Supplementary Information for:

The social leverage effect: Institutions transform weak reputation effects into strong incentives for cooperation

In this document, we introduce a model of reputation-based first- and second-order cooperation. Second-order cooperation consists in contributing to an institution, that then produces additional incentives for first-order cooperation. We introduce the main elements of our model in section 1, and delve into further technical detail in section 2. We define the **institution equilibrium** of our model in this section, an equilibrium in which some individuals contribute to the institution. We characterize the institution equilibrium in three steps, over sections 2-4. We notably compute the equilibrium level of first- and second-order cooperation in section 3, and the necessary and sufficient conditions for this equilibrium in section 4. In section 5, we establish a **baseline equilibrium** for comparison to the results of the institution equilibrium. Finally, we motivate and explain the computation of a **numerical solution** to our model in section 6.

Contents

1 Main elements of the model

In this section, we introduce the main elements of our model, without delving into technical detail (these details are explored in the next section). We notably explain the general structure of our model (section 1.1), how we model reputation (section 1.3), and how we model the institution (section 1.4).

1.1 General structure

We consider interactions between two types of players: a large number $n \gg 1$ of actors, and infinitely many choosers. To avoid lengthy repetitions of the terms chooser and actor, we have assigned a gender to each type of player based on the result of a coin toss. Throughout this text, we will use feminine pronouns (she/her) to refer to actors, and masculine pronouns to refer to choosers (he/him/his).

Our model is composed of n infinitely repeated games—one for each actor. These repeated games occur in parallel and in discrete time, in rounds indexed by the letter $t \in \mathbb{N}$. We refer to all these interactions as the repeated game, or simply the game, and to the interactions that only concern a specific actor as the actor's repeated game. In every round, n choosers are drawn from the infinite chooser population, and assigned to a different actor. At the end of the round, the n actors move on to the next round of their repeated game, while the n choosers for that round exit the game. In other words, actors are long-lived, and play all rounds of their repeated game, while choosers are short-lived, and play only one round of interaction (if and when they are assigned to an actor). We complete the description of our game in section 1.2, by describing the stage game of a generic actor.

Choosers decide whether or not to trust the actor they are assigned to, based on limited information—her reputation, as defined in section 1.3. Their role is to motivate cooperation by actors.

Actors face two types of interactions, and two types of cooperative decisions. Sometimes they can pay to reciprocate a chooser's trust—first-order cooperation—and sometimes they can pay to contribute to an institution, whose purpose is to incentivize reciprocation by every actor—second-order cooperation. The functioning of the institution, which depends on individual contributions, is explained in section 1.4.

To capture the fact that individuals will not all equally be motivated to cooperate, we assume that actors have varying time preferences. Every actor is characterized by a discount factor, which is drawn before the onset of the game. The value of a given actor's discount factor, which we note δ, is hidden to other players. Throughout their repeated game, actors discount future payoffs according to this value. The lifetime payoff of an actor of discount factor δ is equal to the payoff earned in the initial round ($t = 0$), plus δ times the payoff earned in the next round ($t = 1$), plus δ^2 times the payoffs earned in the round after that ($t = 2$), and so on. In section 3, we show that, in our main equilibrium, patient actors (high δ) engage in both forms of cooperation, while impatient actors engage in neither form of cooperation, and actors of intermediate patience only engage in the cheapest form of cooperation.

1.2 Stage game

In this section, we consider a generic actor (without specifying the value of her discount factor). We describe the stage game of this generic actor's infinitely repeated game. This stage game is illustrated by Figure 1. As mentioned above, it is infinitely repeated, for each of the rounds $t \in \mathbb{N}$.

Each round proceeds as follows. First, nature draws between two types of interaction: a trust game with probability q, and the institution game with probability $1 - q$. Then, the actor and the chooser she has been assigned that round play according to the rules of the interaction at hand.

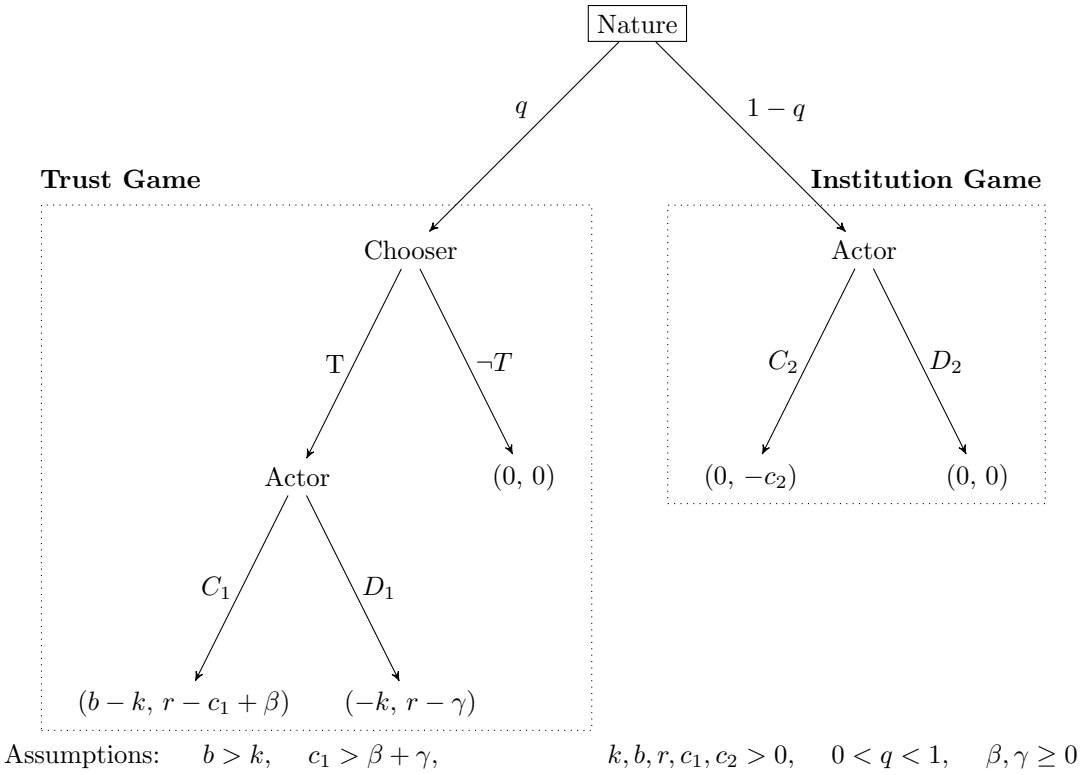

Figure 1: **Stage game**. Nature begins by setting the interaction type: a trust game is drawn with probability q, and the institution game is drawn with probability $1-q$. In the case of a **trust game** (left branch), the actor play an asymmetric prisoner's dilemma with the chooser she has been assigned this round. We assume $b > k$ and $c_1 > \beta + \gamma$ to keep the structure of an asymmetric prisoner's dilemma. For the chooser, playing T instead of $\neg T$ is net beneficial if the actor subsequently plays C_1, but net costly if the actor subsequently plays D_1. For the actor, playing C_1 instead of D_1 is always net costly, despite the effect of the institution. The institution is materialized here by a reward $\beta \geq 0$ granted in the case that the actor plays C_1, and a penalty $\gamma \geq 0$ inflicted in the case that the actor plays D_1. In the **institution game** (right branch), the actor decides whether or not to contribute to the institution by paying $c_2 > 0$. This decision is made in parallel with every other actor who has drawn the institution game that round (we detail the functioning of the institution in section 1.4).

If a trust game is drawn, both players play an asymmetric prisoner's dilemma with two steps. In the first step, the chooser decides whether to trust (i.e. play action T) or not trust ($\neg T$) the actor, putting an early end to the interaction. If he trusts her, the actor then decides whether to reciprocate (C_1) that trust, or cheat (D_1), in the second step. We refer to reciprocation as first-order cooperation, and to cheating as first-order defection (or simply cooperation and defection when there is no ambiguity), which is why we use the labels C_1 and D_1 to designate these two actor actions.

Trust costs $k > 0$ to the chooser, and brings benefit $r > 0$ to the actor. First-order cooperation (reciprocation) costs $c_1 > 0$ to the actor, and brings benefit $b > 0$ to the chooser. We assume $b > k$: the chooser benefits from trusting the actor if she subsequently reciprocates that trust.

In the institution game, the actor plays without the chooser. She can either contribute (C_2) to an institution, whose functioning is described in section 1.4, or free-ride (D_2). We refer to contribution as second-order cooperation, and to free-riding as second-order defection, which is why we use the labels C_2 and D_2 to designate these other two actor actions. Second-order cooperation (contribution) costs $c_2 > 0$ to the actor.

1.3 Reputation

In this section, we continue to consider a generic actor. Throughout her repeated game, the actor faces infinitely many different short-lived choosers, who interact with her with probability q in the relevant round. In any given round, we refer to an actor's current co-player as the current chooser (or simply the chooser, when there is no ambiguity), and to her co-player in the next round as the next chooser.

We restrict the information available to choosers in the following manner. At the end of each round, we assume that the next chooser observes the actor's action in that round with baseline probability p_1 if the actor faced the trust game and was trusted by the current chooser (this probability can be increased through the effect of the institution; see section 1.4), and with fixed probability p_2 if the actor faced the institution game ($0 < p_1 \leq 1$, $0 < p_2 \leq 1$).

We assume that choosers do not observe the actor's behavior in rounds before the one that just ended, and do not observe the behavior of previous choosers.

What this means is that in any round $t \geq 0$, when the current chooser faces the trust game, and therefore the option to trust or not trust the actor, he can be in one of five situations. If the actor did not play in round $t-1$ (because $t = 0$, or because in round $t-1$ she faced the trust game and was not trusted) or if her action was not observed, the chooser does not have access to any information. Otherwise, the chooser has access to one piece of information, pertaining to the actor's action in round $t-1$.

We refer to these five cases as the actor's **reputation**, or, interchangeably, as the information available to the (current) chooser. (Note that the actor's reputation is defined with respect to her current partner.) We note $\mathcal{R} \equiv \{\emptyset, \mathcal{C}_1, \mathcal{D}_1, \mathcal{C}_2, \mathcal{D}_2\}$ the set of possible actor reputations, $\emptyset$ referring to the case of an empty reputation (i.e., the case when the chooser in a given round has no information), $\mathcal{C}_1$ referring to the case when the chooser has observed the actor playing C_1 in the previous round, and so on. We note $\mathcal{R}^* \equiv \mathcal{R} \setminus \{\emptyset\}$ the set of non-empty reputations.

1.4 Mechanism of the institution

In this section, we consider the entire actor population, and the entire repeated game.

An institution collects the expected contribution of the actor (in the institution game), and transforms it into incentives for first-order cooperation (in the trust game). The institution considered here is not a player; for a given set of parameter values, its functioning is fixed. However, it relies on the actor's behavior in the institution game: if the actor never contributes, the institution cannot provide any incentives for first-order cooperation. As detailed below (see section 2), the actor's strategy is allowed to vary with her discount factor and her current reputation. A priori, her behavior is probabilistic: knowing her strategy, her reputation and the population distribution of discount factors (but not the actor's personal discount factor), one can compute the probability that the actor will play C_2 when faced with the institution game.

We note f_2 that probability. In a given round, the institution receives an amount $(1-q)f_2c_2$ in expectation—this expected contribution being calculated at the beginning of a round (before either game is drawn), knowing the actor's strategy, her reputation that round, and the population distribution of discount factors.

In any given round, we assume that the institution receives this expected contribution with certainty. Remember that we have an infinite population in mind, as explained in section 1.1. With an infinite population of actors, each round would see a fraction $(1-q)f_2$ of the total population pay c_2 to contribute to the institution. (Note that we will show that the actor's strategy is stationary in every subgame perfect equilibrium, i.e. that her behavior does not depend on her current reputation. Were we to extend to an infinite-player model, we could define f_2 in such an equilibrium without having to keep track of each individual's reputation.)

We take the amount received by the institution, multiply it by a factor $\rho > 0$, and split the result $\rho(1-q)f_2c_2$ between three types of incentives: a reward for cooperation $\beta \geq 0$, a penalty for defection $\gamma \geq 0$, and an increase $\pi_1 \geq 0$ in the baseline probability of observation in the trust game. In the same round, if the actor faces the trust game and is trusted by the chooser, she earns total payoff $r - c_1 + \beta$ if she plays C_1, $r - \gamma$ if she plays D_1, and is observed with total probability $p_1 + \pi_1$. We assume that:

$$\rho(1-q)f_2c_2 = q(\beta + \gamma + c_1\pi_1) \tag{1.1}$$

To interpret this equation, remember again that we have an infinite population in mind. With an infinite population of actors, each round would see a fraction q of trust games which can be incentivized by the funds from the multiplied amount $\rho(1-q)f_2c_2$. We assume that incentives produced by the institution apply equally to every trust game, and that their sum is equal to this multiplied amount (a factor of conversion c_1 is applied to the probability π_1).

Note that we use Greek letters to refer to the institution and the incentives it creates throughout the model. ρ is a measure of the institution's efficiency: for every dollar in total contribution, ρ dollars are created to incentivize first-order cooperation. β, γ and π_1 are left unspecified: with this general model, we can consider different types of institutions. For instance, a purely punishing institution is obtained by taking $\beta = \pi_1 = 0$; in that case, the total contribution is entirely allocated to punishing defectors, who are inflicted a penalty of $\gamma = \rho f_2 c_2(1-q)/q$. A purely monitoring institution is obtained by taking $\beta = \gamma = 0$; in that case, the probability of observation in the trust game increases by $\pi_1 = \rho f_2(c_2/c_1)(1-q)/q$.

Accounting for the effect of the institution in a given round, the net cost of cooperation is equal to the total payoff of defectors minus the total payoff of cooperators, that is: $(r-\gamma) - (r - c_1 + \beta) = c_1 - (\beta + \gamma)$. We assume that, even after accounting for the effect of the institution, cooperation remains costly for actors, that is:

$$c_1 - (\beta + \gamma) > 0$$

In addition, we assume that the likelihood of observation in the trust game remains below 1, i.e. that: $p_1+\pi_1 \leq 1$.

2 Technical assumptions and strategy space

In this section, we go into further technical detail, and rigorously define histories and the strategy space. Importantly, we assume that the chooser behaves as if many rounds of the game have already been played, and introduce his long-run payoffs (section 2.1.5). Using general arguments and relatively short demonstrations, we show that there can only be three types of subgame perfect equilibria: an uncooperative equilibrium, in which actors never cooperate nor contribute; an institution equilibrium, in which actors cooperate and contribute with positive probability; and non-institution equilibria, in which only first-order cooperation occurs with positive probability.

2.1 History equivalence classes and strategies

2.1.1 Chooser history equivalence classes and strategy space

The chooser only plays in rounds in which the trust game is drawn. Because we strongly restrict the information available to the chooser, chooser histories of the repeated game can be divided in five equivalence classes, depending on the actor's reputation in the eyes of the (current) chooser. We note $\mathcal{H}_{ch} \mid R$ the equivalence class attained when the actor's reputation is $R \in \mathcal{R}$, and note $\mathcal{H}_{ch} \mid \mathcal{R}$ the set comprised of the five equivalence classes for histories of the repeated game; the set of chooser histories $\mathcal{H}_{ch}$ is the union of those equivalence classes.

For simplicity, we equate $\mathcal{R}$ with $\mathcal{H}_{ch} \mid \mathcal{R}$. That is, we define chooser strategy directly as a function of actor reputation, rather than as a function of the history equivalence class. A pure strategy for a chooser specifies whether to trust or not trust the actor depending on her reputation; it is a map:

$$\sigma_{ch} : \mathcal{R} \to \{T, \neg T\}$$

We restrict to the set of chooser strategies $\mathcal{S}_{ch}$ which is pure for non-empty reputations, i.e. the set of strategies following which the chooser plays either T or $\neg T$ with certainty given any information $R \in \mathcal{R}^*$. We note $\sigma^*_{ch} \equiv \sigma_{ch} \mid_{\mathcal{R}^*}$ the restriction of a chooser's strategy to the non-empty information set. There are $2^4 = 16$ possible values for σ^*_{ch}, and an infinite number of possible chooser strategies since we allow choosers to mix between T and $\neg T$ given $\emptyset$. When the chooser plays according to a strategy $\sigma_{ch} \in \mathcal{S}_{ch}$, we note θ the probability that she trusts given $\emptyset$; a chooser's strategy is completely described by σ^*_{ch} and $\theta \in [0, 1]$.

