to put up with their cheating. The most fortunate society is one that contains few scoundrels, but manages to coordinate on and

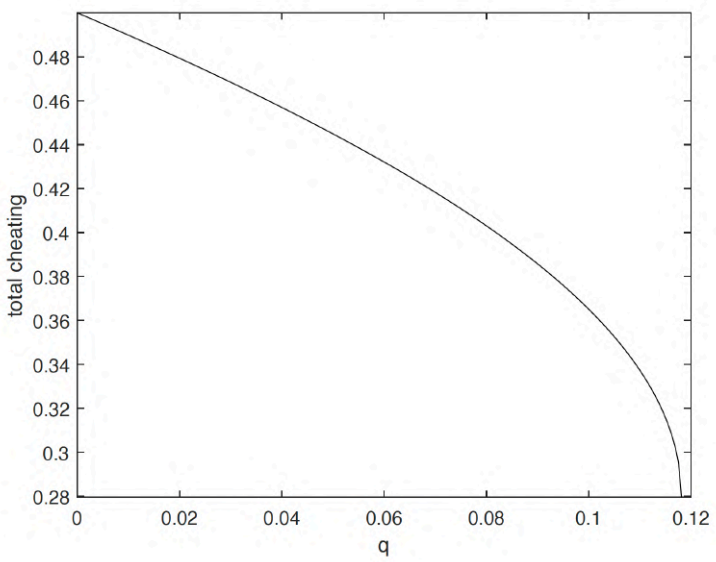

Figure 4: The total instance of cheating (vertical axis), by both scoundrels and responsive responders, in the bad equilibrium, as a function of the proportion of scoundrels q (horizontal axis); for the case $\theta = 1.5$. As the proportion of scoundrels increases, total cheating diminishes, until the proportion of scoundrels nears 0.12, at which point the bad and unstable equilibria vanish and only the good equilibrium remains. At this point, the incidence of cheating drops from about 0.28 to about 0.12.

preserve the good equilibrium, despite its fragility. A less fortunate society is that which still has few scoundrels, but is trapped at the bad equilibrium. The latter society might well welcome more scoundrels, both because the cheating of the additional scoundrels may be more than overwhelmed by inducing responsive responders to cheat less, and because eventually the number of scoundrels may increase to the point that only the good equilibrium remains.

Figure 4 shows the total incidence of cheating as a function of the proportion of scoundrels, for a society with $\theta = 1.5$ that is trapped at the bad equilibrium (when the latter exists). With no scoundrels, half of the agents cheat. As the proportion of scoundrels increases, total cheating diminishes, until the proportion of scoundrels nears 12 percent. Here, the unstable equilibrium and the bad equilibrium coincide, a fraction of about 18 percent of responsive responders cheat, and the total incidence of cheating, including the scoundrels, is about 28 percent. A further increase in the number of scoundrels then gives a discontinuous drop in the incidence of cheating, as society switches to the sole remaining equilibrium, the good one.

4. Robustness

We now turn to the second of our resilience questions. Beginning with a society that has settled on one of the stable equilibria, suppose the beliefs of a small fraction of the society's members are radically changed, interpreted as an infusion of agents accustomed to the other stable equilibrium. Will the original equilibrium survive, or will the infusion disrupt the standing equilibrium and prompt the society to converge to the other equilibrium?

4.1. Assimilation or Disruption?

We continue to suppose that the social cost of cheating is sufficiently high and there are sufficiently few scoundrels (i.e., $q < \hat{q}(\theta)$) so that we have three equilibria. What happens when some outsiders, characterized by the behavior and perceptions of a society in the bad equilibrium, merge into a society characterized by the good equilibrium? One can interpret this as a case in which a high-trust country (or organization, profession, culture, social group, and so on) is opened to immigration (or membership, or participation, and so on) from agents accustomed to the bad equilibrium. Will the newcomers be assimilated, and will their behavior converge to that of the good equilibrium? Or will the newcomers upset the social norm and cause everyone's behavior to settle on the bad equilibrium?

To address these questions, we suppose that a population initially in the good equilibrium is shocked by the injection of a fraction $\lambda \leq 1/2$ of outsiders whose perception and behavior

is taken from the bad equilibrium. Refer to the members of the original population, who are now in proportion $1-\lambda$, as insiders and give them subscript 1, and the invaders as outsiders, with subscript 0. The basic equations for our system are then:[6]

$$
\begin{aligned}
s &= (1-\lambda)^2\zeta_{11} + (1-\lambda)\lambda\zeta_{10} + \lambda(1-\lambda)\zeta_{01} + \lambda^2\zeta_{00} \\
\\
\zeta_{11} &= \min\{\max\{0, x_1 - f(s_1)\}, 1\} \\
\zeta_{10} &= \min\{\max\{0, x_1 - f(s_0)\}, 1\} \\
\zeta_{01} &= \min\{\max\{0, x_0 - f(s_1)\}, 1\} \\
\zeta_{00} &= \min\{\max\{0, x_0 - f(s_0)\}, 1\} \\
\\
x_1 &= \max\left\{f(s_1), \frac{1}{2} + \frac{1}{2}f(s_1)\right\} \\
\\
x_0 &= \max\left\{f(s_0), \frac{1}{2} + \frac{1}{2}f(s_0)\right\}
\end{aligned}
\tag{11}
$$

The variables s_1 and s_0 identify the proportion of cheating on the part of responsive responders perceived by insiders (s_1) and outsiders (s_0), and hence are the counterparts of s_P from Section 3. Insider and outsider proposers make offers that are optimal given their perceptions, and hence x_1 is the offer made by insiders and x_0 the offer made by outsiders. Responders make their decisions of whether to cheat based on the offer they face and their perception of the prevalence of cheating.

The proportion of responsive responders cheating in an interaction depends on both the identity of the proposer and the identity of the responsive responder, and so we have four cheating probabilities to keep track of. ζ_{10}, for example, is the proportion of cheating when an inside proposer interacts with an outside responsive responder. The value of any ζ_{ij} can never go above 1 in equilibrium, so that the outer minimum in the specification of the four realizations of ζ_{ij} is redundant in equilibrium, but ζ_{ij} can hit the upper bound of 1 in an out-of-equilibrium combination of a proposer who expects little cheating and hence makes a large offer with a responsive responder who expects a great deal of cheating and hence a low

[6]Implicit in this formulation is an assumption that insiders and outsiders mix randomly. We could alternatively imagine that outsiders are more likely to meet outsiders.

(social) cost of cheating.

The variable s identifies the realized incidence of cheating among responsive responders. Each of the four terms corresponds to one of the four possible matches, involving either an inside or outside proposer and an inside or outside responder, and gives the probability of such a match multiplied by the proportion of cheating in the match.

We assume that the perceptions s_1 of insiders and s_0 of outsiders both move toward the realization s, according to the dynamic system:[7]

$$\begin{aligned} \dot{s}_1(t) &= \delta\{s(t) - s_1(t)\} \\ \dot{s}_0(t) &= \delta\{s(t) - s_0(t)\}. \end{aligned} \tag{12}$$

4.2. Convergence

We first establish that perceptions converge. The intuition is the following. First, the dynamical system (11)—(12) implies that

$$s_0(t) = s_1(t) + e^{-\delta t}(s_0(0) - s_1(0)). \tag{13}$$

Therefore, the *difference* between $s_0(t)$ and $s_1(t)$ goes to zero as t grows, i.e., the perceptions s_1 and s_0 of insiders and outsiders approach each other. This is expected—both groups are adjusting their perceptions toward a common (though moving) level of realized cheating. Second, once these perceptions are sufficiently close, we essentially have the dynamic system described in Section 3 and pictured in Figure 3, which converges to one of the two stable equilibria. In Section A.3 we prove:

Lemma 1*: The dynamical system (11)-(12) converges, with* $\lim_{t\to\infty} s_1(t) = \lim_{t\to\infty} s_0(t)$ *and with both equal to either* s_g*,* s_u*, or* s_b*.*

4.3. The Fragility of the Good Equilibrium

When scoundrels are scarce, the good equilibrium is especially vulnerable to invasion. If there are sufficiently few scoundrels, an arbitrarily small fraction λ of invaders from the bad