We thus allow the chooser to mix given $\emptyset$. We return to this issue in section **??** in which we calculate the value of θ in equilibrium. Our calculation shows that restricting to pure strategies would lead us not to consider certain equilibria—under certain parameter conditions, the chooser benefits from deviation to trusting given $\theta = 0$ and from deviation to trusting given $\theta = 1$ (because the value of θ influences actor strategy). In fact, the equilibrium value of θ will be an important value, capturing the baseline level of trust in that equilibrium (see section 6).

2.1.2 Actor strategy space

The actor does not always have an opportunity to act. In each round, there are three possibilities: either the trust game is drawn and the chooser plays T, in which case the actor has an opportunity to play C_1 or play D_1; or the trust game is drawn and the chooser plays $\neg T$, in which case the actor does not play this round; or the institution game is drawn, in which case the actor has an opportunity to play C_2 or play D_2. We note $\mathcal{T}$, $\neg\mathcal{T}$ and $\mathcal{I}$ the corresponding events, in the above order.

We restrict actor strategy space in accordance to the restriction applied to chooser strategy space, taking into account only what is relevant to the chooser once either event $\mathcal{T}$ or $\mathcal{I}$ has occurred, and the actor decides between playing C_1 and D_1, or C_2 and D_2, respectively. In other words, since choosers play only according to reputation, we do not need to consider the entire set of possible histories for the actor; we only need to consider elements of the set $\mathcal{R} \times \{\mathcal{T}\}$, and elements of the set $\mathcal{R} \times \{\mathcal{I}\}$. (Since the actor does not play after event $\neg\mathcal{T}$, we do not define actor strategy following that event).

A pure strategy for the actor specifies whether to reciprocate or cheat after being trusted in the trust game, and whether to contribute or free-ride in the institution game, depending on the actor's (current) reputation, and her

(fixed) discount factor; it is comprised of two maps:

$$\begin{aligned}\sigma_{act}: \ & \mathcal{R} \times \{\mathcal{T}\} \times (0,1) \to \{C_1, D_1\} \\ & \mathcal{R} \times \{\mathcal{I}\} \times (0,1) \to \{C_2, D_2\}\end{aligned}$$

We restrict to the set $\mathcal{S}_{act}$ of pure strategies for the actor.

2.1.3 Actor initial reputation

Recall that we have precisely defined the actor's reputation in any given round $t \geq 1$, but not in the initial round. In the initial round, we assume that the reputation of an actor of discount factor $\delta \in (0,1)$ corresponds to $\sigma_{act}(\emptyset, \mathcal{T}, \delta)$ with probability ε; to $\sigma_{act}(\emptyset, \mathcal{I}, \delta)$ with probability ε; and that it is equal to $\emptyset$ with probability $1 - 2\varepsilon$. Take for instance an actor of discount factor δ who would reciprocate if initially trusted (and given empty reputation), and free-ride if the institution game is initially drawn (and given empty reputation). That actor's initial reputation is $\mathcal{C}_1$ with probability ε, $\mathcal{D}_2$ with probability ε, and $\emptyset$ with probability $1 - 2\varepsilon \approx 1$. Note that we show below that the actor's strategy is stationary (Lemma 2.4)—the fact that this definition relies on behavior given $\emptyset$ (rather than another value for the actor's initial reputation) is moot.

2.1.4 Continuation strategy profile

For every $R \in \mathcal{R}$, the continuation game associated with R is defined as the infinitely repeated game in which the chooser initially has information R, corresponding to the actor's initial reputation. The continuation game associated with R occurs each time the actor attains reputation R at the end of the previous round.

In the continuation game associated with R, the chooser plays directly after. For every strategy profile σ, we note $\sigma\,|_R$ the continuation strategy profile induced by R.

The actor plays after histories of the form $\{R, \mathcal{T}\}$ and $\{R, \mathcal{I}\}$. For every strategy profile σ, and every $(R, \mathcal{X}) \in \mathcal{R} \times \{\mathcal{T}, \mathcal{I}\}$, we note $\sigma\,|_{R,\mathcal{X}}$ the continuation strategy profile induced by $(R, \mathcal{X})$.

2.1.5 Chooser long-run payoffs

For every σ and R, we note $u(\sigma\,|_R)$ the expected payoff of the chooser in the continuation game. This is the payoff that the chooser can expect to gain in the current round, given that the trust game is drawn, when players play according to σ, and the chooser has information R on the actor.

As we will see in more detail, the actor's reputation follows a Markov process. Given any strategy profile σ, the actor's reputation in a given round $t \geq 1$ is a function of the prescribed behavior of the chooser and the actor in round $t - 1$. We can then describe the actor's reputation in a given round $t \geq 1$ as a probability distribution over $\mathcal{R}$, depending on σ and the distribution of actor time preferences—in fact, we give this precise description for the two equilibria of interest in sections 3 and 5, and show that the reputation of any given actor of patience δ reaches a stationary state.

An important issue for our model is that the value of information for the chooser can depend on the round number t. This is an issue because we do not allow the chooser to observe t—instead, we assume that he makes decisions based solely on the information $R \in \mathcal{R}$ at his disposal. To model the chooser's decisions in an evolutionary equilibrium, we need a measure of her payoffs that does not depend on the round number. For every σ and R, we note $u^\infty(\sigma\,|_R)$ the **chooser long-run payoffs**, that is, the expected payoff of the chooser in the continuation game assuming that many rounds of the game have already been played, and that the actor's reputation can be approximated using the steady state of the Markov chain.

As we will show, in our two equilibria of interest, the value of a non-empty reputation $R \in \mathcal{R}^*$ is not a function of t, but the value of $\emptyset$ is. Our approximation—considering chooser payoffs in the long-run, i.e. assuming that the chooser plays after a large number of rounds $t \gg 1$ with quasi-certainty—is thus only needed for the empty reputation $\emptyset$.

Intuitively, since actor strategy is stationary (Lemma 2.4 below), every non-empty reputation gives information about the actor's discount factor δ, and therefore her future actions. Since $\mathcal{C}_1$ and $\mathcal{D}_1$ perfectly indicate whether an actor cooperated or defected in the past, in both equilibria of interest, it always pays to trust given $\mathcal{C}_1$ and distrust given $\mathcal{D}_1$. In addition, since the actor's access to the institution game is not constrained by her past reputation (she does not need to be 'trusted' by the institution to free-ride or contribute), $\mathcal{C}_2$ and $\mathcal{D}_2$ always give the same information about her δ, and are thus stationary predictors of whether her δ is such that she will cooperate or defect in the future.

In contrast, whether or not an actor's reputation is empty depends on her past reputation: if she previously defected and was observed, achieving reputation $\mathcal{D}_1$, she is certain not to be trusted in the next round in equilibrium—when an actor who achieved reputation $\mathcal{C}_1$ will be trusted again with probability q, if the trust game is drawn again. Over time, $\emptyset$ comes to yield information about the actor: whereas in round 0, cooperators and defectors are equally as likely to have an empty reputation, in any round $t \geq 1$, defectors are more likely to have an empty reputation.

2.1.6 Chooser inferences given the null event

Null reputations are also an issue. Certain reputations can occur with null probability, whether during the game (in a given round t), or in the steady state (which following the above assumption will be the decisive case). If for instance the actor always reciprocates the chooser's trust, and always contributes to the institution, when given the chance, whatever her discount factor and her current reputation, then $\mathcal{D}_1$ and $\mathcal{D}_2$ are both null events.

In such a case, we assume that choosers do not learn any information. Given a reputation that occurs with probability 0 (at a given point in time t or in the steady state), the actor's action in the trust game is drawn according to the distribution of possibilities (at that same point in time). In the example above, the actor can only have reputation $\mathcal{C}_1$, $\mathcal{C}_2$ or $\emptyset$ at a given point in time. We then assume that the chooser behaves as if the actor would also reciprocate given the two null reputations $\mathcal{D}_1$ and $\mathcal{D}_2$; he would therefore trust. Given null information, the chooser behaves according to the distribution of possible behaviors over all possible reputations and discount factors.

As we will see, this assumption is of little effect—it leads us to take a conservative view of the domain of existence of the baseline equilibrium below (Proposition 5.3), without this having any effect on the outputs of our model (which are all null outside of the conservative domain of existence).

2.1.7 Actor continuation payoff

The actor earns payoffs throughout the game. When the actor's discount factor is δ, we normalize her lifetime payoffs by multiplying payoffs in each round by $(1-\delta)$. For every δ, σ and R, we note $U_\delta(\sigma \mid_R)$ the lifetime expected payoff of the actor starting from the continuation game associated with R. Since the actor begins with empty reputation, $U_\delta(\sigma) \equiv U_\delta(\sigma \mid_\emptyset)$ is the actor's expected payoff over the entire game.

In addition, we define two other classes of continuation payoffs for the actor, relevant to the histories after which she actually plays, in the trust and institution game respectively. For every δ, σ and R, we note $U_\delta(\sigma \mid_{R,\mathcal{T}})$ the lifetime expected payoff of the actor given history $(R, \mathcal{T})$, and $U_\delta(\sigma \mid_{R,\mathcal{I}})$ the lifetime expected payoff of the actor given history $(R, \mathcal{I})$. These correspond to the lifetime's payoff of the actor in the continuation game associated with R, once even $\mathcal{T}$ or $\mathcal{I}$ has occurred (hence not comprising the benefit of being trusted by the chooser in the first case).

2.1.8 Objective and equilibrium concept

A strategy profile $\sigma = (\sigma_{ch}, \sigma_{act})$ is a Nash equilibrium of the repeated game if both players' strategy is a best response to the other's, i.e. if:

$$\begin{aligned} &\forall \sigma'_{ch} \in \mathcal{S}_{ch}, && u^\infty(\sigma) \geq u^\infty(\sigma'_{ch}, \sigma_{act}) \\ &\forall \sigma'_{act} \in \mathcal{S}_{act}, \forall \delta \in (0,1), && U_\delta(\sigma) \geq U_\delta(\sigma_{ch}, \sigma'_{act}) \end{aligned}$$

In lieu of considering all possible Nash equilibria, we consider a more restrictive equilibrium concept—namely, subgame perfection. A strategy profile $\sigma = (\sigma_{ch}, \sigma_{act})$ is a subgame perfect equilibrium of the repeated game if:

$$\begin{aligned} &\forall \sigma'_{ch} \in \mathcal{S}_{ch}, \forall R \in \mathcal{R}, && u^\infty(\sigma \mid_R) \geq u^\infty((\sigma'_{ch}, \sigma_{act}) \mid_R) \\ &\forall \sigma'_{act} \in \mathcal{S}_{act}, \forall R \in \mathcal{R}, \forall \delta \in (0,1), && U_\delta(\sigma \mid_{R,\mathcal{T}}) \geq U_\delta((\sigma_{ch}, \sigma'_{act}) \mid_{R,\mathcal{T}}) \\ & && U_\delta(\sigma \mid_{R,\mathcal{I}}) \geq U_\delta((\sigma_{ch}, \sigma'_{act}) \mid_{R,\mathcal{I}}) \end{aligned}$$

A Nash equilibrium is subgame perfect if, for every possible continuation game, the induced strategy profile is a Nash equilibrium—even when considering unrealized histories; that is, histories which occur with null probability. Here, seeing the restricting assumptions we have made on histories and therefore strategy space, a subgame perfect equilibrium is a strategy profile such that there are no profitable deviations for either player: (i) even when considering reputations that occur with null probability, because the actor never accomplishes a certain actions—e.g., $\mathcal{C}_1$ given that the actor always defects; and (ii) even when considering unrealized combinations of reputation and event $\mathcal{T}$, because the chooser never trusts given certain reputations—e.g. $(\mathcal{D}_1, \mathcal{T})$ given that the chooser does no trust given information $\mathcal{D}_1$.

Objective: Our goal is to characterize the set $\mathcal{S}$ of strategy profiles which:

(i) belong the set $\mathcal{S}_{ch} \times \mathcal{S}_{act}$, following our restrictive assumptions, and

(ii) are subgame perfect equilibrium of the repeated game.

By restricting to subgame perfect equilibria, we restrict to Nash equilibria which are stable to trembles, that is to either player mistakenly playing an unprescribed action with a small, positive probability (Selten, 1983). Arguably, this is the relevant concept when one is interested in endpoints of an evolutionary process—assuming that mistakes or misunderstandings will occur with non-null probability.

Note that restricting to subgame perfect equilibria also leads to getting rid of certain functionally equivalent Nash equilibria. For instance, consider the *uncooperative* strategy profile, defined as the strategy profile whereby: (i) the chooser always plays $\neg T$, whatever the history $R \in \mathcal{R}$; and (ii) the actor always plays D_1 in the trust game, whatever the history $(R, \mathcal{T}) \in \mathcal{R} \times \{\mathcal{T}\}$, and always plays D_2 in the institution game, whatever the history $(R, \mathcal{I}) \in \mathcal{R} \times \{\mathcal{I}\}$. This strategy profile is always a Nash equilibrium because trust is assumed to be costly for the chooser, and first- and second-order cooperation are assumed to be costly for the actor (it is in fact subgame perfect for those reasons). Yet, there are many neutral deviations available. The chooser may for instance deviate to playing T given history $\mathcal{C}_1$. Since $\mathcal{C}_1$ occurs with null probability when the actor plays according to (ii), this unilateral deviation is payoff-neutral. Similarly, $\mathcal{T}$ occurs with null probability when the chooser plays according to (i). The actor may deviate to playing C_1 given history $(\mathcal{D}_1, \mathcal{T})$ without affecting her payoffs (or in fact given any history of the form $(R, \mathcal{T})$ and any discount factor). In both cases, the obtained strategy profile is also a Nash equilibrium, which is functionally equivalent to the one under consideration.

However, only the uncooperative strategy profile is subgame perfect. For instance, the functionally equivalent Nash equilibrium obtained when the actor cooperates given $(\mathcal{D}_1, \mathcal{T})$ instead of defecting is not subgame perfect: given history $(\mathcal{D}_1, \mathcal{T})$, the actor strictly benefits from deviation back to defecting.

Finally, for simplicity of notations, we assume that the actor plays C_1 if she is indifferent between C_1 and D_1, and C_2 if she is indifferent between C_2 and D_2. In sections 3 and 5, we show that the actor can only be indifferent between two such options if her discount rate takes a precise value, which occurs with probability 0, since discount factors are distributed continuously. We also assume that the chooser plays T if she is indifferent between T and $\neg T$. In section 4 and 5, we show that such indifference similarly occurs with probability 0. Note that we do not assume that the chooser plays T given

2.2 Three types of subgame perfect equilibrium

2.2.1 Condition for trust

We begin by deriving a general condition for trust.

Lemma 2.1: Condition fo inferring trust

$\forall \sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}, \forall R \in \mathcal{R}$,

$$\sigma_{ch}(R) = T \iff \mathbf{P}^{\infty}(\sigma_{act}(\mathcal{T}, R, \delta) = C_1 \mid R, \delta \sim \Delta) \geq \frac{k}{b} \tag{2.1}$$

In any subgame perfect equilibrium, the chooser trusts given sufficiently good predictors of the actor's cooperation, and distrusts given sufficiently bad predictors of the actor's cooperation.

Proof. Let us consider a subgame perfect equilibrium $\sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}$, and a certain reputation $R \in \mathcal{R}$.

Since the actor's strategy given $\mathcal{T}$ and R depend on the value of her discount factor δ, and since her δ remains hidden, for any $t \geq 1$, we integrate over the entire distribution of discount factors Δ to calculate the probability that she cooperates given that she is trusted and that her reputation is R. We note this probability:

$$\mathbf{P}^{t}(\sigma_{act}(\mathcal{T}, R, \delta) = C_1 \mid R, \delta \sim \Delta)$$

As explained above in section 2.1.5, this probability can depend on the round number t, and will in fact depend on t in the equilibria of interest when $R = \emptyset$. We note $\mathbf{P}^{\infty}(\sigma_{act}(\mathcal{T}, R, \delta) = C_1 \mid R, \delta \sim \Delta)$ the obtained value in the reputational steady state.