[7] We specify the system directly in terms of the perceptions and realized cheating concerning *responsive* responders. What is observable, however, is the overall fraction of cheating, so an alternative specification of the dynamic system would envisage the perception of *total* cheating adjusting towards the realized *total* cheating: $\frac{d[(1-q)s_k(t)+q]}{dt} = \delta\{((1-q)s(t)+q) - ((1-q)s_k(t)+q)\}$, with $k = 0, 1$. Clearly, for any given q, this would give (12).

equilibrium is capable of disrupting the good equilibrium: eventually all the agents converge to the beliefs and behavior of the bad equilibrium. Section A.4 in the Appendix proves:

Proposition 4*: Consider the dynamic system (11)-(12) with the initial conditions* $s_1(0) = s_g = 0$ *and* $s_0(0) = s_b$ *(i.e. a system in which the insiders initially believe themselves to be in the good equilibrium and outsiders in the bad equilibrium). For any* $\lambda > 0$ *there exists a* $q^* > 0$ *such that, for any* $q \leq q^*$ *it will be the case that* $\lim_{t\to\infty} s_1(t) = \lim_{t\to\infty} s_0(t) = s_b$*, i.e. the system converges to the bad equilibrium.*

The intuition behind this result is that, as the proportion q of scoundrels decreases, the basin of attraction of the good equilibrium becomes smaller (cf. Proposition 3 and Figure 3). Even a small influx of agents whose behavior initially matches the one prevailing in the bad equilibrium then suffices to catapult the system into the basin of attraction of the bad equilibrium. The proof addresses the complications arising out of the fact that this intuition is gleaned from a system suffering a small shock to the perception s_P shared by all agents, whereas Proposition 4 refers to a large shock to the perceptions of a small group of agents.

Proposition 4 directs our attention to the fate of the good equilibrium in the face of small invasions. Section A.5 proves an expected monotonicity result for such invasions:

Proposition 5*: [5.1] If the good equilibrium survives an invasion of size* $\lambda \leq 1/2$*, it survives any invasion of size* $\lambda' < \lambda$*. Similarly, if the good equilibrium is disrupted by an invasion of size* $\lambda < 1/2$*, it is disrupted by any invasion of size* $\lambda' \in [\lambda, 1/2]$*.*

[5.2] There is at most one value $\lambda \in [0, 1/2]$ *such that an invasion of size* λ *gives convergence to the unstable equilibrium.*

There are thus two possibilities. It may be that any invasion of size $\lambda \leq 1/2$ is unable to disrupt the good equilibrium. This will be the case for relatively large values of q, i.e., when there are many scoundrels. Alternatively, when q is sufficiently small, the interval $[0, 1/2]$ is partitioned by a value λ^*, with smaller invasions being assimilated to the good equilibrium, invasions of size λ^* leading to the unstable equilibrium, and larger invasions disrupting the good equilibrium and leading to the bad equilibrium.

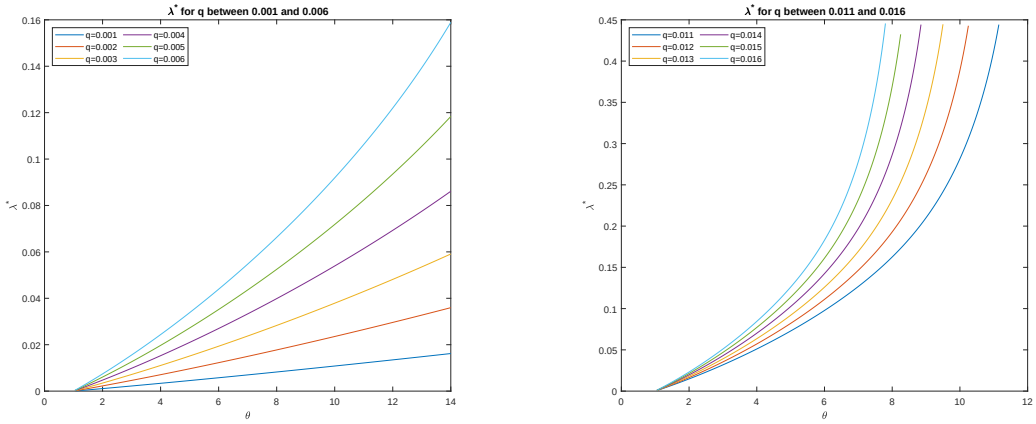

Figure 5: Each panel shows, for the proportions q of scoundrels indicated, the minimum invasion size λ^* (vertical axis) needed to disrupt the good equilibrium, as a function of θ (horizontal axis). The good equilibrium can withstand larger incursions when θ is larger and when there are more scoundrels. (Note the change in scale on the vertical axis in moving from panel (a) to (b).) Note also that when q is larger, as in panel (b), the range of θ that are consistent with the presence of multiple equilibria shrinks, with smaller and smaller value of θ required to ensure that only the good equilibrium exists.

Figure 5 shows the values of λ^* for selected values of q and θ.[8]

4.4. The Robustness of the Bad Equilibrium

We now turn this reasoning around. Consider the system (11)-(12) describing the dynamics of the perceptions of a mixture of agents, of which a fraction $(1-\lambda)$ start with perception $s_g = 0$ and a fraction λ start with perception s_b. If we now define $\hat{\lambda} = 1 - \lambda$, one readily sees that this is the same system as the one in which a proportion $1 - \hat{\lambda}$ of insiders whose initial perception is s_b and a proportion $\hat{\lambda}$ of outsiders whose initial perception is $s_g = 0$.

Proposition 4 then gives:

Corollary 1*: Consider the dynamic system (11)-(12), assuming that for a fraction* $1 - \lambda$ *of insiders* $s_1(0) = s_b$ *and for a fraction* λ *of outsiders* $s_0(0) = s_g = 0$ *(i.e. a system in which the insiders initially believe themselves to be in the bad equilibrium and the outsiders in the*

[8]The value λ^* is computed as the value of λ such that, when its 15th decimal digit is reduced by 1, the limit to which the system converges shifts from the bad equilibrium to the good equilibrium. This and the following figure are based on MatLab simulations of discrete approximations of our continuous dynamic system.

good equilibrium). For any $\lambda < 1$ there exists a $q^* > 0$ such that, for any[9] *$q \leq q^*$ it will be the case that $\lim_{t\to\infty} s_1(t) = \lim_{t\to\infty} s_0(t) = s_b$, i.e. the system converges back to the high cheating equilibrium.*

The proof is almost immediate and can be found in Section A.6. The intuition mirrors that of Proposition 4. As scoundrels become scarce, the basin of attraction of the bad equilibrium becomes large. It accordingly takes a large invasion of agents accustomed to the good equilibrium to disrupt the bad equilibrium. In the extreme, as q approaches zero, the basin of attraction of the bad equilibrium consumes the entire unit interval, allowing the bad equilibrium to withstand arbitrarily large invasions. Putting these results together, when scoundrels are scarce, the good equilibrium is upset by perturbations to bad behavior on the part of a tiny fraction of agents, while a large fraction of the population can shift to good behavior without disrupting the bad equilibrium.

5. Discussion

Trust can be fragile. A high-trust equilibrium can be easily disrupted by the injection of even a few bad apples, while a low-trust equilibrium can stubbornly resist the appearance of trusting agents. In another version of the common saying, trust takes years to build, seconds to break, and forever to repair.

The basic force behind these results is the convexity of the cost of cheating. When very few responsive responders cheat, a cheater is almost certain to be a scoundrel, and hence cheating is punished heavily. However, it initially takes only a modicum of cheating by responsive responders before a cheater is much less likely to be a scoundrel, and so the cost of cheating initially drops very rapidly as the incidence of cheating increases, and then subsequently falls less and less rapidly. The fewer the scoundrels, the more pronounced the effect of having even a few responsive responders among the ranks of cheaters, and so the more pronounced this convexity.