On average, the chooser obtains long-run payoff $-k + \mathbf{P}^\infty(\sigma_{act}(\mathcal{T}, R, \delta) = C_1 \mid R, \delta \sim \Delta) \times b$ if he trusts, in which case he pays k and receives b in return if and only if the actor plays C_1. If he distrusts, he gains payoff 0 with certainty.

In a subgame perfect equilibrium, the chooser thus trusts given R if and only if:

$$-k + \mathbf{P}^\infty(\sigma_{act}(\mathcal{T}, R, \delta) = C_1 \mid R, \delta \sim \Delta) \times b \geq 0$$
$$\mathbf{P}^\infty(\sigma_{act}(\mathcal{T}, R, \delta) = C_1 \mid R, \delta \sim \Delta) \geq \frac{k}{b}$$

□

Note that we explicitly condition on the distribution of discount factors Δ in the above probability. This is to be explicit about the fact that this probability, which involves the letter δ on the left, is not calculated for any specific discount factor value, but over the entire distribution. Every probability in this document is taken given the distribution Δ—we only keep the notation $\mid \delta \sim \Delta$ in those cases where this type of ambiguity arises.

Note also that we indicate with a superscript when a probability depends on the round number t, or is taken in the steady state ∞. When no superscript is mentioned, the probability does not depend on time (and only depends on the distribution Δ). Condition (2.1) depends on the steady state value of this probability, because we have assumed that the chooser behaves so as to optimize his long-run payoffs (section 2.1.5). In section 4, we show that every non-empty reputation is a stationary predictor of the actor's behavior, and we explain how we determine the equilibrium value of θ, the probability that the chooser trusts given empty reputation.

2.2.2 Conditions for cooperation

We derive two general conditions for first- and second-order cooperation, respectively.

Lemma 2.2: Condition for first-order cooperation

$\forall \sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}, \forall R \in \mathcal{R}, \forall \delta \in (0,1)$,

$$\sigma_{act}(R, \mathcal{T}, \delta) = C_1 \iff c_1 - (\beta + \gamma) \leq \delta \times (p_1 + \pi_1) \times (U_\delta(\sigma \mid_{\mathcal{C}_1}) - U_\delta(\sigma \mid_{\mathcal{D}_1})) \tag{2.2}$$

In any subgame perfect equilibrium, whatever her current reputation R, and whatever her discount factor δ, the actor cooperates when given the opportunity to do so if and only if the immediate, net cost of cooperation $c_1 - \gamma - \beta$ is smaller than her future (i.e. multiplied by her δ) benefit of achieving reputation $\mathcal{C}_1$ rather than $\mathcal{D}_1$, if observed (with probability $p_1 + \pi_1$).

Proof. Let us consider a subgame perfect equilibrium $\sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}$. Given any history of the form $(R, \mathcal{T})$, σ_{act} will prescribe playing C_1 if and only if her continuation payoff were she to play C_1 is greater or equal than her continuation payoff were she to play D_1 (recall that we just assumed that the actor plays C_1 if indifferent between both options).

If the actor cheats, she gains $r - \gamma$ in the current round. Her reputation starting in the next round is then $\mathcal{D}_1$ with probability $p_1 + \pi_1$, and $\emptyset$ with probability $1 - (p_1 + \pi_1)$. If she reciprocates, she gains only $r - c_1 + \beta < r - \gamma$, and achieves future reputation $\mathcal{C}_1$ with probability $p_1 + \pi_1$, and, again, $\emptyset$ with probability $1 - (p_1 + \pi_1)$. Following the actor's action, the continuation game associated with $\mathcal{C}_1$, $\mathcal{D}_1$ or $\emptyset$ occurs, depending on the actor's chosen action and the outcome of observation.

Given $(R, \mathcal{T})$, σ_{act} will then prescribe playing C_1 if and only if:

$$r - \gamma + \delta[(p_1 + \pi_1)U_\delta(\sigma \mid_{\mathcal{D}_1}) + (1 - p_1 + \pi_1)U_\delta(\sigma \mid_{\emptyset})] \leq r - c_1 + \beta + \delta[(p_1 + \pi_1)U_\delta(\sigma \mid_{\mathcal{C}_1}) + (1 - p_1 + \pi_1)U_\delta(\sigma \mid_{\emptyset})]$$

Simplifying, this is equivalent to:

$$r - \gamma + \delta \times (p_1 + \pi_1) \times U_\delta(\sigma \mid_{\mathcal{D}_1}) \leq r - c_1 + \beta + \delta \times (p_1 + \pi_1) \times U_\delta(\sigma \mid_{\mathcal{C}_1})$$

Re-arranging, we obtain the proposed inequality:

$$c_1 - (\beta + \gamma) \leq \delta \times (p_1 + \pi_1) \times (U_\delta(\sigma \mid_{\mathcal{C}_1}) - U_\delta(\sigma \mid_{\mathcal{D}_1})) \tag{2.2}$$

□

Lemma 2.3: Condition for second-order cooperation

$\forall \sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}, \forall R \in \mathcal{R}, \forall \delta \in (0,1)$,

$$\sigma_{act}(R, \mathcal{I}, \delta) = C_2 \iff c_2 \leq \delta \times p_2 \times (U_\delta(\sigma\,|_{\mathcal{C}_2}) - U_\delta(\sigma\,|_{\mathcal{D}_2})) \tag{2.3}$$

In any subgame perfect equilibrium, whatever her current reputation R, and whatever her discount factor δ, the actor contributes when given the opportunity to do so if and only if the immediate cost of second-order cooperation c_2 is smaller than her future (i.e. multiplied by her δ) benefit of achieving reputation $\mathcal{C}_2$ rather than $\mathcal{D}_2$, if observed (with probability p_2).

Proof. The proof is analogous to the one above. For any $\sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}$, given any history of the form $(R, \mathcal{I})$, the actor gains 0 if she plays D_2 in the current round, following which she attains reputation $\mathcal{D}_2$ with probability p_2. If she plays C_2, she pays c_2, but instead attains reputation $\mathcal{C}_2$ with probability p_2. Whatever her action, she attains empty reputation with probability $1 - p_2$.

Comparing the continuation payoffs in both cases (similarly to before, $\delta \times (1 - p_2) \times U_\delta(\sigma\,|_\emptyset)$ is a factor in both continuation payoffs, and cancels out), we deduce that σ_{act} will prescribe playing C_2 if and only if:

$$0 + \delta \times p_2 \times U_\delta(\sigma\,|_{\mathcal{D}_2}) \leq -c_2 + \delta \times p_2 \times U_\delta(\sigma\,|_{\mathcal{C}_2})$$

Re-arranging, we immediately obtain the proposed inequality. □

2.2.3 Actor strategy is stationary

Using Lemmas (2.2-2.3), we deduce that the actor's strategy must be stationary.

Lemma 2.4: Actor strategy is stationary

$\forall \sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}, \forall (R, R') \in \mathcal{R}^2, \forall \delta \in (0,1)$,

$$\sigma_{act}(R, \mathcal{T}, \delta) = \sigma_{act}(R', \mathcal{T}, \delta)$$
$$\sigma_{act}(R, \mathcal{I}, \delta) = \sigma_{act}(R', \mathcal{I}, \delta)$$

In any subgame perfect equilibrium, the actor's strategy is stationary—she cooperates and contributes depending only on the value of her discount factor δ, and not on the value of her current reputation R (i.e. the history equivalence class).

Proof. This immediately from the above two lemmas, as the actor's current reputation R is absent from both equations (2.2) and (2.3). In a subgame perfect equilibrium, using the reasoning in Lemma 2.2, we deduce that the actor plays C_1 or D_1 given the opportunity to do so (i.e. given $\mathcal{T}$), depending solely on the value of her discount factor δ, and on the lifetime benefit of achieving reputation $\mathcal{C}_1$ rather than reputation $\mathcal{D}_1$—which does not depend on her current reputation R (but instead, notably depends on the chooser's strategy, as shown below). Using the reasoning in Lemma 2.3, we similarly deduce that the actor plays C_2 or D_2 depending only on δ and the lifetime benefit of achieving reputation $\mathcal{C}_2$ rather than reputation $\mathcal{D}_2$: the actor's strategy does not depend on her current reputation. □

2.2.4 Possible equilibrium chooser strategies for non-empty reputations σ^*_{ch}

Using Lemmas (2.2-2.4), we deduce that reputation must incentivize first-order cooperation for it to occur, and that reputation must incentivize both forms of cooperation for second-order cooperation to occur—and therefore for the institution to meaningully receive contributions. This restricts the possibilities for σ^*_{ch}.

Lemma 2.5: If cooperation occurs, it is incentivized by reputation

$\forall \sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}, \exists I \subset (0,1), \mathbf{P}(\delta \in I) > 0, \forall R \in \mathcal{R}, \forall \delta \in I,$

$$\sigma_{act}(R, \mathcal{T}, \delta) = C_1 \implies \begin{cases} \sigma_{ch}(\mathcal{C}_1) = T \\ \sigma_{ch}(\mathcal{D}_1) = \neg T \end{cases}$$

In any subgame perfect equilibrium in which the actor cooperates over a non-trivial interval of values for her discount factor, reputation incentivizes cooperation—the chooser trusts given $\mathcal{C}_1$, and distrusts given $\mathcal{D}_1$.

Proof. This follows from the above lemmas, and the fact that net cost of cooperation is positive: $c_1 - \gamma - \beta > 0$. Given the opportunity to do so, playing C_1 instead of D_1 leads to an immediate payoff loss, which can only be upset by a future gain, that will be reflected in the difference $U_\delta(\sigma \mid_{\mathcal{C}_1}) - U_\delta(\sigma \mid_{\mathcal{D}_1})$, as visible in equation (2.2).

For cooperation to occur over a non-trivial interval of values for the actor's discount factor, this difference must therefore be positive; which requires that choosers trust given $\mathcal{C}_1$ and distrust given $\mathcal{D}_1$. Note that we will calculate this difference in continuation payoffs in the following section. Here, it suffices to note that $U_\delta(\sigma \mid_{\mathcal{C}_1}) - U_\delta(\sigma \mid_{\mathcal{D}_1}) = 0$ if choosers play the same action given both of these reputations, and that $U_\delta(\sigma \mid_{\mathcal{C}_1}) - U_\delta(\sigma \mid_{\mathcal{D}_1}) \leq 0$ if they trust given $\mathcal{D}_1$ and distrust given $\mathcal{C}_1$. □

Lemma 2.6: If contribution occurs, both forms of cooperation are incentivized by reputation

$\forall \sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}, \exists I \subset (0,1), \mathbf{P}(\delta \in I) > 0, \forall R \in \mathcal{R}, \forall \delta \in I,$

$$\sigma_{\text{act}}(R, \mathcal{I}, \delta) = C_2 \implies \begin{cases} \sigma_{\text{ch}}(\mathcal{C}_1) = T & \text{and} \quad \sigma_{\text{ch}}(\mathcal{C}_2) = T \\ \sigma_{\text{ch}}(\mathcal{D}_1) = \neg T & \text{and} \quad \sigma_{\text{ch}}(\mathcal{D}_2) = \neg T \end{cases}$$

In any subgame perfect equilibrium in which the actor contributes over a non-trivial interval of values for her discount factor, reputation incentivizes both forms of cooperation—the chooser trusts given $\mathcal{C}_1$ and $\mathcal{C}_2$, and distrusts given $\mathcal{D}_1$ and $\mathcal{D}_2$.

Proof. Let $\sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}$, be a subgame perfect equilibrium such that there exists a non-trivial interval $I \subset (0,1)$, with $\mathbf{P}(\delta \in I) > 0$, over which the actor contributes, whatever her reputation.

As in the proof just above, using Lemma 2.3 and 2.4, and the fact that $c_2 > 0$, we deduce that the chooser must trust given $\mathcal{C}_2$ and distrust given $\mathcal{D}_2$. Indeed, given the opportunity to do so, playing C_2 instead of D_2 leads to an immediate payoff loss, which can only be upset by a future gain, that will be reflected in the difference $U_\delta(\sigma \mid_{\mathcal{C}_2}) - U_\delta(\sigma \mid_{\mathcal{D}_2})$, as visible in equation (2.3). We conclude as above (and calculate this difference in the next section).

To show that the chooser must also incentivize first-order cooperation, we begin by recalling that the actor's strategy is stationary following Lemma 2.4, and then calculate the chooser's expected payoff from playing T given information $\mathcal{C}_1$, and given information $\mathcal{D}_1$. Note that we will return to, and rigorously calculate, the chooser's payoff in section 4; here, it suffices to note that this payoff must be positive in the first case, and negative in the second case.

Let $I_{C_2} \supset I$ be the subset of $(0,1)$ over which she contributes—that is, the maximum subset of $(0,1)$ such that $\forall R \in \mathcal{R}, \forall \delta \in I_{C_2}, \sigma_{act}(R, \mathcal{T}, \delta) = C_2$—and let $I_{D_2} = (0,1) \setminus I_{C_2}$ be the subset of $(0,1)$ over which she free-rides. By assumption, $\mathbf{P}(\delta \in I_{C_2}) \geq \mathbf{P}(\delta \in I) > 0$, meaning that the event $\mathcal{C}_2$ also occurs with positive probability since $1 - q > 0$ and since $p_2 > 0$ (this event occurs when the actor's δ is in I_{C_2}, and after every round in which the institution game is drawn, and the actor's action observed). Using equation (2.3), we further note that we must have $\mathbf{P}(\delta \in I_{D_2}) > 0$—given sufficiently small values of δ, this equation cannot be satisfied. It follows that the event $\mathcal{D}_2$ also occurs with positive probability.

Let I_{C_1} be the subset of $(0,1)$ over which she cooperates, and $I_{D_1} = (0,1) \setminus I_{C_1}$ be the subset of $(0,1)$ over which she defects. Since the chooser trusts given $\mathcal{C}_2$, as we just showed, and since $k > 0$, we deduce that $\mathbf{P}(\delta \in I_{C_1}) > 0$: the actor must cooperate with positive probability, otherwise trusting would be net costly. Every time the chooser has information $\mathcal{C}_1$ (which also occurs with positive probability), he can then infer with certainty that the actor's discount factor belongs to I_{C_1}: the actor can only achieve reputation $\mathcal{C}_1$ if she (is trusted and then) cooperates (and then is observed), which requires $\delta \in I_{C_1}$.

In other words, given information $\mathcal{C}_1$, the actor reciprocates with certainty. If the chooser trusts, he gains payoff:

$$-k + 1 \times b = b - k > 0$$

This payoff is greater than 0, which is the payoff the chooser would gain from not trusting. Since σ is subgame perfect, the chooser must trust given $\mathcal{C}_1$.

We conclude by, similarly, showing that $\mathbf{P}(\delta \in I_{D_1}) > 0$: since the chooser distrusts given $\mathcal{D}_2$, the actor must cheat with positive probability—otherwise, trusting given $\mathcal{D}_2$ would yield payoff $-k + b > 0$, and it would be strictly beneficial to deviate to doing so.

Given information $\mathcal{D}_1$, the actor cheats then with certainty. If the chooser trusts, he gains payoff:

$$-k + 0 \times b = -k < 0$$

This payoff is smaller than 0, which is the payoff the chooser would gain from not trusting. Since σ is subgame perfect, the chooser must distrust given $\mathcal{D}_1$.

□

2.2.5 Three types of subgame perfect equilibrium

The last two lemmas restrict the possibilities for σ^*_{ch}. We deduce that there are in fact only three cases that can be subgame perfect (we show that these cases are indeed subgame perfect under certain conditions in the following sections).

First case: reputation does not incentivize either form of cooperation. Following Lemmas 2.5 and 2.6, the actor then cooperates and contributes with probability 0. It follows that the chooser will never trust in a subgame perfect equilibrium—this is the uncooperative equilibrium that we have already seen, and that we will not come back to.

Second case: reputation incentivizes first-order cooperation, but not second-order cooperation. In such a situation, following Lemma 2.6, the actor contributes with probability 0. In section 5 we derive the conditions under which we obtain a subgame perfect equilibrium in which the chooser trusts given $\mathcal{C}_1$ and distrusts given $\mathcal{D}_1$, but either does not trust given $\mathcal{C}_2$ or does not distrust given $\mathcal{D}_2$. We show that the actor must cooperate with positive probability in such a situation, and calculate that probability.

Third and final case: reputation incentivizes first- and second-order cooperation. As we just saw, this means that the chooser trusts given $\mathcal{C}_1$ and $\mathcal{C}_2$, and distrusts given $\mathcal{D}_1$ and $\mathcal{D}_2$—in other words, this fully determines σ^*_{ch}. The only thing left to determine is the probability $\theta \in [0, 1]$ that he trusts given empty reputation.

We call such a situation the **institution equilibrium**—this is the only case in which the institution receives any contributions, and can have any effect on first-order cooperation. In section 3, we show that for every possible θ, the actor's strategy is fully determined. In section 4, we determine the equilibrium value of θ, which in the parameter range we study in section 6 is unique—prompting us to call this strategy profile 'the' institution equilibrium, rather than 'an' institution equilibrium. In section 4, we also derive necessary and sufficient conditions for the existence of the institution equilibrium.

3 Institution equilibrium: actor strategy

In this section, we characterize the actor's strategy in the institution equilibrium, in which we have shown that σ^*_{ch} is determined, but that the chooser may trust given empty reputation with any probability $\theta \in [0, 1]$. We show that, necessarily, the actor cooperates if and only if her discount rate exceeds a certain threshold $\hat{\delta}_1(\theta)$ (in section 3.3), and that she contributes if and only if her discount rate exceeds another threshold $\hat{\delta}_2(\theta)$ (in section 3.4). Knowing θ, we can then calculate the probability of cooperation and contribution—that is, the fraction of cooperators and contributors for the infinite population model we have in mind. In addition, we define and calculate the actor's normalized payoff in the institution equilibrium, as well as the steady state of her reputation, and the long-run level of cooperation.

3.1 Objective and simplifying notations

3.1.1 Objective

Throughout this section, we assume that the chooser trusts given $\mathcal{C}_1$ and $\mathcal{C}_2$, distrusts given $\mathcal{D}_1$ and $\mathcal{D}_2$, and trusts given $\emptyset$ with probability $\theta \in [0, 1]$. We note such a strategy $\sigma^{inst,\theta}_{ch}$. As we just saw, following Lemma 2.6, this is the

form that the chooser's strategy must take for an institution to be established—the actor only contributes to the institution with positive probability if reputation incentivizes both forms of cooperation.

We assume the existence of a subgame perfect strategy $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, that we call the institution equilibrium. Our objective in this section is to characterize the actor's strategy σ_{act}, and show that it is uniquely determined. Later, in section 4, we will show that θ is also uniquely determined, and derive the conditions under which a subgame perfect institution equilibrium does in fact exist.

3.1.2 Reputational benefit

Note that when the chooser plays according to $\sigma_{ch}^{inst,\theta}$, reputation incentivizes first- and second-order cooperation *equally*—for the actor, the benefit of attaining reputation $\mathcal{C}_1$ rather than $\mathcal{D}_1$ is the same as the benefit of attaining reputation $\mathcal{C}_2$ rather than $\mathcal{D}_2$ (in our set up, reputational incentives are the fruit of only one decision, namely the chooser's decision to trust the actor).

Given discount factor $\delta \in (0,1)$, we note $U_\delta^G \equiv U_\delta(\sigma \mid_{\mathcal{C}_1}) = U_\delta(\sigma \mid_{\mathcal{C}_2})$ the actor's lifetime payoff in a continuation game associated with $\mathcal{C}_1$ or $\mathcal{C}_2$—in which case we say that the actor is in *good standing*. Similarly, we note $U_\delta^B \equiv U_\delta(\sigma \mid_{\mathcal{D}_1}) = U_\delta(\sigma \mid_{\mathcal{D}_2})$, and say that the actor is in *bad standing* when her reputation is $\mathcal{D}_1$ or $\mathcal{D}_2$. Finally, we note $U_\delta^\emptyset = U_\delta(\sigma \mid_\emptyset)$, and say that the actor is in *null standing* when her reputation is $\emptyset$.

We define the **reputational benefit** of good behavior (or simply reputational benefit) R_δ to be the difference in continuation payoffs given good vs. bad standing, i.e. we define:

$$R_\delta \equiv U_\delta^G - U_\delta^B$$

$U_\delta^G, U_\delta^B, U_\delta^\emptyset$, and R_δ are functions of the actor's discount factor δ, as well as the strategy profile σ—which we have fixed in this section, and omit for concision.

3.1.3 Normalized actor payoff

Since the actor starts out in null standing, given δ, her payoff is the discounted sum: $U_\delta(\sigma) = U_\delta^\emptyset$. Her normalized payoff $\overline{U}_\delta$ is obtained by multiplying the discounted sum by $(1-\delta)$:

$$\overline{U}_\delta \equiv (1-\delta) U_\delta^\emptyset \tag{3.1}$$

In our numerical results, in section 6, we rely on the expected value of $\overline{U}_\delta$ given a certain distribution of discount factors—that is, the average normalized payoff of an actor in the infinite population model.

3.1.4 Condensed payoffs

We note $\mathbf{r_C} \equiv r - c_1 + \beta$ the stage payoff of the actor if she is trusted and reciprocates, taking into account potential institutional reward $\beta \geq 0$. Similarly, we note $\mathbf{r_D} \equiv r - \gamma$ her stage payoff if is trusted and cheats, $\mathbf{c_1} \equiv \mathbf{r_D} - \mathbf{r_C} = r - (\beta + \gamma)$ the cost of cooperation, and $\mathbf{p_1} = p_1 + \pi_1$ the probability of cooperation—all of which take into account the effect of the institution in the proposed equilibrium σ.

Using our new notations, for any R and δ, we can rewrite the conditions for first- and second-order cooperation in (Lemmas 2.2 and 2.3) as:

$$\sigma_{act}(R, \mathcal{T}, \delta) = C_1 \iff \mathbf{c_1} \leq \delta \times \mathbf{p_1} \times R_\delta \tag{2.2'}$$

$$\sigma_{act}(R, \mathcal{I}, \delta) = C_2 \iff c_2 \leq \delta \times p_2 \times R_\delta \tag{2.3'}$$

We use these condensed payoffs to reduce the size our calculations in the proofs below.

3.1.5 Partition of the set of possible discount factors $(0,1)$

As we have seen, the actor's strategy is stationary (2.4): she reciprocates given $\mathcal{T}$ and contributes given $\mathcal{I}$ depending only on the value of her discount factor δ–or, more simply, she reciprocates and contributes depending only on her δ. We note $I_{C_1} \subset (0,1)$ the subset of discount factors for which σ_{act} prescribes reciprocation, that is the maximum interval such that, $\forall R \in \mathcal{R}, \forall \delta \in I_{C_1}, \sigma_{act}(R, \mathcal{T}, \delta) = C_1$. We note $I_{D_1} \equiv (0,1) \setminus I_{C_1}$ the subset of discount factor for which σ_{act} prescribes cheating; and analogously define I_{C_2} and I_{D_2}.

The intersections of these intervals partition the set of possible discount factors $(0,1)$ into four, depending on the two actions prescribed by σ_{act} in both games—that is, considering all four possibilities for the actor's strategy given her discount factor δ. For instance, if $\delta \in I_{C_1} \cap I_{C_2}$, the actor will play C_1 and C_2 each time she is given the opportunity to, throughout the repeated game.

$I_{C_1}, I_{C_2}, I_{D_1}$, and I_{D_2} are functions of the specific strategy profile σ that we are considering, which we once again omit from the notation for concision. Below, we show that these sets take a simple expression in a subgame perfect equilibrium, and that everything can be reduced to two threshold values $\hat{\delta}_1(\theta)$ and $\hat{\delta}_2(\theta)$.

3.2 Derivation of the reputational benefit R_δ

Using a bit of algebra (Lemma 3.1) and the notations introduced above, we derive the value of the actor's reputational benefit R_δ as a function of δ (Lemma 3.2).

Lemma 3.1: Continuation payoff given empty reputation

Assume that the chooser plays $\sigma_{ch}^{inst,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$. For any $\delta \in (0,1)$, we have:

$$U_\delta^\emptyset = \theta U_\delta^G + (1-\theta) U_\delta^B \tag{3.2}$$

$$U_\delta^\emptyset - U_\delta^B = \theta R_\delta \tag{3.3}$$

$$U_\delta^G - U_\delta^\emptyset = (1-\theta) R_\delta \tag{3.4}$$

Proof. Take our strategy profile $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma_{ch}^{inst,\theta}$. Since the actor's strategy is stationary (Lemma 2.4), the only thing that affects the actor's continuation payoff is her current standing, and her discount factor—the latter defining her strategy.

When the actor is in good standing, when the trust game is drawn she is trusted with probability 1, and then reciprocates depending only on the value of her discount factor δ. When the institution game is drawn, her standing does not affect her ability to act, and her action depends once again only on her δ.

When she is in bad standing, the only thing that changes is that she is distrusted with probability 1 when the trust game is drawn—and when she is in null standing, the only thing that changes is that she is trusted with probability probability θ in such a case, and distrusted with probability $1-\theta$.

In other words, when she is in null standing, with probability θ, everything is as if she is in good standing, and with probability $1-\theta$, everything is as if she is in bad standing. We deduce:

$$U_\delta^\emptyset = \theta U_\delta^G + (1-\theta) U_\delta^B \tag{3.2}$$

The other two equations are deduced from this equation and the definition of R_δ, using $U_\delta^G = R_\delta + U_\delta^B$ to obtain (3.3), and $U_\delta^B = U_\delta^G - R_\delta$ to obtain (3.4). □

Lemma 3.2: Reputational benefit

Assume that the chooser plays $\sigma_{ch}^{inst,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$. For any $\delta \in (0,1)$, we have:

$$R_\delta = \begin{cases} \frac{q(r-\gamma)}{1+q\delta(p_1+\pi_1)\theta} & \text{if } \delta \in I_{D_1} \\ \frac{q(r-c_1+\beta)}{1-q\delta(p_1+\pi_1)(1-\theta)} & \text{if } \delta \in I_{C_1} \end{cases} \tag{3.5}$$

Proof. Take our strategy profile $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma_{ch}^{inst,\theta}$. As we have seen, the intersections $I_{C_1} \cap I_{C_2}$, $I_{C_1} \cap I_{D_2}$, $I_{D_1} \cap I_{C_1}$, and $I_{D_1} \cap I_{D_2}$ partition the set of possible discount factors into four. In addition, the actor's strategy is the same over any of these sets—for instance, for any $\delta \in I_{C_1} \cap I_{C_2}$, the actor cooperates and contributes throughout the game, whatever her reputation, and hence whatever the history.

Using this partition, we can calculate R_δ four times, each time taking the actor's strategy as fixed—as we will show, only behavior in the trust game affects the value of R_δ, meaning that we only need to partition between I_{C_1} and I_{D_1}.

To begin, let us consider the case $\delta \in I_{C_1} \cap I_{C_2}$, in which the actor always cooperates at both orders. Given bad standing, she isn't trusted in the trust game, earning payoff 0; following which her standing becomes null, with certainty (since she does not act). In the institution game, she contributes, earning payoff $-c_2$; following which her

standing becomes good if she is observed, with probability p_2, and null otherwise, with probability $1 - p_2$. Since the trust game is drawn with probability q and the institution game with probability $1 - q$, we deduce:

$$U_\delta^B = q \times (0 + \delta \times [1 \times U_\delta^\emptyset]) + (1 - q) \times (-c_2 + \delta \times [p_2 \times U_\delta^G + (1 - p_2) \times U_\delta^\emptyset]) \qquad \forall \delta \in I_{C_1} \cap I_{C_2}$$

Given good standing, the actor is trusted in the trust game, and reciprocates that trust, earning payoff $\mathbf{r_C}$ instead of 0; following which she achieves good standing with probability $\mathbf{p_1}$, and null standing with probability $1 - \mathbf{p_1}$. In other words:

$$U_\delta^G = q \times (\mathbf{r_C} + \delta \times [\mathbf{p_1} \times U_\delta^G + (1 - \mathbf{p_1}) \times U_\delta^\emptyset]) + (1 - q) \times (-c_2 + \delta \times [p_2 \times U_\delta^G + (1 - p_2) \times U_\delta^\emptyset]) \quad \forall \delta \in I_{C_1} \cap I_{C_2}$$

Subtracting U_δ^B to U_δ^G, the right parts of each expressions simplify—since the actor's standing does not affect her ability to act in the institution game, her action in that game (here, C_2) does not affect her reputational benefit. We obtain:

$$\begin{aligned} R_\delta &= q \times (\mathbf{r_C} + \delta \times [\mathbf{p_1} \times U_\delta^G + (1 - \mathbf{p_1}) \times U_\delta^\emptyset]) - q \times (0 + \delta \times [1 \times U_\delta^\emptyset]) && \forall \delta \in I_{C_1} \cap I_{C_2} \\ R_\delta &= q \times (\mathbf{r_C} + \delta \times \mathbf{p_1}(U_\delta^G - U_\delta^\emptyset)) && \forall \delta \in I_{C_1} \cap I_{C_2} \end{aligned}$$

We obtain the same expression for $\delta \in I_{C_1} \cap I_{D_2}$—since, once again, R_δ is unaffected by behavior in the institution game (in this case, the right parts of U_δ^B and U_δ^G are both equal to $(1 - q) \times (0 + \delta \times [p_2 \times U_\delta^B + (1 - p_2) \times U_\delta^\emptyset])$). In other words, the above expression is valid for every $\delta \in I_{C_1}$.

Replacing $U_\delta^G - U_\delta^\emptyset$ using equation (3.4), we deduce:

$$\begin{aligned} R_\delta &= q \times (\mathbf{r_C} + \delta \times \mathbf{p_1} \times (1 - \theta) R_\delta(\sigma)) && \forall \delta \in I_{C_1} \\ R_\delta \times [1 - q \times \delta \times \mathbf{p_1} \times (1 - \theta)] &= q \times \mathbf{r_C} && \forall \delta \in I_{C_1} \\ R_\delta &= \frac{q \times \mathbf{r_C}}{1 - q \times \delta \times \mathbf{p_1} \times (1 - \theta)} && \forall \delta \in I_{C_1} \\ R_\delta &= \frac{q \times (r - c_1 + \beta)}{1 - q \times \delta \times (p_1 + \pi_1) \times (1 - \theta)} && \forall \delta \in I_{C_1} \end{aligned}$$

This proves the lower part of the (3.5).

To prove the remaining upper part of the equation, and prove the lemma, we calculate R_δ directly when $\delta \in I_{D_1}$. Given bad standing, the actor isn't trusted in the trust game, earning payoff 0; following which her standing becomes null, with certainty. Given good standing, the actor is trusted in the trust game, earning payoff $\mathbf{r_D}$; following which her standing becomes bad with probability $\mathbf{p}_1$, and null with probability $1 - \mathbf{p}_1$. Her behavior in the institution game does not matter to R_δ, since it simplifies in both cases $\delta \in I_{D_1} \cap I_{C_2}$ and $\delta \in I_{D_1} \cap I_{D_2}$.