This convexity is intuitive. One readily notices and characterizes as a scoundrel the only person who litters in setting that everyone else preserves as pristine, or the only person who attempts to jump a queue that everyone else scrupulously maintains, or the only person who breaks a traffic law that everyone else respects. In contrast, there is little to be concluded from seeing a person commit such an act when many people do so.

[9] Recall that we are interested in the case in which there are three equilibria. So we must also have $q^* < \hat{q}(\theta)$ where $\hat{q}(\theta)$ is as in (10).

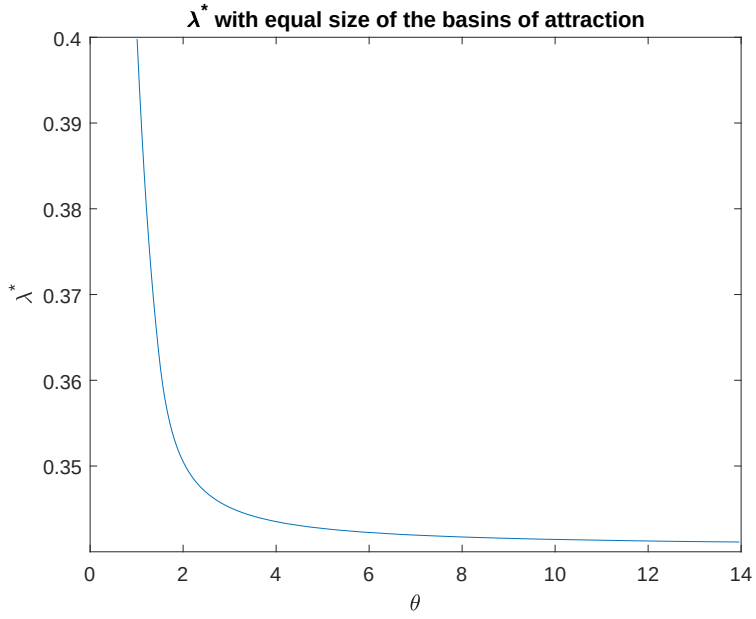

Figure 6: For each value of the cheating-cost parameter θ, the proportion of scoundrels q is set so that the good and bad equilibria have equal-sized basins of attraction. We then numerically calculate λ^*, the size of infusion of agents from the bad equilibrium just sufficient to disrupt the good equilibrium. As θ increases above 1, the proportion of scoundrels required to equalize the sizes of basins of attraction of the good and bad equilibrium decreases, making the cost-of-cheating function more convex, and hence reducing the infusion of agents accustomed to the bad equilibrium that suffices to disrupt the good equilibrium.

This convexity plays a dual role. As we have seen in Section 3, it ensures that as the proportion of scoundrels gets small, the basin of attraction of the good equilibrium shrinks and that of the bad equilibrium expands. This gives us a first sense in which trust becomes fragile when there are few scoundrels. Section 4 then showed that when there are few scoundrels, an arbitrarily small invasion of agents accustomed to the bad equilibrium can disrupt the good equilibrium. One might think that this is nothing more than a manifestation of the small basin of attraction of the good equilibrium. To see that this is not the case, Figure 6 reports results of the following exercise.

For various values of the cheating-cost parameter θ, we set the proportion of scoundrels q

so that the good and bad equilibria have equal-sized basins of attraction. We then numerically calculate λ^*, the size of infusion of agents from the bad equilibrium just sufficient to disrupt the good equilibrium, for each of these cases.

The first thing to note is that over the relevant range the value of λ^* is always well below 0.5 (it is in fact below 0.4). This, using Propositions 4 and Corollary 1, implies that when the basins of attractions are of equal size it is always the case that more outsiders are needed to disrupt the bad equilibrium than are needed to disrupt the good equilibrium.

Moreover, if the ability of such infusions to disrupt the bad equilibrium depended only on the sizes of the basins of attraction, the value of λ^* in Figure 6 would be constant in θ. Instead, as θ increases the size of the lethal infusion decreases. In particular, as θ increases, the proportion of scoundrels required to equalize the sizes of basins of attraction of the good and bad equilibrium decreases, making the cost-of-cheating function more convex, and hence reducing the infusion of agents accustomed to the bad equilibrium that suffices to disrupt the good equilibrium.

Figure 7 portrays the path of the dynamic systems, for two cases in which an invasion disrupts the good equilibrium and induces convergence to the bad equilibrium. In each case, s_1 initially equals $s_g = 0$ (insider perception and behavior are initially consistent with the good equilibrium) and s_0 initially equals s_b (outsider perception and behavior are initially consistent with the bad equilibrium).

Two aspects of these dynamics stand out. First, s_1 initially increases as insiders adjust to the more-than-expected cheating carried out by outsiders. However, the outsiders' perceived level of cheating s_0 falls, as they meet less cheating than expected when matched with insiders. The outsiders' perceived level of cheating falls much more dramatically, reflecting their smaller share of the population, and hence the realized level of cheating, s, on balance falls (only imperceptibly at the beginning). There thus initially appears to be overwhelming evidence that the population is adjusting toward the good equilibrium. However, in both cases the direction of s eventually reverses (after some seeming indecision in the right panel) and the population converges to the bad equilibrium. Second, the adjustment of the aggregate level of cheating not only need not be monotonic, but can be complicated, in the right panel reversing direction three times.

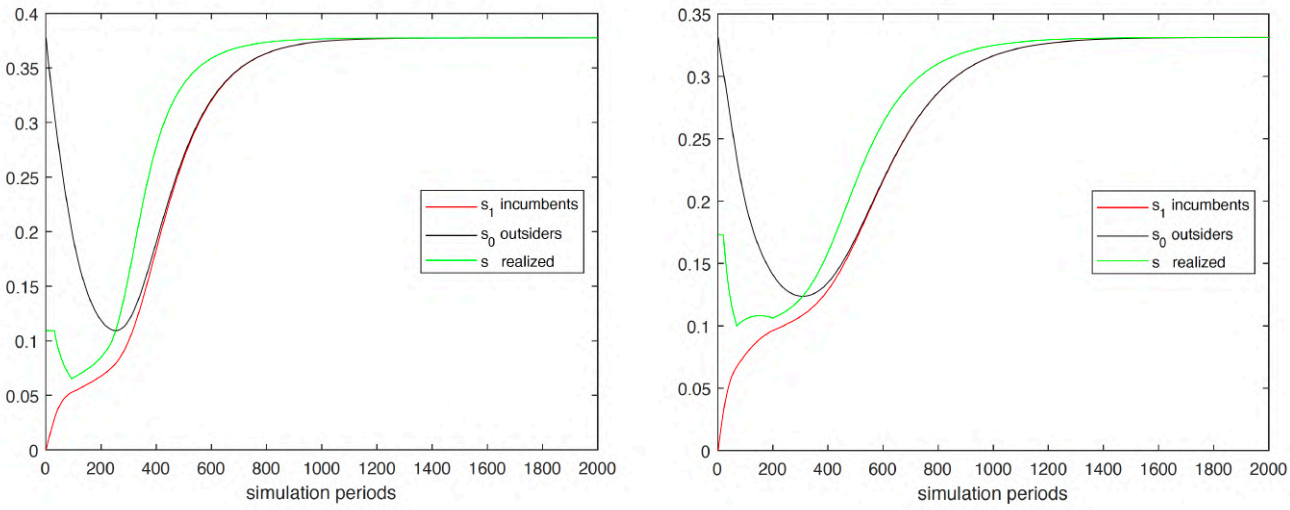

Figure 7: Depiction of the dynamics for two cases in which an invasion of agents accustomed to the bad equilibrium leads the population to converge to the bad equilibrium. Each panel shows the paths of the insider perception of cheating s_1 (red), outsider perception of cheating s_0 (black) and the realized perceptions of cheating s (green). The parameters underlying the left panel are $\theta = 2, \lambda = 0.118, q = 0.05$, those in the right panel are $\theta = 2, \lambda = 0.2, q = 0.0634$. The kinks in the paths arise as various of the min and max operators in (11) come into play.