Subtracting her prospects in the trust game given bad standing to those same prospects given good standing, we deduce:

$$\begin{aligned} R_\delta &= q \times (\mathbf{r_D} + \delta \times [\mathbf{p_1} \times U_\delta^B + (1 - \mathbf{p_1}) \times U_\delta^\emptyset]) - q \times (0 + \delta \times [1 \times U_\delta^\emptyset]) && \forall \delta \in I_{D_1} \\ R_\delta &= q \times (\mathbf{r_D} + \delta \times \mathbf{p_1}(U_\delta^B - U_\delta^\emptyset)) && \forall \delta \in I_{D_1} \end{aligned}$$

And, replacing $U_{\delta_-}^B - U_{\delta_-}^\emptyset$ using equation (3.3), we deduce:

$$\begin{aligned} R_\delta &= q \times (\mathbf{r_D} + \delta \times \mathbf{p_1} \times (-\theta R_\delta)) && \forall \delta \in I_{D_1} \\ R_\delta \times [1 + q \times \delta \times \mathbf{p_1} \times \theta] &= q \times \mathbf{r_D} && \forall \delta \in I_{D_1} \\ R_\delta &= \frac{q \times \mathbf{r_D}}{1 + q \times \delta \times \mathbf{p_1} \times \theta} && \forall \delta \in I_{D_1} \\ R_\delta &= \frac{q \times (r - \gamma)}{1 + q \times \delta \times (p_1 + \pi_1) \times \theta} && \forall \delta \in I_{D_1} \end{aligned}$$

□

3.3 Threshold discount factor for first-order cooperation

Using the general condition for cooperation (2.2) (shown in Lemma 2.2), we deduce that, in the institution equilibrium, the actor cooperates if she is sufficiently patient, and defects if she is sufficiently impatient—the indifference point being captured by a single threshold value $\hat{\delta}_1(\theta)$.

Proposition 3.1: Threshold discount factor for cooperation

Assume that the chooser plays $\sigma_{ch}^{inst,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$. Then:

$$I_{D_1} =]0, \hat{\delta}_1(\theta)[\cap (0,1)$$
$$I_{C_1} = [\hat{\delta}_1(\theta), 1[\cap (0,1)$$

where:

$$\hat{\delta}_1(\theta) \equiv \frac{c_1 - (\beta + \gamma)}{(p_1 + \pi_1)q[r - \gamma - \theta(c_1 - (\beta + \gamma)]} \tag{3.6}$$

In the institution equilibrium, the actor always reciprocates if she is sufficiently patient ($\delta \geq \hat{\delta}_1(\theta)$), and always cheats if she is sufficiently impatient ($\delta < \hat{\delta}_1(\theta)$)—the value of the threshold discount factor being a function of the likelihood that the chooser trusts given empty reputation θ.

Proof. Take our strategy profile $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma_{ch}^{inst,\theta}$.

Using the general condition for cooperation (2.2), shown in Lemma 2.2, our condensed payoffs, and the definition of I_{C_1}, we have:

$$\delta \in I_{C_1} \iff \mathbf{c_1} \leq \delta \times \mathbf{p_1} \times R_\delta$$

In the previous lemma, we showed that:

$$R_\delta = \frac{q \times \mathbf{r_C}}{1 - q \times \delta \times \mathbf{p_1} \times (1 - \theta)} \qquad \forall \delta \in I_{C_1}$$

Replacing, we deduce:

$$\begin{aligned}
\delta \in I_{C_1} &\iff \mathbf{c_1} \leq \delta \times \mathbf{p_1} \times \frac{q \times \mathbf{r_C}}{1 - q \times \delta \times \mathbf{p_1} \times (1 - \theta)} \\
&\iff \mathbf{c_1} \times [1 - q \times \delta \times \mathbf{p_1} \times (1 - \theta)] \leq \delta \times \mathbf{p_1} \times q \times \mathbf{r_C} \\
&\iff \mathbf{c_1} \leq \delta \times [\mathbf{p_1} \times q \times \mathbf{r_C} + q \times \mathbf{p_1} \times (1 - \theta) \times \mathbf{c_1}] \\
&\iff \delta \geq \frac{\mathbf{c_1}}{\mathbf{p_1} \times q \times (\mathbf{r_C} + (1 - \theta) \times \mathbf{c_1})}
\end{aligned}$$

Using $\mathbf{r_C} + 1 \times \mathbf{c_1} = \mathbf{r_D}$, and replacing our condensed payoffs, we deduce:

$$\begin{aligned}
\delta \in I_{C_1} &\iff \delta \geq \frac{\mathbf{c_1}}{\mathbf{p_1} \times q \times (\mathbf{r_D} - \theta \times \mathbf{c_1})} \\
&\iff \delta \geq \frac{c_1 - (\beta + \gamma)}{(p_1 + \pi_1) \times q \times (r - \gamma - \theta \times (c_1 - (\beta + \gamma)))}
\end{aligned}$$

This proves that I_{C_1} is composed of all the possible discount factor values that are greater or equal to the threshold $\hat{\delta}_1(\theta)$, as given by condition (3.6), and therefore $I_{C_1} = [\hat{\delta}_1, 1[\cap(0,1)$. Since $I_{D_1} = (0,1) \setminus I_{C_1}$, we deduce $I_{D_1} =]0, \hat{\delta}_1[\cap(0,1)$, proving the proposition.

For good measure, let us note that we obtain the same result when focusing on I_{D_1}. The same general condition for cooperation shows that:

$$\delta \in I_{D_1} \iff \mathbf{c_1} > \delta \times \mathbf{p_1} \times R_\delta$$

Using the expression $R_\delta = \frac{q \times \mathbf{r_D}}{1 + q \times \delta \times \mathbf{p_1} \times \theta}$, valid over the entire set I_{D_1}, we deduce:

$$\begin{aligned}
\delta \in I_{D_1} &\iff \mathbf{c_1} > \delta \times \mathbf{p_1} \times \frac{q \times \mathbf{r_D}}{1 + q \times \delta \times \mathbf{p_1} \times \theta} \\
&\iff \mathbf{c_1} \times [1 + q \times \delta \times \mathbf{p_1} \times \theta] > \delta \times \mathbf{p_1} \times q \times \mathbf{r_D} \\
&\iff \mathbf{c_1} > \delta \times \mathbf{p_1} \times q \times [\mathbf{r_D} - \theta \times \mathbf{c_1}] \\
&\iff \delta < \frac{\mathbf{c_1}}{\mathbf{p_1} \times q \times (\mathbf{r_D} - \theta \times \mathbf{c_1})} \\
&\iff \delta < \hat{\delta}_1(\theta)
\end{aligned}$$

□

We refer to $\hat{\delta}_1(\theta)$ as the **difficulty of cooperation in the institution equilibrium for** θ. In the infinite population model we have in mind, sufficiently impatient actors always cheat, and since $\hat{\delta}_1(\theta) > 0$, the fraction of cheaters is always positive. Sufficiently patient actors reciprocate, although we need $\hat{\delta}_1(\theta) < 1$ for this to occur with positive probability—cooperation cannot be 'too difficult'.

3.4 Threshold discount factor for second-order cooperation

Using the general condition for contribution (2.3) (shown in Lemma 2.3), we show, similarly, that the actor contributes if she is sufficiently patient, and free-rides if she is sufficiently impatient.

Similarly to above, we refer to the indifference point $\hat{\delta}_2(\theta)$ as the **difficulty of second-order cooperation in the institution equilibrium for** θ. There are two cases, depending on whether this threshold is smaller, or greater, than $\hat{\delta}_1(\theta)$—that is, on whether second-order cooperation can be said to be 'easier', or 'more difficult', than first-order cooperation.

Proposition 3.2: Threshold discount factor for second-order cooperation

Assume that the chooser plays $\sigma_{ch}^{inst,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$. Then:

$$I_{D_2} =]0, \hat{\delta}_2(\theta)[\cap (0,1)$$
$$I_{C_2} = [\hat{\delta}_2(\theta), 1[\cap (0,1)$$

where:

$$\hat{\delta}_2(\theta) \equiv \begin{cases} \frac{c_2}{q[p_2(r-\gamma)-(p_1+\pi_1)\theta c_2]} & \text{if } \hat{\delta}_2(\theta) < \hat{\delta}_1(\theta) \\ \frac{c_2}{q[p_2(r-c_1+\beta)+(p_1+\pi_1)(1-\theta)c_2]} & \text{if } \hat{\delta}_2(\theta) \geq \hat{\delta}_1(\theta) \end{cases} \tag{3.7}$$

In the institution equilibrium, the actor always contributes if she is sufficiently patient ($\delta \geq \hat{\delta}_2(\theta)$), and always free-rides if she is sufficiently impatient ($\delta < \hat{\delta}_2(\theta)$)—the value of the threshold discount factor being a function of the likelihood that the chooser trusts given empty reputation θ, as well as whether second-order cooperation is more difficult than first-order cooperation ($\hat{\delta}_2(\theta) \geq \hat{\delta}_1(\theta)$), or less difficult ($\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta)$) ($\hat{\delta}_1(\theta)$ being defined in Proposition 3.1).
What's more:

$$\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta) \iff \frac{c_2}{p_2} < \frac{\mathbf{c_1}}{\mathbf{p_1}} = \frac{c_1 - (\beta+\gamma)}{p_1 + \pi_1} \tag{3.8}$$

The easiest form of cooperation is the one with the lowest ratio of cost divided by observability.

Proof. Take our strategy profile $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma_{ch}^{inst,\theta}$.

Using equation (2.3) and a similar reasoning to the proof above, we deduce that the actor will contribute given $\mathcal{I}$ if and only if her discount factor exceeds the threshold $\hat{\delta}_2(\theta)$ satisfying the equation:

$$c_2 = \hat{\delta}_2(\theta) p_2 R_{\hat{\delta}_2(\theta)}(\sigma)$$

There are two possibilities, depending on whether this threshold is smaller than $\hat{\delta}_1(\theta)$—in which case we must use the upper part of (3.5) to replace $R_{\hat{\delta}_2(\theta)}(\sigma)$—or larger than this threshold—in which case we must use the lower part of this equation.

First case (second-order cooperation is easier): when $\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta)$, the critical reputational benefit is obtained using the upper part of (3.5). We obtain:

$$c_2 = \hat{\delta}_2(\theta) p_2 \frac{q(r-\gamma)}{1 + \hat{\delta}_2(\theta)(p_1+\pi_1)q\theta} \qquad \text{if } \hat{\delta}_2(\theta) < \hat{\delta}_1(\theta)$$
$$\hat{\delta}_2(\theta) = \frac{c_2}{q[p_2(r-\gamma) - (p_1+\pi_1)\theta c_2]} \qquad \text{if } \hat{\delta}_2(\theta) < \hat{\delta}_1(\theta)$$

Second case (second-order cooperation is more difficult): when $\hat{\delta}_2(\theta) \geq \hat{\delta}_1(\theta)$, we obtain:

$$c_2 = \hat{\delta}_2(\theta) p_2 \frac{q(r - c_1 + \beta)}{1 - q\hat{\delta}_2(\theta)(p_1 + \pi_1)(1 - \theta)} \qquad \text{if } \hat{\delta}_2(\theta) \geq \hat{\delta}_1(\theta)$$

$$\hat{\delta}_2(\theta) = \frac{c_2}{q[p_2(r - c_1 + \beta) + (p_1 + \pi_1)(1 - \theta)c_2]} \qquad \text{if } \hat{\delta}_2(\theta) \geq \hat{\delta}_1(\theta)$$

Bringing both equations together, we have proven condition (3.7).

Using condition (3.6), proven in the previous Proposition, the upper part of (3.7), as well as the condensed notations introduced above, we deduce that if $\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta)$, then we must have:

$$\begin{aligned}
\frac{c_2}{q[p_2 \mathbf{r_D} - \mathbf{p_1}\theta c_2]} &< \frac{\mathbf{c_1}}{\mathbf{p_1} q[\mathbf{r_D} - \theta \mathbf{c_1}]} \\
c_2 \mathbf{p_1}(\mathbf{r_D} - \theta \mathbf{c_1}) &< \mathbf{c_1}(p_2 \mathbf{r_D} - \mathbf{p_1}\theta c_2) \\
c_2 \mathbf{p_1} \mathbf{r_D} &< \mathbf{c_1} p_2 \mathbf{r_D} \\
\frac{c_2}{p_2} &< \frac{\mathbf{c_1}}{\mathbf{p_1}}
\end{aligned}$$

Using the bottom part of (3.7), we deduce that the above is also a sufficient condition. Second-order cooperation is easier than first-order cooperation if and only if its cost divided by the relevant probability of observation p_2 is smaller than the net cost of first-order cooperation divided by the relevant total probability of observation $\mathbf{p_1}$. In other words:

$$\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta) \iff \frac{c_2}{p_2} < \frac{\mathbf{c_1}}{\mathbf{p_1}} = \frac{c_1 - \gamma - \beta}{p_1 + \pi_1} \tag{3.8}$$

□

3.5 Normalized actor payoff

Using our previous results, we calculate the normalized payoff of the actor given δ in the institution equilibrium.

Lemma 3.3: Normalized actor payoff

Assume that the chooser plays $\sigma_{ch}^{inst,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$. For any $\delta \in (0,1)$, we have:

$$\overline{U}_\delta = \begin{cases} [1 - (1-q)p_2\delta]\theta R_\delta & \text{if } \delta < \hat{\delta}_2(\theta) \\ (1-q)(-c_2) + [\theta + (1-q)p_2\delta(1-\theta)]R_\delta & \text{if } \delta \geq \hat{\delta}_2(\theta) \end{cases} \tag{3.9}$$

Proof. Take our strategy profile $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma_{ch}^{inst,\theta}$. We consider an actor of discount factor $\delta \in (0,1)$.

We begin by calculating the actor's continuation payoff given bad standing—in which case, she is never trusted this round, and her action in the trust game has no bearing for her immediate payoffs. Following Proposition (3.2), there are two cases: either $\delta < \hat{\delta}_2(\theta)$, and the actor will free-ride if the institution game is drawn, or $\delta \geq \delta_2(\theta)$, and the actor will contribute if the institution game is drawn.

First case: $\delta < \hat{\delta}_2(\theta)$. If the trust game is drawn, the actor does not play and achieves null standing with certainty. If the institution game is drawn, the actor free-rides, and achieves bad standing if she is observed, and null standing otherwise. Thus:

$$\begin{aligned}
U_\delta^B &= q \times (0 + \delta \times [1 \times U_\delta^\emptyset]) + (1-q) \times (0 + \delta \times [p_2 \times U_\delta^B + (1-p_2) \times U_\delta^\emptyset) && \forall \delta < \hat{\delta}_2(\theta) \\
U_\delta^B &= [q + (1-q)(1-p_2)] \times \delta U_\delta^\emptyset + (1-q)p_2 \times \delta U_\delta^B && \forall \delta < \hat{\delta}_2(\theta) \\
U_\delta^B[1 - (1-q)p_2 \times \delta] &= [1 - (1-q)p_2] \times \delta U_\delta^\emptyset && \forall \delta < \hat{\delta}_2(\theta)
\end{aligned}$$

Since $U_\delta^B = U_\delta^\emptyset - \theta R_\delta$, following equation (3.3), we deduce:

$$\begin{aligned}
(U_\delta^\emptyset - \theta R_\delta)[1 - (1-q)p_2 \times \delta] &= [1 - (1-q)p_2] \times \delta U_\delta^\emptyset && \forall \delta < \hat{\delta}_2(\theta) \\
U_\delta^\emptyset([1 - (1-q)p_2 \times \delta] - [1 - (1-q)p_2] \times \delta) &= [1 - (1-q)p_2\delta]\theta R_\delta && \forall \delta < \hat{\delta}_2(\theta) \\
U_\delta^\emptyset(1 - \delta) &= [1 - (1-q)p_2\delta]\theta R_\delta && \forall \delta < \hat{\delta}_2(\theta) \\
\overline{U}_\delta &= [1 - (1-q)p_2\delta]\theta R_\delta && \forall \delta < \hat{\delta}_2(\theta)
\end{aligned}$$

This proves the upper part of (3.9). To prove the lower part of the equation, and thus the lemma, we consider the second case: $\delta \geq \hat{\delta}_2(\theta)$. In this case, in contrast to before, the actor contributes when the trust game, thus achieving good standing if observed. Her payoff given bad standing is then:

$$U_\delta^B = q \times (0 + \delta \times [1 \times U_\delta^\emptyset]) + (1-q) \times (-c_2 + \delta \times [p_2 \times U_\delta^G + (1-p_2) \times U_\delta^\emptyset] \qquad \forall \delta \geq \hat{\delta}_2(\theta)$$
$$U_\delta^B = [1-(1-q)p_2]\delta U_\delta^\emptyset + (1-q)(-c_2 + p_2 \times \delta U_\delta^G) \qquad \forall \delta \geq \hat{\delta}_2(\theta)$$

By subtracting $(1-q)(p_2 \times \delta U_\delta^B)$ from both sides of the equation, and using $U_\delta^G - U_\delta^B = R_\delta$, we deduce:

$$U_\delta^B[1-(1-q)p_2\delta] = [1-(1-q)p_2]\delta U_\delta^\emptyset + (1-q)(-c_2) + (1-q)p_2 \times \delta R_\delta \qquad \forall \delta \geq \hat{\delta}_2(\theta)$$

And, using $U_\delta^B = U_\delta^\emptyset - \theta R_\delta$ once again, which follows from (3.3), we deduce:

$$(U_\delta^\emptyset - \theta R_\delta)[1-(1-q)p_2\delta] = [1-(1-q)p_2]\delta U_\delta^\emptyset + (1-q)(-c_2) + [(1-q)p_2\delta]R_\delta \qquad \forall \delta \geq \hat{\delta}_2(\theta)$$
$$U_\delta^\emptyset(1-\delta) = (1-q)(-c_2) + ([(1-q)p_2\delta] + \theta[1-(1-q)p_2\delta]]R_\delta \qquad \forall \delta \geq \hat{\delta}_2(\theta)$$
$$\overline{U}_\delta = (1-q)(-c_2) + [\theta + (1-q)p_2\delta(1-\theta)]R_\delta \qquad \forall \delta \geq \hat{\delta}_2(\theta)$$

□

3.6 Steady state of the actor's reputation

In the institution equilibrium, given δ, the actor's reputation follows a Markov process—depending on the actor's actions, i.e. depending on whether δ is compared to both thresholds $\hat{\delta}_1(\theta)$ and $\hat{\delta}_2(\theta)$. For instance, if $\delta \geq \max\{\hat{\delta}_1(\theta), \hat{\delta}_2(\theta)\}$, then the actor cooperates in both games, and in at the end of a given round, her reputation becomes $\emptyset$, $\mathcal{C}_1$ or $\mathcal{C}_2$, depending on the game that is drawn that round, whether or not she is observed, and her reputation at the beginning of the round (which affects her ability to cooperate)—but not on her reputation in any round before.

Here, we characterize the steady state of the actor's reputation, which we also call her long-run reputation, assuming that $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$—in which case, following Proposition 3.2, first-order cooperation is more difficult than second-order cooperation ($\hat{\delta}_1(\theta) \geq \hat{\delta}_2(\theta)$), and there are only three cases to consider: $\delta \geq \hat{\delta}_1(\theta)$, in which case the actor reciprocates and contributes; $\delta < \hat{\delta}_2(\theta)$, in which case the actor cheats and free-rides; and $\hat{\delta}_2(\theta) \leq \delta < \hat{\delta}_1(\theta)$, in which case the actor contributes but does not reciprocate.

We use the steady state of the actor's reputation to calculate and plot the long-run level of cooperation, defined just below. In the main document and our numerical resolution (section 6), we assume $c_2 < \mathbf{c_1}$ and $p_2 > \mathbf{p_1}$, guaranteeing that first-order cooperation be more difficult than second-order cooperation, which is the case of interest for us. Were we to allow the converse to be true, we would need to prove a fourth lemma to replace 3.5, as the actor would then reciprocate but not contribute given an intermediary δ.

Lemma 3.4: Long-run reputation for a high patience actor

Assume that the chooser plays $\sigma_{ch}^{inst,\theta}$, with $\theta \in [0,1]$, that $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, and that $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$. For any $\delta \geq \hat{\delta}_1(\theta)$, the actor's reputation follows a Markov process with three states: $\mathcal{C}_1$, $\mathcal{C}_2$ and $\emptyset$. The steady state $\pi^H \equiv (\pi_{\mathcal{C}_1}^H, \pi_{\mathcal{C}_2}^H, \pi_\emptyset^H)$ of the actor's reputation is:

$$\pi_{\mathcal{C}_1}^H \equiv q\mathbf{p}_1 \frac{(1-q)p_2(1-\theta)+\theta}{1-q(1-\theta)\mathbf{p}_1} \tag{3.10}$$
$$\pi_{\mathcal{C}_2}^H \equiv (1-q)p_2 \tag{3.11}$$
$$\pi_\emptyset^H \equiv 1 - q\mathbf{p}_1 \frac{(1-q)p_2(1-\theta)+\theta}{1-q(1-\theta)\mathbf{p}_1} - (1-q)p_2 \tag{3.12}$$

Proof. Take our strategy profile $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma_{ch}^{inst,\theta}$. We assume $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$, and consider a discount factor $\delta \geq \hat{\delta}_1(\theta)$, assuming such a value exists.

The actor plays C_1 and C_2 (or doesn't play) throughout the game. Her reputation can thus take three values during her life: $\mathcal{C}_1$, $\mathcal{C}_2$, and $\emptyset$. For every $t \geq 0$, we note $\pi_{\mathcal{C}_1}^H(t)$, $\pi_{\mathcal{C}_2}^H(t)$ and $\pi_\emptyset^H(t)$ the probability that this high patience actor is in state $\mathcal{C}_1$, $\mathcal{C}_2$, and $\emptyset$, respectively. By assumption, $\pi_{\mathcal{C}_1}^H(0) = \varepsilon = \pi_{\mathcal{C}_2}^H(0)$, and $\pi_\emptyset^H(0) = 1 - 2\varepsilon$.

Entering into any round $t \geq 1$, the actor has reputation $\mathcal{C}_1$ if and only if: in the previous round $t-1$, (i) the trust game was drawn, with probability q, (ii) she was trusted, with probability 1 if her reputation was $\mathcal{C}_1$ or $\mathcal{C}_2$, and probability θ if her reputation was $\emptyset$, and (iii) she was observed, with probability $\mathbf{p_1}$.

It follows that, for any $t \geq 1$:

$$\pi^H_{\mathcal{C}_1}(t) = q \times [1 \times \pi^H_{\mathcal{C}_1}(t-1) + 1 \times \pi^H_{\mathcal{C}_2}(t-1) + \theta \times \pi^H_{\emptyset}(t-1)] \times \mathbf{p_1}$$

Entering into any round $t \geq 1$, the actor has reputation $\mathcal{C}_2$ if and only if: (i) the institution game was drawn, with probability $1-q$, and (ii) she was observed, with probability p_2. It follows that, for all $t \geq 1$:

$$\pi^H_{\mathcal{C}_2}(t) = (1-q) \times p_2$$

These equations show that the actor's reputation follows a Markov process, as, for any $t \geq 1$, $\pi^H_{\mathcal{C}_1}(t)$, $\pi^H_{\mathcal{C}_2}(t)$ and $\pi^H_{\emptyset}(t) = 1 - \pi^H_{\mathcal{C}_1}(t) - \pi^H_{\mathcal{C}_2}(t)$ each only depend on the probabilities for round $t-1$.

Using both of the above equations, the steady state probabilities $\pi^H_{\mathcal{C}_1}, \pi^H_{\mathcal{C}_2}$, and $\pi^H_{\emptyset}$ must verify:

$$\begin{aligned}
\pi^H_{\mathcal{C}_1} &= q(\pi^H_{\mathcal{C}_1} + \pi^H_{\mathcal{C}_2} + \theta\pi^H_{\emptyset})\mathbf{p_1} \\
\pi^H_{\mathcal{C}_2} &= (1-q)p_2 \\
\pi^H_{\emptyset} &= 1 - \pi^H_{\mathcal{C}_1} - \pi^H_{\mathcal{C}_2}
\end{aligned}$$

Replacing in the first equation, we deduce:

$$\begin{aligned}
\pi^H_{\mathcal{C}_1} &= q(\pi^H_{\mathcal{C}_1} + \pi^H_{\mathcal{C}_2} + \theta(1 - \pi^H_{\mathcal{C}_1} - \pi^H_{\mathcal{C}_2}))\mathbf{p_1} \\
\pi^H_{\mathcal{C}_1}(1 - q(1-\theta)\mathbf{p_1}) &= q(\pi^H_{\mathcal{C}_2}(1-\theta) + \theta)\mathbf{p_1} \\
\pi^H_{\mathcal{C}_1} &= q\mathbf{p_1}\frac{\pi^H_{\mathcal{C}_2}(1-\theta) + \theta}{1 - q(1-\theta)\mathbf{p_1}}
\end{aligned}$$

Using $\pi^H_{\mathcal{C}_2} = (1-q)p_2$, we deduce all three proposed equations. □

Lemma 3.5: Long-run reputation for a medium patience actor

Assume that the chooser plays $\sigma_{ch}^{inst,\theta}$, with $\theta \in [0,1]$, that $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, and that $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$. For any $\delta \in [\hat{\delta}_2(\theta), \hat{\delta}_1(\theta)[$, the actor's reputation follows a Markov process with three states: $\mathcal{D}_1$, $\mathcal{C}_2$ and $\emptyset$. The steady state $\pi^M \equiv (\pi^M_{\mathcal{D}_1}, \pi^M_{\mathcal{C}_2}, \pi^M_{\emptyset})$ of the actor's reputation is:

$$\pi^M_{\mathcal{D}_1} \equiv q\mathbf{p_1}\frac{(1-q)p_2(1-\theta) + \theta}{1 + q\theta\mathbf{p_1}} \tag{3.13}$$

$$\pi^M_{\mathcal{C}_2} \equiv (1-q)p_2 \tag{3.14}$$

$$\pi^M_{\emptyset} \equiv 1 - q\mathbf{p_1}\frac{(1-q)p_2(1-\theta) + \theta}{1 + q\theta\mathbf{p_1}} - (1-q)p_2 \tag{3.15}$$

Proof. Take our strategy profile $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma_{ch}^{inst,\theta}$. We assume $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$, and consider a discount factor $\delta \in [\hat{\delta}_2(\theta), \hat{\delta}_1(\theta)[$, assuming such a value exists.

The actor plays D_1 and C_2 (or doesn't play) throughout the game. Her reputation can thus take three values during her life: $\mathcal{D}_1$, $\mathcal{C}_2$, and $\emptyset$. For every $t \geq 0$, we note $\pi^M_{\mathcal{D}_1}(t)$, $\pi^M_{\mathcal{C}_2}(t)$ and $\pi^M_{\emptyset}(t)$ the probability that this medium patience actor is in state $\mathcal{D}_1$, $\mathcal{C}_2$, and $\emptyset$, respectively. By assumption, $\pi^M_{\mathcal{D}_1}(0) = \varepsilon = \pi^M_{\mathcal{C}_2}(0)$, and $\pi^M_{\emptyset}(0) = 1 - 2\varepsilon$.

Entering into any round $t \geq 1$, the actor has reputation $\mathcal{D}_1$ if and only if: in the previous round $t-1$, (i) the trust game was drawn, with probability q, (ii) she was trusted, with probability 0 if her reputation was $\mathcal{D}_1$, probability 1 if her reputation was $\mathcal{C}_2$, and probability θ if her reputation was $\emptyset$, and (iii) she was observed, with probability $\mathbf{p_1}$.

It follows that, for any $t \geq 1$:

$$\pi^M_{\mathcal{D}_1}(t) = q \times [0 \times \pi^M_{\mathcal{D}_1}(t-1) + 1 \times \pi^M_{\mathcal{C}_2}(t-1) + \theta \times \pi^M_{\emptyset}(t-1)] \times \mathbf{p_1}$$

Entering into any round $t \geq 1$, the actor has reputation $\mathcal{C}_2$ if and only if: (i) the institution game was drawn, with probability $1-q$, and (ii) she was observed, with probability p_2. It follows that, for all $t \geq 1$:

$$\pi^M_{\mathcal{C}_2}(t) = (1-q) \times p_2$$

These equations show that the actor's reputation follows a Markov process, as, for any $t \geq 1$, $\pi^M_{\mathcal{C}_1}(t)$, $\pi^M_{\mathcal{C}_2}(t)$ and $\pi^M_{\emptyset}(t) = 1 - \pi^M_{\mathcal{D}_1}(t) - \pi^M_{\mathcal{C}_2}(t)$ each only depend on the probabilities for round $t-1$.

Using both of the above equations, the steady state probabilities $\pi^M_{\mathcal{D}_1}, \pi^M_{\mathcal{C}_2}$, and $\pi^M_{\emptyset}$ must verify:

$$\begin{aligned}\pi^M_{\mathcal{D}_1} &= q(\pi^M_{\mathcal{C}_2} + \theta\pi^M_{\emptyset})\mathbf{p}_1\\ \pi^M_{\mathcal{C}_2} &= (1-q)p_2\\ \pi^M_{\emptyset} &= 1 - \pi^M_{\mathcal{D}_1} - \pi^M_{\mathcal{C}_2}\end{aligned}$$

Replacing in the first equation, we deduce:

$$\begin{aligned}\pi^M_{\mathcal{D}_1} &= q(\pi^M_{\mathcal{C}_2} + \theta(1 - \pi^M_{\mathcal{D}_1} - \pi^M_{\mathcal{C}_2}))\mathbf{p}_1\\ \pi^M_{\mathcal{D}_1}(1 + q\theta\mathbf{p}_1) &= q(\pi^M_{\mathcal{C}_2}(1-\theta) + \theta)\mathbf{p}_1\\ \pi^M_{\mathcal{D}_1} &= q\mathbf{p}_1\frac{\pi^M_{\mathcal{C}_2}(1-\theta) + \theta}{1 + q\theta\mathbf{p}_1}\end{aligned}$$

Using $\pi^M_{C_2} = (1-q)p_2$, we deduce all three proposed equations.

□

Lemma 3.6: Long-run reputation for a low patience actor

Assume that the chooser plays $\sigma^{inst,\theta}_{ch}$, with $\theta \in [0,1]$, that $\sigma = (\sigma^{inst,\theta}_{ch}, \sigma_{act}) \in \mathcal{S}$, and that $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$. For any $\delta < \hat{\delta}_2(\theta)$, the actor's reputation follows a Markov process with three states: $\mathcal{D}_1$, $\mathcal{D}_2$ and $\emptyset$. The steady state $\pi^L \equiv (\pi^L_{\mathcal{D}_1}, \pi^L_{\mathcal{D}_2}, \pi^L_{\emptyset})$ of the actor's reputation is:

$$\pi^L_{\mathcal{D}_1} \equiv q\theta\mathbf{p}_1\frac{1-(1-q)p_2}{1+q\theta\mathbf{p}_1} \tag{3.16}$$

$$\pi^L_{\mathcal{D}_2} \equiv (1-q)p_2 \tag{3.17}$$

$$\pi^L_{\emptyset} \equiv 1 - q\theta\mathbf{p}_1\frac{1-(1-q)p_2}{1+q\theta\mathbf{p}_1} - (1-q)p_2 \tag{3.18}$$

Proof. Take our strategy profile $\sigma = (\sigma^{inst,\theta}_{ch}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma^{inst,\theta}_{ch}$. We assume $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$, and consider a discount factor $\delta < \hat{\delta}_2(\theta)$.

The actor plays D_1 and D_2 (or doesn't play) throughout the game. Her reputation can thus take three values during her life: $\mathcal{D}_1$, $\mathcal{D}_2$, and $\emptyset$. For every $t \geq 0$, we note $\pi^L_{\mathcal{D}_1}(t)$, $\pi^L_{\mathcal{D}_2}(t)$ and $\pi^L_{\emptyset}(t)$ the probability that this low patience actor is in state $\mathcal{D}_1$, $\mathcal{D}_2$, and $\emptyset$, respectively. By assumption, $\pi^L_{\mathcal{D}_1}(0) = \varepsilon = \pi^L_{\mathcal{D}_2}(0)$, and $\pi^L_{\emptyset}(0) = 1 - 2\varepsilon$.