The idea that trust can be fragile is familiar. The more surprising finding to emerge from this exploration is that, perhaps paradoxically, trust can be more robust when there are more agents in the economy who never trust. Intuitively, this is because social disapproval is heaped on cheaters who do so without a good reason. The more reckless cheaters there are—the more scoundrels, as we call them—the more likely a person observed cheating is one of them, and therefore the more his cheating is socially sanctioned. Scoundrels can thus be valuable for two reasons. Increasing the number of scoundrels may convert an economy with multiple equilibria into an economy with a unique (good) equilibrium. As we have seen in Section 3, if the former economy is coordinated on the less trusting of the multiple equilibria, the increase in scoundrels can lead to an increase in trust. In addition, an increase in the number of scoundrels can render an economy coordinated on the most trusting of multiple equilibria better able to withstand perturbations to that equilibrium. The most fortunate economy is one that has few scoundrels, and hence multiple equilibria, but that has coordinated on the

high-trust equilibrium. But the higher is the level of trust in the good equilibrium (i.e., the fewer scoundrels), the more precarious is the equilibrium itself.

Our analysis points to steps that might mitigate this fragility. If we broadened the purview of our analysis to accommodate either multiple or continuous arrivals of outsiders, then we expect that an economy whose good equilibrium would be disrupted by a moderate influx of outsiders accustomed to the bad equilibrium could accommodate an even larger number of such additions if they occur sufficiently slowly. This moderated flow would allow previous arrivals to have time to adjust and thus keep the system within the basin of attraction of the good equilibrium, even as the flow of new arrivals continues. Taking steps to hasten the adjustment of perceptions would allow the good equilibrium to withstand a larger influx of outsiders, but taking steps to reduce the number of scoundrels would have the reverse effect. We can expect an institution devoid of scoundrels (perhaps Minnesota?) to have more difficulty accommodating arrivals accustomed to the bad equilibrium than a somewhat grittier one (perhaps New York?).

We have worked throughout with the simple specification of the social cost of cheating given by (1) and simple, symmetric adjustment dynamic given by (12). We believe that, if anything, the more realistic components we might build into these specifications would reinforce our basic finding that trust is likely to be fragile. For example, we expect violations of trust in a high-trust environment to be more visible and more salient than episodes of trust in a low-trust environment. If so, the tendency of shocks to disrupt a high-trust equilibrium will be exacerbated.

References

Anderlini, L., and D. Terlizzese (2017): "Equilibrium Trust," *Games and Economic Behavior*, 102, 624–644.

Arrow, K. J. (1974): *The limits of organization*. WW Norton & Company.

Berg, J., J. Dickhaut, and K. McCabe (1995): "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, 10, 122–142.

Jackson, M. O. (2020): "A Typology of Social Capital and Associated Network Measures," *Social Choice and Welfare*, 54, 311–336.

Levitsky, S., and D. Ziblatt (2019): *How democracies die*. Crown.

Putnam, R. D. (2000): *Bowling Alone*. New York: Simon and Schuster.

Slovic, P. (1993): "Perceived risk, trust, and democracy," *Risk analysis*, 13(6), 675–682.

——— (1999): "Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield," *Risk analysis*, 19, 689–701.

Appendix

A.1. Proof of Proposition 1

From (3), if $\theta < 1$, then $s^* < 0$. Hence the equilibrium conditions (4)–(6) reduce to

$$\begin{aligned} s &= \max\{0, x - f(s)\} \\ &= \max\left\{0, \frac{1}{2} - \frac{1}{2}f(s)\right\} \\ &= \max\left\{0, \frac{1}{2} - \frac{1}{2}\frac{\theta q}{q + (1-q)s}\right\} \\ &= \frac{1}{2} - \frac{1}{2}\frac{\theta q}{q + (1-q)s}. \end{aligned}$$

Given $\theta < 1$, this equation has only one positive (real) solution. ■

A.2. Proof of Proposition 3

Straightforward manipulations of (8) and (9), in the range consistent with $q < \hat{q}(\theta)$, imply that s_u is an increasing function of q and θ, while s_b is a decreasing function of q and θ, with

$$\lim_{q \to 0} s_u(q) = 0 \qquad \text{and} \qquad \lim_{q \to 0} s_b(q) = \frac{1}{2}.$$

■

A.3. Proof of Lemma 1

We first note that for $s_u \leq s \leq s_b$, we have that

$$\frac{1}{2} - \frac{1}{2}f(s) \geq s \tag{A.1}$$

with a strict inequality except at the two boundaries, while for both $s < s_u$ and $s > s_b$, it is true that

$$\frac{1}{2} - \frac{1}{2}f(s) < s. \tag{A.2}$$

We can write $s(t) = h(s_0(t), s_1(t))$ and then write the dynamical system (11)-(12) as

$$\begin{aligned} \dot{s}_1(t) &= \delta\{h(s_0(t), s_1(t)) - s_1(t)\} \\ \dot{s}_0(t) &= \delta\{h(s_0(t), s_1(t)) - s_0(t)\}, \end{aligned} \tag{A.3}$$

where the function $h(s_0(t), s_1(t))$ is derived from (11) and gives the realized proportion of cheating by responsive responders, $s(t)$, as a function of the current state of the perceptions by outsiders and incumbents, respectively, $(s_0(t), s_1(t))$. In the following argument, we repeatedly use the facts that the function h is

uniformly continuous on $[0,1]^2$, and that along the diagonal $s_1(t) = s_0(t) = s$, the function h is given by

$$h(s,s) = \begin{cases} 0 & s \leq s^* \\ \frac{1}{2} - \frac{1}{2}f(s) & s \geq s^*, \end{cases}$$

and hence, as implied by (A.1) and (A.2), we have

$$\begin{aligned} h(s,s) - s &< 0 \qquad s < s_u \\ h(s,s) - s &= 0 \qquad s = s_u \qquad \text{(A.4)} \\ h(s,s) - s &> 0 \qquad s_u < s < s_b \qquad \text{(A.5)} \\ h(s,s) - s &= 0 \qquad s = s_b \qquad \text{(A.6)} \\ h(s,s) - s &< 0 \qquad s > s_b. \qquad \text{(A.7)} \end{aligned}$$

Fix a sufficiently small $\eta > 0$. Then there exists $\varepsilon(\eta) > 0$ such that

$$\begin{aligned} s \in [\eta, s_u - \eta] &\implies h(s,s) - s < -\varepsilon(\eta) \qquad \text{(A.8)} \\ s \in [s_u + \eta, s_b - \eta] &\implies h(s,s) - s > \varepsilon(\eta) \qquad \text{(A.9)} \\ s \in [s_b + \eta, 1] &\implies h(s,s) - s < -\varepsilon(\eta). \qquad \text{(A.10)} \end{aligned}$$

Let $\|\cdot\|$ denote the sup norm. There exists $\gamma(\eta) > 0$ sufficiently small such that $\|(s_0, s_1) - (s_0, s_0)\| < \gamma(\eta)$ implies[10]

$$\begin{aligned} |h(s_0, s_1) - h(s_0, s_0)| &< \frac{\varepsilon(\eta)}{4} \\ |s_0 - s_1| &< \frac{\varepsilon(\eta)}{4}, \end{aligned} \qquad \text{(A.11)}$$

which in turn imply, using the triangle inequality,

$$|h(s_0, s_1) - s_0) - (h(s_0, s_0) - s_0)| < \frac{\varepsilon(\eta)}{2} \qquad \text{(A.12)}$$

$$|(h(s_0, s_1) - s_1) - (h(s_0, s_0) - s_0)| < \frac{\varepsilon(\eta)}{2}. \qquad \text{(A.13)}$$

Hence, whenever $\|(s_0, s_1) - (s_0, s_0)\| < \gamma(\eta)$, we can combine (A.12) and (A.13) with (A.8)–(A.10), to establish the following implications:

$$s_0, s_1 \in [\eta, s_u - \eta] \implies \left[h(s_0, s_1) - s_0 < -\frac{\varepsilon(\eta)}{2}, \quad h(s_0, s_1) - s_1 < -\frac{\varepsilon(\eta)}{2} \right] \qquad \text{(A.14)}$$

[10] The first inequality follows from the absolute continuity of h. We can ensure the second by taking $\gamma(\eta)$ to be sufficiently small.

$$s_0, s_1 \in [s_u + \eta, s_b - \eta] \implies \left[h(s_0, s_1) - s_0 > \frac{\varepsilon(\eta)}{2}, \quad h(s_0, s_1) - s_1 > \frac{\varepsilon(\eta)}{2}\right] \tag{A.15}$$

$$s_0, s_1 \in [s_b + \eta, 1] \implies \left[h(s_0, s_1) - s_0 < -\frac{\varepsilon(\eta)}{2}, \quad h(s_0, s_1) - s_1 < -\frac{\varepsilon(\eta)}{2}\right]. \tag{A.16}$$

From (13), we see that there exists $T(\eta)$ such that for all $t > T(\eta)$, we have $\|(s_0, s_1) - (s_0, s_0)\| < \min\{\eta, \gamma(\eta)\}$. The preceding three implications then imply two possibilities:

- For all $t > T(\eta)$, s_0 and s_1 are both within 2η of s_u.
- There is a time $t' > T$ at which at least one of s_0 or s_1 differ from s_u by more than 2η. Then both s_0 and s_1 differ from s_u by more than η. Hence, (A.14)–(A.16) imply that there exists a time $t'' \geq t'$ such that for all $t > t''$, either both s_0 and s_1 differ from s_g by at most 2η or both s_0 and s_1 differ from s_b by at most 2η.

Since this holds for any $\eta > 0$, we have convergence. ■

A.4. Proof of Proposition 4

We begin with a preliminary result.

Lemma A.1: *Let $q < \hat{q}(\theta)$, so that there are 3 distinct equilibria. If at some finite time t it is the case that $s_1(t) = s_u$, the dynamic system (11)-(12), with initial conditions $s_1(0) = 0$ and $s_0(0) = s_b$, converges to s_b.*

Proof: Using (13) we can write the dynamics entirely in terms of $s_1(t)$ and t, for a given λ:

$$\begin{aligned}
&\dot{s}_1(t) = \\
&\delta \Big\{ (1-\lambda)^2 \min\{1, \max\{0, \max\{f(s_1(t)), \frac{1}{2} + \frac{1}{2} f(s_1(t))\} - f(s_1(t))\}\} \\
&+\lambda(1-\lambda) \min\{1, \max\{0, \max\{f(s_1(t)), \frac{1}{2} + \frac{1}{2} f(s_1(t))\} - f(s_1(t) + e^{-\delta t} s_b)\}\} \\
&+\lambda(1-\lambda) \min\{1, \max\{0, \max\{f(s_1(t) + e^{-\delta t} s_b), \frac{1}{2} + \frac{1}{2} f(s_1(t) + e^{-\delta t} s_b)\} - f(s_1(t))\}\} \\
&+\lambda^2 \min\{1, \max\{0, \max\{f(s_1(t) + e^{-\delta t} s_b), \frac{1}{2} + \frac{1}{2} f(s_1(t) + e^{-\delta t} s_b)\} - f(s_1(t) + e^{-\delta t} s_b)\}\} - s_1(t) \Big\}.
\end{aligned} \tag{A.17}$$

Assume now that, at some finite t, $s_1(t) \geq s_u$. For any $\theta > 1$ this implies that

$$s_0(t) > s_1(t) \geq s_u > s^* > 0.$$

As long as $s_1(t) < s_b$ (which is strictly larger than s_u, given that $q < \hat{q}(\theta)$), we can simplify the dynamics, since all the inner max appearing in (A.17) are solved by the second of the two expressions. More in detail, in the expression multiplied by $(1-\lambda)^2$ we have:

$$\max\left\{f(s_1(t)), \frac{1}{2} + \frac{1}{2}f(s_1(t))\right\} - f(s_1(t)) \;=\; \frac{1}{2} - \frac{1}{2}f(s_1(t)).$$

In the first of the two expressions multiplied by $\lambda(1-\lambda)$ we have:

$$\max\left\{f(s_1(t)), \frac{1}{2} + \frac{1}{2}f(s_1(t))\right\} - f(s_1(t) + e^{-\delta t}s_b) = \frac{1}{2} + \frac{1}{2}f(s_1(t)) - f(s_1(t) + e^{-\delta t}s_b).$$

In the second of the two expressions multiplied by $\lambda(1-\lambda)$ we have:

$$\max\left\{f(s_1(t) + e^{-\delta t}s_b), \frac{1}{2} + \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b)\right\} - f(s_1(t)) = \frac{1}{2} + \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b) - f(s_1(t)).$$

Note that

$$\frac{1}{2} + \frac{1}{2}f(s_1(t)) - f(s_1(t) + e^{-\delta t}s_b) > 0$$

and since $s_1(t) > s^*$ this implies that

$$1 > f(s_1(t)) > f(s_1(t) + e^{-\delta t}s_b).$$

Therefore, the first of the two expressions multiplied by $\lambda(1-\lambda)$ reduces to

$$\frac{1}{2} + \frac{1}{2}f(s_1(t)) - f(s_1(t) + e^{-\delta t}s_b).$$

Since

$$\frac{1}{2} + \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b) - f(s_1(t))$$

cannot be signed, the second of the two expressions in (A.17) multiplied by $\lambda(1-\lambda)$ only reduces to

$$\max\left\{0, \frac{1}{2} + \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b) - f(s_1(t))\right\}.$$

Finally, for the expression in (A.17) multiplied by λ^2 we have:

$$\max\{f(s_1(t) + e^{-\delta t}s_b), \frac{1}{2} + \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b)\} - f(s_1(t) + e^{-\delta t}s_b) = \frac{1}{2} - \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b).$$

Putting together all these observations about the four components of the right side of (A.17) we get

$$\begin{aligned}
&\dot{s}_1(t) = \\
&\delta\left\{(1-\lambda)^2\left(\frac{1}{2}-\frac{1}{2}f(s_1(t))\right)\right. \\
&+\lambda(1-\lambda)\left(\frac{1}{2}+\frac{1}{2}f(s_1(t))-f(s_1(t)+e^{-\delta t}s_b)+\max\left\{0,\frac{1}{2}+\frac{1}{2}f(s_1(t)+e^{-\delta t}s_b)-f(s_1(t))\right\}\right) \\
&\left.+\lambda^2\left(\frac{1}{2}-\frac{1}{2}f(s_1(t)+e^{-\delta t}s_b)\right)-s_1(t)\right\} \geq \qquad (A.18)\\
&\qquad \delta\left\{(1-\lambda)^2\left(\frac{1}{2}-\frac{1}{2}f(s_1(t))\right)+\lambda(1-\lambda)\left(1-\frac{1}{2}f(s_1(t))-\frac{1}{2}f(s_1(t)+e^{-\delta t}s_b)\right)+\right. \\
&\qquad\qquad \left.\lambda^2\left(\frac{1}{2}-\frac{1}{2}f(s_1(t)+e^{-\delta t}s_b)\right)-s_1(t)\right\} \\
&\qquad = \delta\left\{\frac{1}{2}-\frac{1}{2}f(s_1(t))+\frac{\lambda}{2}(f(s_1(t))-f(s_1(t)+e^{-\delta t}s_b))-s_1(t)\right\},
\end{aligned}$$

where the middle inequality results from neglecting the max operator.

Given that $s_u \leq s_1(t) < s_b$, we know that

$$s_1(t) \leq \frac{1}{2}-\frac{1}{2}f(s_1(t)) \quad \Leftrightarrow \quad \frac{1}{2}-\frac{1}{2}f(s_1(t))-s_1(t) \geq 0.$$

Moreover, since f is decreasing, for any finite t we have

$$f(s_1(t))-f(s_1(t)+e^{-\delta t}s_b) > 0$$

Hence $\dot{s}_1(t) > 0$ for all $s_u \leq s_1(t) < s_b$. Since we know that the system converges, it must then be that $s_1(t)$ converges to s_b. ■

The proof of Proposition 4 now proceeds in four steps.