Entering into any round $t \geq 1$, the actor has reputation $\mathcal{D}_1$ if and only if: in the previous round $t-1$, (i) the trust game was drawn, with probability q, (ii) she was trusted, with probability 0 if her reputation was $\mathcal{D}_1$ or $\mathcal{D}_2$, and probability θ if her reputation was $\emptyset$, and (iii) she was observed, with probability $\mathbf{p}_1$.

It follows that, for any $t \geq 1$:

$$\pi^L_{\mathcal{D}_1}(t) = q \times [0 \times \pi^L_{\mathcal{D}_1}(t-1) + 0 \times \pi^L_{\mathcal{D}_2}(t-1) + \theta \times \pi^L_{\emptyset}(t-1)] \times \mathbf{p}_1$$

Entering into any round $t \geq 1$, the actor has reputation $\mathcal{D}_2$ if and only if: (i) the institution game was drawn, with probability $1-q$, and (ii) she was observed, with probability p_2. It follows that, for all $t \geq 1$:

$$\pi^L_{\mathcal{D}_2}(t) = (1-q) \times p_2$$

These equations show that the actor's reputation follows a Markov process, as, for any $t \geq 1$, $\pi^L_{\mathcal{C}_1}(t)$, $\pi^L_{\mathcal{D}_2}(t)$ and $\pi^L_{\emptyset}(t) = 1 - \pi^L_{\mathcal{D}_1}(t) - \pi^L_{\mathcal{D}_2}(t)$ each only depend on the probabilities for round $t-1$.

Using both of the above equations, the steady state probabilities $\pi^L_{\mathcal{D}_1}, \pi^L_{\mathcal{D}_2}$, and $\pi^L_{\emptyset}$ must verify:

$$\begin{aligned}\pi^L_{\mathcal{D}_1} &= q\theta\pi^L_{\emptyset}\mathbf{p}_1\\ \pi^L_{\mathcal{D}_2} &= (1-q)p_2\\ \pi^L_{\emptyset} &= 1 - \pi^L_{\mathcal{D}_1} - \pi^L_{\mathcal{D}_2}\end{aligned}$$

Replacing in the first equation, we deduce:

$$\begin{aligned}\pi^L_{\mathcal{D}_1} &= q\theta(1-\pi^L_{\mathcal{D}_1}-\pi^L_{\mathcal{D}_2})\mathbf{p}_1\\ \pi^L_{\mathcal{D}_1}(1+q\theta\mathbf{p}_1) &= q\theta(1-\pi^L_{\mathcal{D}_2})\mathbf{p}_1\\ \pi^L_{\mathcal{D}_1} &= q\theta\mathbf{p}_1\frac{1-\pi^L_{\mathcal{D}_2}}{1+q\theta\mathbf{p}_1}\end{aligned}$$

Using $\pi^L_{D_2} = (1-q)p_2$, we deduce all three proposed equations.

□

3.7 Long-run level of cooperation

We deduce the long-run level of cooperation $\overline{LC}$, defined as the probability that, in the steady state, the actor is trusted and then cooperates, taking only the distribution of time preferences as given (and not any specific value δ). $\overline{LC}$ can be understood as the long-run level of cooperation in the trust game. With an infinite population, this is the fraction of actor-chooser pairings where the chooser trusts and the actor reciprocates that trust, once a large number of rounds have already been played.

The long-run level of cooperation is the main output we rely on our main document and our numerical resolution, in section 6.

Lemma 3.7: Long-run level of cooperation

Assume that the chooser plays $\sigma^{inst,\theta}_{ch}$, with $\theta \in [0,1]$, that $\sigma = (\sigma^{inst,\theta}_{ch}, \sigma_{act}) \in \mathcal{S}$, and that $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$. Then:

$$\overline{LC} = \mathbf{P}(\delta \geq \hat{\delta}_1(\theta)) \times (\pi^H_{\mathcal{C}_1} + \pi^H_{\mathcal{C}_2} + \theta\pi^H_{\emptyset}) \tag{3.19}$$

Proof. Take our strategy profile $\sigma = (\sigma^{inst,\theta}_{ch}, \sigma_{act}) \in \mathcal{S}$, which we have assumed is subgame perfect, with the chooser playing according to $\sigma^{inst,\theta}_{ch}$. We assume $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$.

As we have shown in Proposition 3.1, in any given round, the actor cooperates if and only if her discount rate is greater than $\hat{\delta}_1(\theta)$, which occurs with probability $\mathbf{P}(\delta \geq \hat{\delta}_1(\theta))$. In that case, her reputation alternates between $\mathcal{C}_1$, $\mathcal{C}_2$ and $\emptyset$, in which case, given that the trust game is drawn, she is respectively trusted with probability 1, 1, and θ.

The steady state of her reputation is given by Lemma 3.4, and the above formula immediately follows. □

4 Institution equilibrium: chooser strategy and domain of existence

In this section, we characterize the domain of existence of the institution equilibrium. Previously, in section 2, we showed that σ^*_{ch} is determined whenever second-order cooperation occurs with non-null probability, and that the only remaining degree of liberty for the chooser is the probability that he trusts given empty reputation θ. In section 3, we showed that the actor's strategy is then fully determined, and can be described by two thresholds, $\hat{\delta}_1(\theta)$ and $\hat{\delta}_2(\theta)$.
Here, we show how to compute the equilibrium value of θ, and derive the conditions under which the chooser does not benefit from deviation from σ^*_{ch}. We deduce the domain of existence of the institution equilibrium.

4.1 Objective

Throughout this section, we assume that, whatever her reputation, the actor cooperates if and only if her discount factor is greater or equal than $\hat{\delta}_1(\theta)$, and contributes if and only if her discount factor is greater or equal than $\hat{\delta}_2(\theta)$, where $\theta \in [0,1]$. We note such a strategy $\sigma^{inst,\theta}_{act}$. As we just saw, following Propositions 3.1 and 3.2, this is the form that the actor's strategy must take in the institution equilibrium in which the chooser trusts given empty reputation with probability θ, i.e. plays according to $\sigma^{inst,\theta}_{ch}$.

In other words, we have shown that the institution equilibrium must be a strategy profile of the form $\sigma^\theta \equiv (\sigma^{inst,\theta}_{ch}, \sigma^{inst,\theta}_{act}) \in \mathcal{S}$, where $\theta \in [0,1]$. In this section, we characterize the domain of existence of the institution equilibrium, by showing how to compute the equilibrium value of θ (Lemma **??**), thus specifying the strategy profile.

We then derive the conditions under which this strategy profile—the institution equilibrium—is subgame perfect (Proposition 4.2).

4.2 Predictive value of $R \in \mathcal{R}$

We begin by deriving the predictive value of a piece of information R—the extent to which the chooser can infer that the actor will cooperate given reputation R, which will tell us whether or not he trusts in equilibrium. We do so for all non-empty reputations, at any point in time (Lemma 4.1), at in the steady state for the empty reputation (Lemma 4.2).

Lemma 4.1: Inferring trustworthiness from non-empty reputations

Assume that both players play according to the strategy profile $\sigma^\theta = (\sigma_{ch}^{inst,\theta}, \sigma_{act}^{inst,\theta})$ for a certain $\theta \in [0,1]$, such that $\hat{\delta}_1(\theta) < 1$ and $\hat{\delta}_2(\theta) < 1$. For any $t \geq 0$:

$$\begin{aligned}
\mathbf{P}^t(\sigma_{act}^{inst,\theta}(\mathcal{T}, \mathcal{D}_1, \delta) = C_1 \mid \mathcal{D}_1, \delta \sim \Delta) &= 0 \\
\mathbf{P}^t(\sigma_{act}^{inst,\theta}(\mathcal{T}, \mathcal{C}_1, \delta) = C_1 \mid \mathcal{C}_1, \delta \sim \Delta) &= 1 \\
\mathbf{P}^t(\sigma_{act}^{inst,\theta}(\mathcal{T}, \mathcal{D}_2, \delta) = C_1 \mid \mathcal{D}_2, \delta \sim \Delta) &= \mathbf{P}(\delta \geq \hat{\delta}_1(\theta) \mid \delta < \hat{\delta}_2(\theta)) \\
\mathbf{P}^t(\sigma_{act}^{inst,\theta}(\mathcal{T}, \mathcal{C}_2, \delta) = C_1 \mid \mathcal{C}_2, \delta \sim \Delta) &= \mathbf{P}(\delta \geq \hat{\delta}_1(\theta) \mid \delta \geq \hat{\delta}_2(\theta))
\end{aligned}$$

In the institution equilibrium, every non-empty reputation is a stationary predictor of cooperation.

Proof. Take $\theta \in [0,1]$, and consider the strategy profile $\sigma^\theta = (\sigma_{ch}^{inst,\theta}, \sigma_{act}^{inst,\theta})$. We assume that $\hat{\delta}_1(\theta) < 1$ and $\hat{\delta}_2(\theta) < 1$.

Since $\hat{\delta}_1(\theta) < 1$ and $\hat{\delta}_2(\theta) < 1$, the actor cooperates and contributes with positive probability. In addition, since $\hat{\delta}_1(\theta) > 0$ and $\hat{\delta}_2(\theta) > 0$ are both always true, the actor also defects and free-rides with positive probability.

We begin by noting that all four conditions are true by assumption for $t = 0$, because of how we have modeled the actor's initial reputation.

If $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$, we can then separate the cases where the actor is high, medium and low patience as in Lemmas 3.4-3.6, each of which occurs with positive probability. As in the proofs of those lemmas, for any $t \geq 1$, we can define the probability that the high type is in reputation $\mathcal{C}_1, \mathcal{C}_2$ or $\emptyset$, that the medium type is in reputation $\mathcal{D}_1, \mathcal{C}_2$ or $\emptyset$, and that the low type is in reputation $\mathcal{D}_1$, $\mathcal{D}_2$ or $\emptyset$.

If $(c_2/p_2) \geq (\mathbf{c_1}/\mathbf{p}_1)$, we can also similarly define these probabilities—the only difference being that the medium type will then alternate between reputations $\mathcal{C}_1$, $\mathcal{D}_2$ and $\emptyset$, and that there may be a 0 probability that the actor is of medium patience (in the equality case).

For the first two conditions, all we need to note is that, in either of those two cases, for any $t \geq 1$:

$$\begin{aligned}
\mathbf{P}^t(\mathcal{D}_1) &> 0 \\
\mathbf{P}^t(\mathcal{C}_1) &> 0
\end{aligned}$$

Starting from round 1, we are guaranteed that the reputations $\mathcal{C}_1$ and $\mathcal{D}_1$ occur with positive probability. Since the actor's strategy is stationary, she can only attain reputation $\mathcal{C}_1$ (resp. $\mathcal{D}_1$), if she is sufficiently patient (resp. impatient), in which case she can be expected to reciprocate again (cheat again) if given the chance.

In other words, we immediately deduce that, for any $t \geq 1$:

$$\begin{aligned}
\mathbf{P}^t(\sigma_{act}^{inst,\theta}(\mathcal{T}, \mathcal{D}_1, \delta) = C_1 \mid \mathcal{D}_1, \delta \sim \Delta) &= 0 \\
\mathbf{P}^t(\sigma_{act}^{inst,\theta}(\mathcal{T}, \mathcal{C}_1, \delta) = C_1 \mid \mathcal{C}_1, \delta \sim \Delta) &= 1
\end{aligned}$$

To obtain the last two equations, let us note that, whatever the case above and the types of actors obtained, for any $t \geq 1$:

$$\begin{aligned}
\mathbf{P}^t(\mathcal{D}_2) &= (1-q)p_2 \times \mathbf{P}(\delta \geq \hat{\delta}_2(\theta)) \\
\mathbf{P}^t(\mathcal{C}_2) &= (1-q)p_2 \times \mathbf{P}(\delta < \hat{\delta}_2(\theta))
\end{aligned}$$

When the institution game is drawn, with probability $(1-q)$, the actor acts, and with probability p_2, her reputation is updated to the non-empty reputation which reflects her action, and therefore whether or not her discount factor

is greater than the threshold for second-order cooperation. The probability of obtaining either $\mathcal{C}_2$ or $\mathcal{D}_2$ does not depend on t, because the actor's ability to play in the institution game does not depend on her current reputation.

We deduce, for instance (recall that we only include the notation $\mid \delta \sim \Delta$ when we need to, to be explicit about not considering any specific value of δ):

$$\begin{aligned}\mathbf{P}^t(\sigma_{act}^{inst,\theta}(\mathcal{T},\mathcal{C}_2,\delta)=C_1 \mid \mathcal{C}_2,\delta\sim\Delta) &= \frac{\mathbf{P^t}(\sigma_{act}^{inst,\theta}(\mathcal{T},\mathcal{C}_2,\delta)=C_1,\mathcal{C}_2 \mid \delta\sim\Delta)}{\mathbf{P}^t(\mathcal{C}_2)}\\ &= \frac{\mathbf{P}(\delta\geq\hat{\delta}_1(\theta),\delta\geq\hat{\delta}_2(\theta))\times(1-q)p_2}{(1-q)p_2\times\mathbf{P}(\delta\geq\hat{\delta}_2(\theta))}\end{aligned}$$

This is obtained using the fact that an actor who plays C_1 can only hold reputation $\mathcal{C}_2$ in round t if she contributes in the institution game—i.e. if $\delta\geq\hat{\delta}_2(\theta)$—and if she was previously observed in that game, with probability $(1-q)p_2$. Simplifying, we deduce:

$$\begin{aligned}\mathbf{P}^t(\sigma_{act}^{inst,\theta}(\mathcal{T},\mathcal{C}_2,\delta)=C_1 \mid \mathcal{C}_2,\delta\sim\Delta) &= \frac{\mathbf{P}(\delta\geq\hat{\delta}_1(\theta),\delta\geq\hat{\delta}_2(\theta))}{\mathbf{P}(\delta\geq\hat{\delta}_2(\theta))}\\ &= \mathbf{P}(\delta\geq\hat{\delta}_1(\theta)\mid\delta\geq\hat{\delta}_2(\theta))\end{aligned}$$

We similarly show the third condition, by considering a cooperative actor whose current reputation is $\mathcal{D}_2$. Note that one of these conditional probabilities will always be equal to 0, depending on which of the thresholds is greater. When for instance second-order cooperation is easier than first-order cooperation, i.e. when $\hat{\delta}_2(\theta)<\hat{\delta}_1(\theta)$, then $\mathbf{P}(\delta\geq\hat{\delta}_1(\theta)\mid\delta<\hat{\delta}_2(\theta))=0=\mathbf{P}(\delta\geq\hat{\delta}_1(\theta),\delta<\hat{\delta}_2(\theta))$.

□

Lemma 4.2: Inferring trustworthiness from the empty reputation

Assume that both players play according to the strategy profile $\sigma^\theta=(\sigma_{ch}^{inst,\theta},\sigma_{act}^{inst,\theta})$ for a certain $\theta\in[0,1]$, such that $\hat{\delta}_2(\theta)<\hat{\delta}_1(\theta)<1$. We note $\mathbf{P}^\infty(C_1\mid\emptyset)\equiv\mathbf{P}^\infty(\sigma_{act}^{inst,\theta}(\mathcal{T},\emptyset,\delta)=C_1\mid\emptyset,\delta\sim\Delta)$ the long-run probability that the actor cooperates if trusted, given that her reputation is empty. Then:

$$\mathbf{P}^\infty(C_1\mid\emptyset)=\frac{\mathbf{P}(\delta\geq\hat{\delta}_1(\theta))\pi_\emptyset^H}{\mathbf{P}(\delta\geq\hat{\delta}_1(\theta))\pi_\emptyset^H+\mathbf{P}(\hat{\delta}_1(\theta)>\delta\geq\hat{\delta}_2(\theta))\pi_\emptyset^M+\mathbf{P}(\hat{\delta}_2(\theta)>\delta)\pi_\emptyset^L} \tag{4.1}$$

In the institution equilibrium, when second-order cooperation remains easier than first-order cooperation, the informative value of the empty reputation in the steady state can be calculated using the long-run reputation of the high, medium and low patience actors, as defined in Lemmas 3.4-3.6.