Step 1: *Bounding s_0 from below for an initial interval of time*

First, fix $\lambda \leq 1/2$, θ and a value of $0 < q < \hat{q}(\theta)$, to guarantee that there are three equilibria (to simplify the notation, we will denote this as $\hat{q}$). Recall the dynamics

$$\begin{aligned}
\dot{s}_1 &= \delta(s-s_1) \\
\dot{s}_0 &= \delta(s-s_0). \qquad (A.19)
\end{aligned}$$

Recall that s_{00} is the amount of cheating that occurs when an outsider proposer meets an outside responder. At time 0, we have $s_{00} = s_b$, where we recall that the latter is the level of cheating characterizing the bad equilibrium. Then in general we have, using (11),

$$s \geq \lambda^2 s_{00},$$

and hence

$$\begin{aligned} \dot{s}_1 &\geq \delta(\lambda^2 s_{00} - s_1) \\ \dot{s}_0 &\geq \delta(\lambda^2 s_{00} - s_0). \end{aligned} \tag{A.20}$$

Now we note that, as long as $s_0 > s^*$(which initially must be the case given that $s_0(0) = s_b > s^*$), we have

$$s_{00} = \frac{1}{2} - \frac{1}{2}\frac{\theta q}{q + (1-q)s_0},$$

and so we can write

$$\begin{aligned} \dot{s}_1 &\geq \delta\left(\lambda^2\left(\frac{1}{2} - \frac{1}{2}\frac{\theta q}{q + (1-q)s_0}\right) - s_1\right) \\ \dot{s}_0 &\geq \delta\left(\lambda^2\left(\frac{1}{2} - \frac{1}{2}\frac{\theta q}{q + (1-q)s_0}\right) - s_0\right). \end{aligned} \tag{A.21}$$

The right hand side in (A.21) is larger than the expression we obtain by setting to 0 the s_0 that appears in the denominator. Hence we have

$$\dot{s}_0 \geq \delta\left(\lambda^2\left(\frac{1}{2} - \frac{1}{2}\theta\right) - s_0\right)$$

for all $q \in (0, \hat{q})$.

Hence, for any $\eta > 0$, there exists a time t_η such that $s_0(t) \geq s_b - \eta$ for all $t \in [0, t_\eta]$.

Step 2: *Bounding s_1 from below at a given point in time*

Consider now (A.21). The expression within the inner brackets is increasing in s_0 and decreasing in q. Therefore, over the interval $[0, t_\eta]$, replacing s_0 by its lower bound of $s_b - \eta$, and again s_b by its lower bound[11] of $(1 - 3\hat{q})/(4(1-\hat{q}))$, we reduce that expression. We also reduce it replacing q by its upper bound of $\hat{q}$. Combining these changes we obtain a lower bound on the right side of (A.21) that implies

$$\dot{s}_1(t) \geq \delta\left(\lambda^2\left(\frac{1}{2} - \frac{1}{2}\frac{\theta\hat{q}}{\hat{q}(\frac{1}{4} + \eta) + (\frac{1}{4} - \eta)}\right) - s_1(t)\right).$$

[11] See (8) and (10).

It is a bit tedious but straightforward to verify that, for any $\theta > 1$ it must be that

$$\frac{1}{2}\left(1 - \frac{\theta\hat{q}}{\frac{1}{4}(\hat{q}+1)}\right) > 0.$$

We can then choose η sufficiently small so that

$$\frac{1}{2}\left(1 - \frac{\theta\hat{q}}{\hat{q}(\frac{1}{4}+\eta) + (\frac{1}{4}-\eta)}\right) > 0.$$

Then we have that

$$\dot{s}_1(t) \geq \delta(A - s_1(t))$$

for some $A > 0$ and for any fixed $q \in (0, \hat{q})$ and all $t \in [0, t_\eta]$.

Hence, there exists a time τ and value $\xi > 0$ such that, for any fixed $q \in (0, \hat{q})$, we have,

$$s_1(\tau) \geq \xi > 0.$$

Step 3*: Pushing s_u below s_1.*

Now let q approach 0. As we do so, $s_u(q) \to 0$. Hence, for all sufficiently small q, at time τ we have $s_1(\tau) > s_u$.

Step 4*: Showing convergence to s_b.*

We can now invoke Lemma A.1 and conclude that $s_1(t)$ converges to s_b. ■

A.5. Proof of Proposition 5

The outline of the argument is as follows.

First, we think of $s(t)$, the realized proportion of cheaters at time t, as a function $s(s_1(t), t, \lambda)$ of $s_1(t)$ (the insiders' perceived level of cheating at time t), t and λ.[12]

Second, we show that for fixed s_1 and t, the smaller is λ the smaller is $s(s_1, t, \lambda))$.

This in turn ensures that, for a fixed s_1 and t, the smaller is λ, the smaller is ds_1/dt.

Third, suppose that the path of $s_1(t)$ induced by λ converges to s_g, the good equilibrium. Then, for a smaller value λ', we get a path in which, at every time t, either the induced value of s_1 is smaller, or (if equal)

[12] In principle, we should write $s_1(t, \lambda)$, but omit the latter argument to conserve on clutter. We need not include $s_0(t)$ as an argument of s, since (from (13)) this can be inferred from s_1 and t.

the derivative ds_1/dt is smaller. Hence, the path induced by the smaller value λ' is always either below or being pushed below that induced by λ, and so the λ' path also converges to 0. Hence, if the path of $s_1(t)$ induced by λ converges to s_g, then so does the path induced by any $\lambda' < \lambda$. A similar argument shows that if the path of $s_1(t)$ induced by λ converges to s_b, then so does the path induced by any $\lambda' > \lambda$. This gives [5.1].

Finally, we show [5.2], that at most one value $\lambda \in [0, 1/2]$ induces convergence to s_u.

We begin with a preliminary result.

Lemma A.2: *Consider two paths of insider perceptions, $s_1(t, \lambda_1)$ and $s_1(t, \lambda_2)$, with $\lambda_1 > \lambda_2$. Suppose both paths converge to s_u. Then, for all t large enough, it must be the case that $s_1(t, \lambda_1) < s_1(t, \lambda_2)$,*

Proof: To simplify the notation, denote by s_1^j the path of the insider perceptions corresponding to λ_j. For a t large enough, we know that the dynamics of $s_1^1(t)$ and $s_1^2(t)$ follow

$$\dot{s}_1^1(t) = \delta \left\{ \frac{1}{2} - \frac{1}{2} f(s_1^1(t)) - s_1^1(t) + \frac{\lambda_1}{2} (f(s_1^1(t)) - f(s_1^1(t) + c)) \right\},$$

and

$$\dot{s}_1^2(t) = \delta \left\{ \frac{1}{2} - \frac{1}{2} f(s_1^2(t)) - s_1^2(t) + \frac{\lambda_2}{2} (f(s_1^2(t)) - f(s_1^2(t) + c)) \right\},$$

where $c = e^{-\delta t} s_b$ is, for a given t, a constant which is common to both paths.

We want to show that, if t is large enough, it cannot be that $s_1^1(t) \geq s_1^2(t)$. Suppose, by way of contradiction, that this is the case. We will show that this implies that

$$\dot{s}_1^1(t) > \dot{s}_1^2(t).$$

This in turn implies that $s_1^1(t)$ and $s_1^2(t)$ would diverge from each other, and therefore they could not both converge to s_u.

If at some (large) t were the case that $s_1^1(t) = s_1^2(t)$, it would easily follow that $\dot{s}_1^1(t) > \dot{s}_1^2(t)$. Starting from t, the path for s_1^1 would then immediately be above the path for s_1^2. We would then need to consider the case $s_1^1(t) > s_1^2(t)$, to which we turn.

We have

$$\dot{s}_1^1(t) - \dot{s}_1^2(t) = \delta \left\{ \frac{1}{2} (f(s_1^2(t)) - f(s_1^1(t))) + s_1^2(t) - s_1^1(t) + \right. \tag{A.22}$$

$$\left. \frac{\lambda_1}{2} (f(s_1^1(t)) - f(s_1^1(t) + c)) - \frac{\lambda_2}{2} (f(s_1^2(t)) - f(s_1^2(t) + c)) \right\}. \tag{A.23}$$

As a preliminary step, we show that

$$\frac{1}{2} (f(s_1^2(t)) - f(s_1^1(t))) > s_1^1(t) - s_1^2(t).$$

Indeed,

$$\frac{1}{2}(f(s_1^2(t)) - f(s_1^1(t))) = \frac{\theta q}{2}\frac{(1-q)(s_1^1(t) - s_1^2(t))}{(q+(1-q)s_1^1(t))(q+(1-q)s_1^2(t))}, \tag{A.23}$$

hence

$$\frac{1}{2}(f(s_1^2(t)) - f(s_1^1(t))) > s_1^1(t) - s_1^2(t)$$

if

$$\frac{\theta q}{2}\frac{(1-q)}{(q+(1-q)s_1^1(t))(q+(1-q)s_1^2(t))} > 1.$$

In turn, given that both $s_1^1(t)$ and $s_1^2(t)$ are smaller than s_u, we have that

$$\begin{aligned}\frac{\theta q}{2}\frac{(1-q)}{(q+(1-q)s_1^1(t))(q+(1-q)s_1^2(t))} &> \frac{\theta q}{2}\frac{(1-q)}{(q+(1-q)s_u)^2} \\ &= \frac{\theta q(1-q)}{2}\frac{(1-2s_u)^2}{(\theta q)^2} \\ &= \frac{(1-q)}{2\theta q}(1-2s_u)^2,\end{aligned}$$

where we used equations (7) and (1) to replace $q + (1-q)s_u$.

Using now the definition of s_u (equation (9)) we have that

$$1 - 2s_u = \frac{1+q+\sqrt{(q+1)^2 - 8\theta q(1-q)}}{2(1-q)}.$$

Therefore,

$$\begin{aligned}\frac{(1-q)}{2\theta q}(1-2s_u)^2 &= \frac{2(1+q)^2 - 8\theta q(1-q) + 2(1+q)\sqrt{(1+q)^2 - 8\theta q(1-q)}}{8\theta q(1-q)} \\ &= \frac{(1+q)^2}{4\theta q(1-q)} - 1 + \frac{(1+q)}{4\theta q(1-q)}\sqrt{(1+q)^2 - 8\theta q(1-q)}.\end{aligned}$$

We need to establish whether the right side is larger than 1. This is equivalent to establish whether

$$\sqrt{(1+q)^2 - 8\theta q(1-q)} > \frac{8\theta q(1-q)}{1+q} - (1+q).$$

Squaring both sides we obtain

$$(1+q)^2 - 8\theta q(1-q) > \frac{(8\theta q(1-q))^2}{(1+q)^2} + (1+q)^2 - 16\theta q(1-q).$$

Simplifying this boils down to

$$(1+q)^2 > 8\theta q(1-q),$$

which is a condition satisfied as long as we have 3 equilibria of the dynamic system. This establishes the

preliminary step

$$\frac{1}{2}(f(s_1^2(t)) - f(s_1^1(t))) > s_1^1(t) - s_1^2(t). \tag{A.24}$$

Rewrite now equation (A.22) as follows:

$$\dot{s}_1^1(t) - \dot{s}_1^2(t) = \\ \delta\left\{\frac{1}{2}\{[f(s_1^2(t))(1-\lambda_2) + f(s_1^2(t)+c)\lambda_2] - [f(s_1^1(t))(1-\lambda_1) + f(s_1^1(t)+c)\lambda_1]\} + s_1^2(t) - s_1^1(t)\right\}. \tag{A.25}$$

The expression within the first pair of square brackets can be written as

$$f(s_1^2(t)) - \lambda_2 k_2, \tag{A.26}$$

where

$$k_2 = \frac{\theta q(1-q)c}{(q+(1-q)s_1^2(t))(q+(1-q)(s_1^2(t)+c)}.$$

Similarly, the expression within the second pair of square brackets can be written as

$$f(s_1^1(t)) - \lambda_1 k_1, \tag{A.27}$$

where

$$k_1 = \frac{\theta q(1-q)c}{(q+(1-q)s_1^1(t))(q+(1-q)(s_1^1(t)+c)},$$

and $k_2 > k_1$.

Therefore, the right side of (A.25) can be written as

$$\delta\left\{\frac{1}{2}\{f(s_1^2(t)) - f(s_1^1(t)) + \lambda_1 k_1 - \lambda_2 k_2\} + s_1^2(t) - s_1^1(t)\right\}.$$

We now show that, when t is sufficiently large, and therefore c is sufficiently small, $\lambda_1 k_1 - \lambda_2 k_2 \geq 0$. This inequality is equivalent to

$$\frac{\lambda_1 - \lambda_2}{\lambda_1} \geq \frac{k_2 - k_1}{k_2} = 1 - \frac{(q+(1-q)s_1^2(t))(q+(1-q)(s_1^2(t)+c)}{q+(1-q)s_1^1(t))(q+(1-q)(s_1^1(t)+c)}.$$

The left side is a positive, constant scalar. As t becomes large the right side approaches 0. For a sufficiently large t this then proves that $\lambda_1 k_1 - \lambda_2 k_2 \geq 0$, which in turn implies, using (A.24), that $\dot{s}_1^1(t) - \dot{s}_1^2(t) > 0$. ■

As we anticipated the actual proof of Proposition 5 is divided into four steps.

Step 1*:* Recalling (11) and using (13) (specialized to the case we are considering) we define

$$s(t) \coloneqq s(s_1(t), t, \lambda) = (1-\lambda)^2\zeta_{11} + \lambda(1-\lambda)(\zeta_{01} + \zeta_{10}) + \lambda^2\zeta_{00} \tag{A.28}$$

where

$$\zeta_{11} = \min\left\{1, \max\{0, \max\{f(s_1(t)), \frac{1}{2} + \frac{1}{2}f(s_1(t))\} - f(s_1(t))\}\right\} \tag{A.29}$$

$$\zeta_{10} = \min\left\{1, \max\{0, \max\{f(s_1(t)), \frac{1}{2} + \frac{1}{2}f(s_1(t))\} - f(s_1(t) + e^{-\delta t}s_b)\}\right\} \tag{A.30}$$

$$\zeta_{01} = \min\left\{1, \max\{0, \max\{f(s_1(t) + e^{-\delta t}s_b), \frac{1}{2} + \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b)\}\ f(s_1(t))\}\right\} \tag{A.31}$$

$$\zeta_{00} = \min\left\{1, \max\{0, \max\{f(s_1(t) + e^{-\delta t}s_b), \frac{1}{2} + \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b)\}\ f(s_1(t) + e^{-\delta t}s_b)\}\right\}. \tag{A.32}$$

For any given $s_1(t)$ and t, we have

$$\begin{aligned} \frac{\partial s}{\partial \lambda} &= -2(1-\lambda)\zeta_{11} + (1-2\lambda)(\zeta_{01} + \zeta_{10}) + 2\lambda\zeta_{00} \\ &= -2\zeta_{11} + (\zeta_{01} + \zeta_{10}) + 2\lambda[\zeta_{11} + \zeta_{00} - (\zeta_{01} + \zeta_{10})] \\ \frac{\partial^2(s)}{\partial^2 \lambda} &= 2(\zeta_{11} + \zeta_{00} - (\zeta_{01} + \zeta_{10})). \end{aligned}$$

Step 2*:* We show that $\partial s/\partial\lambda \geq 0$ in the interval $\lambda \in [0, 1/2]$. Because the second derivative has a constant sign over this interval, it suffices to show that $\partial s/\partial\lambda \geq 0$ for $\lambda = 0$ and $\lambda = \frac{1}{2}$. The corresponding requirements are

$$\begin{aligned} 2\,\zeta_{11} &\leq \zeta_{01} + \zeta_{10} \\ \zeta_{11} &\leq \zeta_{00}. \end{aligned} \tag{A.33}$$

The second of these is almost immediate.[13] For any fixed λ, for all t it is the case that $s_1(t) + e^{-\delta t}s_b \geq s_1(t)$ (in fact the inequality is always strict and tends to an equality as t tends to ∞). If $s_1(t) > s^*$, then also $s_1(t) + e^{-\delta t}s_b > s^*$. Therefore,

$$\zeta_{11} = \frac{1}{2} - \frac{1}{2}f(s_1(t)) < \zeta_{00} = \frac{1}{2} - \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b),$$

since f is decreasing and $f(s_1(t)) < 1$. If $s_1(t) \leq s^*$, there are two possibilities: either $s_1(t) + e^{-\delta t}s_b > s^*$ or $s_1(t) + e^{-\delta t}s_b \leq s^*$. In the first case,

$$\zeta_{11} = 0 < \frac{1}{2} - \frac{1}{2}f(s_1(t) + e^{-\delta t}s_b) = \zeta_{00}.$$

[13]Intuitively, ζ_{11} is the level of cheating when two good agents meet, and ζ_{00} is the level of cheating when two bad agents meet. The second requirement is then the statement that bad agents cheat more than good agents.

In the second case,

$$\zeta_{11} = 0 = \zeta_{00}.$$

Moving to the first, we need $2\,\zeta_{11} \leq \zeta_{01} + \zeta_{10}$. We can simplify the expressions for ζ_{11}, ζ_{01} and ζ_{10} as follows (for notational convenience, we neglect the dependence of s_1 on t and we denote by s_0 the term $s_1(t) + e^{-\delta t} s_b$):

$$\begin{aligned}
\zeta_{11} &= \max\left\{0, \frac{1}{2} - \frac{1}{2} f(s_1)\right\} \\
\zeta_{10} &= \min\left\{1, \max\{f(s_1) - f(s_0), \frac{1}{2} + \frac{1}{2} f(s_1) - f(s_0)\}\right\} \\
\zeta_{01} &= \max\left\{0, \frac{1}{2} + \frac{1}{2} f(s_0) - f(s_1)\right\},
\end{aligned} \tag{A.34}$$

These hold because,

- In equation (A.29) for ζ_{11}, if $s_1 \leq s^*$, the inner maximum is solved by $f(s_1)$, so the whole expression is 0, while if $s_1 > s^*$ the inner maximum is solved by $\frac{1}{2} + \frac{1}{2} f(s_1) < 1$, so the whole expression is $\frac{1}{2} - \frac{1}{2} f(s_1)$;
- In equation (A.30) for ζ_{10}, again, if $s_1 \leq s^*$, the inner maximum is solved by $f(s_1)$, hence we have $f(s_1) - f(s_0)$; this could be bigger than 1, so we cannot neglect the outer minimum. If $s_1 > s^*$ the inner maximum is solved by $\frac{1}{2} + \frac{1}{2} f(s_1)$, so the whole expression is $\frac{1}{2} + \frac{1}{2} f(s_1) - f(s_0)$; since $1 > f(s_1) > f(s_0)$, this is positive;
- In equation (A.31) for ζ_{01}, if $s_0 \leq s^*$, the inner maximum is solved by $f(s_0)$, hence we have $f(s_0) - f(s_1)$; this is negative, so we need to bound the whole expression below by zero. If $s_0 > s^*$ the inner maximum is solved by $\frac{1}{2} + \frac{1}{2} f(s_0)$. We then have $\frac{1}{2} + \frac{1}{2} f(s_0) - f(s_1)$, which also could be negative, since $f(s_1)$ could be bigger than 1 (if $s_1 < s^*$) and anyway is bigger than $f(s_0)$.

The expression $2\,\zeta_{11} \leq \zeta_{01} + \zeta_{10}$ can now be written as

$$\begin{aligned}
\max\{0, 1 - f(s_1)\} \leq\ & \min\left\{1, \max\{f(s_1) - f(s_0), \frac{1}{2} + \frac{1}{2} f(s_1) - f(s_0)\}\right\} \\
& + \max\left\{0, \frac{1}{2} + \frac{1}{2} f(s_0) - f(s_1)\right\}
\end{aligned}. \tag{A.35}$$

If the maximum on the left side of (A.35) is zero, the inequality is satisfied and we have that both conditions in (A.33) are true. Let us then assume that the second maximum on the left side is positive. This is equivalent to $f(s_1) < 1$, and so we now maintain this assumption. This in turn ensures that the minimum in the first term on the right side of (A.35) is not 1 and the first maximum is realized by its second term, and so we have

$$1 - f(s_1) \leq \left[\frac{1}{2} + \frac{1}{2} f(s_1) - f(s_0)\right] + \max\left\{0, \frac{1}{2} + \frac{1}{2} f(s_0) - f(s_1)\right\}.$$

To prove the second condition in (A.33) it then suffices to show that this inequality holds no matter which term in the final maximum is larger, which is equivalent to

$$1 - f(s_1) \leq \begin{cases} \frac{1}{2} + \frac{1}{2}f(s_1) - f(s_0) \\ \\ 1 - \frac{1}{2}f(s_1) - \frac{1}{2}f(s_0). \end{cases} \tag{A.36}$$

The second of these simplifies to $0 \leq (f(s_1) - f(s_0))$, which is always true. We thus need to check the first, which is

$$\frac{1}{2} \leq \frac{3}{2}f(s_1) - f(s_0),$$

or, equivalently,

$$\frac{1}{2}(1 - f(s_1)) \leq f(s_1) - f(s_0).$$

Remember, however, that we are considering the case when 0 is larger than $\frac{1}{2} + \frac{1}{2}f(s_0) - f(s_1)$, and hence $f(s_0) < 2f(s_1) - 1$, which is equivalent to $f(s_1) - f(s_0) > 1 - f(s_1)$. Since we are considering the case $f(s_1) < 1$, we then have

$$\frac{1}{2}(1 - f(s_1)) < 1 - f(s_1) < f(s_1) - f(s_0),$$

which is the first in (A.36). Therefore both conditions in (A.33) are satisfied.

Hence, in the interval $\lambda \in [0, \frac{1}{2}]$, for a fixed s_1 and t, we have $\partial s/\partial\lambda \geq 0$. This in turn ensures, given that $\dot{s_1}(t)$ is increasing in s, that for a fixed s_1 and t, the smaller is λ, the smaller is $\dot{s_1}(t)$.

Step 3*:* Now consider a $\lambda \leq \frac{1}{2}$ such that the path of $s_1(t)$ converges to 0, the good equilibrium and take a smaller value λ'.

At time 0 and initial condition $s_1(0) = 0$, common for both λs, we now know that $s(0, \lambda) > s(0, \lambda')$. Hence the path of s_1 induced by λ' is initially below the path induced by λ.

If the former path always remained below the latter, it would also converge to 0.

By contradiction, suppose it does not converge to 0. Then there must be a (finite) t such that the path induced by λ' crosses, from below, the path induced by λ. At that t, $s_1(t, \lambda') = s_1(t, \lambda)$. Hence, given t and this value for s_1, we have that

$$\frac{ds_1(t, \lambda')}{dt} < \frac{ds_1(t, \lambda)}{dt}.$$

Hence, the path induced by the smaller value λ' is always either below or being pushed below that induced by λ, and so the path induced by λ' also converges to 0. A similar argument shows that if the path of $s_1(t)$ induced by λ converges to s_b, then so does the path induced by any $\lambda > \lambda'$.

Step 4*:* Now consider [5.2]. Suppose we have two paths, $s_1(t, \lambda_1)$ and $s_1(t, \lambda_2)$, with $\frac{1}{2} \geq \lambda_1 > \lambda_2$, both converging to s_u. Our previous steps show that the first path (associated to the larger λ) must always lie at least weakly above the second path (associated to the smaller λ). Using Lemma A.2 we then have a contradiction and hence the proof is now complete. ■

A.6. Proof of Corollary 1

Proposition 4 established that for any $\lambda > 0$ there exists a $q^* > 0$ such that, for any $q \leq q^*$, the system converges to s_b. Defining $\hat{\lambda} = 1 - \lambda$, this also means that for any $\hat{\lambda} < 1$ there exists a q^* such that, for any $q \leq q^*$, the system converges to s_b. This is the claim we wanted to establish. ■