Proof. Take $\theta\in[0,1]$, and consider the strategy profile $\sigma^\theta=(\sigma_{ch}^{inst,\theta},\sigma_{act}^{inst,\theta})$. We assume that $\hat{\delta}_1(\theta)<1$ and $\hat{\delta}_2(\theta)<\hat{\delta}_2(\theta)$.

We are then in the case where second-order cooperation is easier than first-order cooperation. Following Lemmas 3.4-3.6, there are three cases for the actor, all of which occur with positive probability: high patience, with probability $\mathbf{P}(\delta\geq\hat{\delta}_1(\theta))$; medium patience, with probability $\mathbf{P}(\hat{\delta}_1(\theta)>\delta\geq\hat{\delta}_2(\theta))$; and low patience, with probability $\mathbf{P}(\hat{\delta}_2(\theta)>\delta)$.

Using the notations from these lemmas, the probability that the actor has empty reputation in the steady state is equal to:

$$\mathbf{P}^\infty(\emptyset)=\mathbf{P}(\delta\geq\hat{\delta}_1(\theta))\times\pi_\emptyset^H+\mathbf{P}(\hat{\delta}_1(\theta)>\delta\geq\hat{\delta}_2(\theta))\times\pi_\emptyset^M+\mathbf{P}(\hat{\delta}_2(\theta)>\delta)\times\pi_\emptyset^H$$

The probability that the actor cooperates and has empty reputation in the state is equal to the probability that she is of high patience and has empty reputation:

$$\mathbf{P}^\infty(\sigma_{act}^{inst,\theta}(\mathcal{T},\emptyset,\delta)=C_1,\emptyset\mid\delta\sim\Delta)=\mathbf{P}(\delta\geq\hat{\delta}_1(\theta))\times\pi_\emptyset^L$$

The informative value of the empty reputation in the steady state is then equal to:

$$\mathbf{P}^\infty(\sigma_{act}^{inst,\theta}(\mathcal{T},\emptyset,\delta)=C_1\mid\emptyset,\delta\sim\Delta)=\frac{\mathbf{P}^\infty(\sigma_{act}^{inst,\theta}(\mathcal{T},\emptyset,\delta)=C_1,\emptyset\mid\delta\sim\Delta)}{\mathbf{P}^\infty(\emptyset)}$$

Replacing, we deduce the proposed condition. □

4.3 Long-run chooser payoff

Lemma 4.3: Long-run chooser payoff

Assume that both players play according to the strategy profile $\sigma^\theta = (\sigma_{ch}^{inst,\theta}, \sigma_{act}^{inst,\theta})$ for a certain $\theta \in [0,1]$, such that $\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta) < 1$. The long-run probability of each reputation are obtained using the formulas defined in Lemmas 3.4-3.6:

$$\begin{aligned}
\mathbf{P}^\infty(\emptyset) &= \mathbf{P}(\delta \geq \hat{\delta}_1(\theta)) \times \pi_\emptyset^H + \mathbf{P}(\hat{\delta}_1(\theta) > \delta \geq \hat{\delta}_2(\theta)) \times \pi_\emptyset^M + \mathbf{P}(\hat{\delta}_2(\theta) > \delta) \times \pi_\emptyset^L \\
\mathbf{P}^\infty(\mathcal{C}_1) &= \mathbf{P}(\delta \geq \hat{\delta}_1(\theta)) \times \pi_{\mathcal{C}_1}^H \\
\mathbf{P}^\infty(\mathcal{D}_1) &= \mathbf{P}(\hat{\delta}_1(\theta) > \delta \geq \hat{\delta}_2(\theta)) \times \pi_{\mathcal{D}_1}^M + \mathbf{P}(\hat{\delta}_2(\theta) > \delta) \times \pi_{\mathcal{D}_1}^L \\
\mathbf{P}^\infty(\mathcal{C}_2) &= \mathbf{P}(\delta \geq \hat{\delta}_1(\theta)) \times \pi_{\mathcal{C}_2}^H + \mathbf{P}(\hat{\delta}_1(\theta) > \delta \geq \hat{\delta}_2(\theta)) \times \pi_{\mathcal{C}_2}^M \\
\mathbf{P}^\infty(\mathcal{D}_2) &= \mathbf{P}(\hat{\delta}_2(\theta) > \delta) \times \pi_{\mathcal{D}_2}^L
\end{aligned}$$

The long-run payoff of the chooser u^∞, defined as the expected value of the chooser's payoff over every possible reputation in the steady state is equal to:

$$u^\infty \equiv \theta \times \mathbf{P}^\infty(\emptyset)(-k + \mathbf{P}^\infty(C_1 \mid \emptyset)b) + \mathbf{P}^\infty(\mathcal{C}_1)(-k+b) + \mathbf{P}^\infty(\mathcal{C}_2)(-k + \mathbf{P}(\delta \geq \hat{\delta}_1(\theta) \mid \delta \geq \hat{\delta}_2(\theta))b) \quad (4.2)$$

Proof. Take $\theta \in [0,1]$, and consider the strategy profile $\sigma^\theta = (\sigma_{ch}^{inst,\theta}, \sigma_{act}^{inst,\theta})$. We assume that $\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta) < 1$.

In the previous lemma, we used the formulas from Lemmas 3.4-3.6 to calculate $\mathbf{P}^\infty(\emptyset)$. We can similarly calculate the probability of each reputation, taking into account the types of actors that are able to achieve that reputation (e.g., only high patience actors can achieve reputation $\mathcal{C}_1$)—this yields the five conditions above.

We deduce the long-run payoff of the chooser u^∞ by taking into account that the chooser trusts with probability θ given $\emptyset$, and with probability 1 given $\mathcal{C}_1$ and $\mathcal{C}_2$, and by taking into account the probabilities that the actor reciprocates given each of these reputations in the steady state, all three of which we have already calculated. □

4.4 Equilibrium value of θ

The previous Lemma gives the long-run predictive value of $\emptyset$ in the case of interest—the institution equilibrium, given that second-order cooperation remains easier than first-order cooperation. (To derive the formula given $(c_2/p_2) \geq (\mathbf{c_1}/\mathbf{p_1})$, one need just to specify the long-run repuation of a medium actor in this case—i.e. an actor who plays C_1 and D_2).

By assumption, in a subgame perfect equilibrium, the chooser optimizes based on this long-run predictive value. We deduce an algorithm for determining the equilibrium value of θ.

Note that the algorithm can be underspecified: there could be two values $t \in (0,1)$ such that $-k + b \times \mathbf{P}^\infty(\sigma_{act}^{inst,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) = 0$. In the numerical analysis below, in section 6, we consider specific parameter regions (namely, those for which $\mathbf{p}_1$ is not too large) in which this formula is a strictly decreasing function of t—meaning that the algorithm will yield a unique value θ^*, and therefore a unique candidate profile for the institution equilibrium.

Proposition 4.1: Equilibrium value of θ

Assume that both players play according to the strategy profile $\sigma^\theta = (\sigma_{ch}^{inst,\theta}, \sigma_{act}^{inst,\theta})$ for a certain $\theta \in [0,1]$. If $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$ and if $\sigma^\theta \in \mathcal{S}$ is subgame perfect, then θ is given by:

$$\theta = \theta^* \equiv \begin{cases} 0 & \text{if } \forall t \in [0,1], -k + b \times \mathbf{P}^\infty(\sigma_{act}^{inst,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) \leq 0 \\ 1 & \text{if } \forall t \in [0,1], -k + b \times \mathbf{P}^\infty(\sigma_{act}^{inst,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) \geq 0 \\ t & \text{such that } -k + b \times \mathbf{P}^\infty(\sigma_{act}^{inst,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) = 0 \end{cases} \quad (4.3)$$

Proof. Take $\theta \in [0,1]$, and consider the strategy profile $\sigma^\theta = (\sigma_{ch}^{inst,\theta}, \sigma_{act}^{inst,\theta})$. We assume that $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$, guaranteeing that $\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta)$, and that we can calculate the long-run predictive value of the empty reputation as in the Lemma above.

If the chooser trusts given empty reputation in the steady state, he loses k, and receives back b if and only if the actor ends up reciprocating. In other words, on average, he earns:

$$-k + b \times \mathbf{P}^{\infty}(\sigma_{act}^{inst,\theta}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta)$$

If he does not trust, he earns 0 payoff.

If σ^{θ} is subgame perfect, then there cannot be any beneficial deviations for the chooser given $\emptyset$, or the actor. If $\theta = 0$, we deduce that we must have:

$$-k + b \times \mathbf{P}^{\infty}(\sigma_{act}^{inst,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) \leq 0 \qquad \forall t \in [0,1]$$

Indeed, when the chooser plays according to $\sigma_{ch}^{inst,t}$ for a certain $t \in [0,1]$, the actor must play according to $\sigma_{act}^{inst,t}$ in a subgame perfect equilibrium, as we have already shown. Were the above not to be true, there would be a beneficial deviation for the chooser: to trusting with positive probability $t > 0$ such that this inequality is unverified (note that the above is a continuous function of t—if it is positive for $t = 0$, then it is also positive in a vicinity of 0).

Similarly, for $\theta = 1$ to occur in a subgame perfect equilibrium, we must have:

$$-k + b \times \mathbf{P}^{\infty}(\sigma_{act}^{inst,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) \geq 0 \qquad \forall t \in [0,1]$$

Otherwise, when neither of these two conditions are verified, there exists $t \in [0,1]$ such that the chooser is indifferent between trusting and not trusting given $\emptyset$. This proves that θ must be given by the formula defining θ^*. □

4.5 Domain of existence of the institution equilibrium

Proposition 4.2: Domain of existence of the institution equilibrium

If $(c_2/p_2) < (\mathbf{c_1}/\mathbf{p_1})$, the strategy profile $\sigma^{\theta^*} = (\sigma_{ch}^{inst,\theta^*}, \sigma_{act}^{inst,\theta^*})$, where θ^* is defined as in Proposition 4.1, is a subgame perfect equilibrium if and only if:

$$\hat{\delta}_1(\theta^*) < 1 \tag{4.4}$$

$$\mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*)) \geq \frac{k}{b} \tag{4.5}$$

We obtain an institution equilibrium when second-order cooperation is a sufficiently good predictor of first-order cooperation.

Proof. Consider the strategy profile $\sigma^{\theta^*} = (\sigma_{ch}^{inst,\theta^*}, \sigma_{act}^{inst,\theta^*})$, where θ^* is defined as in Proposition 4.1.

We begin by noting that these two conditions are necessary for σ^{θ^*} to be subgame perfect, by contraposition. Indeed, if $\hat{\delta}_1(\theta^*) \leq 1$, then cooperation occurs with probability 0—whatever the actor's reputation, the chooser benefits from playing $\neg T$, in order to avoid paying the cost of trust $k > 0$ for nothing. In particular, the chooser benefits from deviation to playing $\neg T$ given $\mathcal{C}_1$.

If $\mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*)) < \frac{k}{b}$, then the predictive value of $\mathcal{C}_2$ (following Lemma 4.1) is insufficient. The chooser benefits from deviation to playing $\neg T$ given $\mathcal{C}_2$, and earning 0 instead of $-k + b \times \mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*)) < 0$ on average.

This proves that both conditions are necessary. We prove that they are sufficient, by checking that the chooser does not have any beneficial deviations given a non-empty reputation when both conditions hold.

Since $(c_2/p_2) < \mathbf{c_1}/\mathbf{p_1}$, second-order cooperation remains easier than first-order cooperation, and, in particular, we have $\hat{\delta}_2(\theta^*) < 1$. We are in the conditions of Lemma 4.1: every non-empty reputation is a stationary predictor of cooperation, and there are no beneficial deviations given $\mathcal{C}_1$ or $\mathcal{D}_1$ (since these are perfect predictors of cooperation and defection, respectively).

Given $\mathcal{C}_2$, the chooser earns on average $-k + b \times \mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*))$ as detailed above, which is positive or null—he does not benefit from deviation to playing $\neg T$ given $\mathcal{C}_2$.

Finally, given $\mathcal{D}_2$, the chooser does not trust, earning null payoff. Since second-order cooperation is easier, deviation to playing T in such a case is costly: the chooser would then pay k to earn back b with probability $0 = \mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta < \hat{\delta}_2(\theta^*))$.

This proves that the chooser does not have any beneficial deviations given a non-empty reputation. Due to how θ^* is defined, he has no beneficial deviations given $\emptyset$ either. Finally, there are no beneficial deviations for the actor, as we have shown in section 3: σ^{θ^*} is subgame perfect when $\hat{\delta}_1(\theta^*) < 1$ and $\mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*)) \geq \frac{k}{b}$. □

5 Baseline equilibrium

In this section, we characterize the non-institution equilibrium in a specific case—namely $p_2 = 0$. We call the obtained equilibrium the baseline equilibrium; in the next section, we will use it as a basis for comparison of our numerical results in the institution equilibrium.

5.1 Objective

In section 2, we showed that two types of cooperative equilibria were possible: the institution equilibrium, studied in sections 3-4; and 'non-institution equilibria', in which reputation incentivizes first-order cooperation but not second-order cooperation.

In such a situation, the actor never contributes (since $c_2 > 0$), and the chooser must trust given $\mathcal{C}_1$ and distrust given $\mathcal{D}_1$. His behavior given second-order reputations is unspecified: he could trust given both $\mathcal{C}_2$—which is the null event—and $\mathcal{D}_2$—which conveys no information on the actor (it just conveys that the institution game was drawn in the round before). We are left with an additional degree of liberty (in addition to the probability of trust given $\emptyset$) —depending on the average actor's trustworthiness (i.e. on $\mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta))$, where this threshold is defined below), we should obtain one or the other non-institution equilibrium.

Here, however, we are interested in establishing a baseline for comparison of our results in the institution equilibrium. To do so, we 'turn off' information coming from the institution game, by assuming that:

$$p_2 = 0 \tag{5.1}$$

We also restrict the possible reputations for the actor to $\mathcal{R}^b \equiv \{\mathcal{C}_1, \mathcal{D}_1, \emptyset\}$, and assume that both players only take elements of this restricted set in their strategies. Similarly to before, we assume that the actor's decision in the initial trust game is observed beforehand with probability ε: an actor who cooperate if trusted starts out with reputation $\mathcal{C}_1$ with probability ε; a cheaters starts out with reputation $\mathcal{D}_1$ with probability ε; and every actor starts out with empty reputation $\emptyset$ with probability $1 - \varepsilon$.

We note $\mathcal{S}^b$ the set of subgame perfect equilibria of our repeated game under these modified assumptions. We note $\sigma_{ch}^{base,\theta}$ the chooser strategy consisting in playing T given $\mathcal{C}_1$, $\neg T$ given $\mathcal{D}_1$, and trusting with probability θ given $\emptyset$. We note σ_{act} the actor's strategy. We call $\sigma \equiv (\sigma_{ch}^{base,\theta}, \sigma_{act}) \in \mathcal{S}^b$ the subgame perfect equilibrium obtained when the chooser plays according $\sigma_{ch}^{base,\theta}$ the **baseline equilibrium**.

Below, we characterize the baseline equilibrium, by listing, and briefly justifying, our results in the same order than in sections 3-4 (among other things, we show that it is uniquely determined as a function of the parameters, like the institution equilibrium, justifying our use of the singular).

Note that we use the superscript b for the main characteristics of the baseline equilibrium (the threshold discount factor, the equilibrium value of θ), and specify that we are in the baseline for each our main results (which are propositions). We do not do this for other values that will be useful in our numerical calculations to not overcharge these notations (e.g., the reputational benefit), which are detailed in the lemmas and proofs below—every such notation refers only to values computed in this section, and not those computed in previous sections for the institution equilibrium.

5.2 Threshold discount factor for first-order cooperation

As we have shown, the actor plays a stationary strategy in a subgame perfect equilibrium, which can be described using two intervals of discount factors I_{D_1} and I_{C_1}, over which the actor respectively cheats if trusted and reciprocates if trusted, whatever her current reputation.

Proposition 5.1: Threshold discount factor for cooperation in the baseline

Assume that the chooser plays $\sigma_{ch}^{base,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{inst,\theta}, \sigma_{act}) \in \mathcal{S}^b$. Then patient ($\delta \geq \hat{\delta}_1^b(\theta)$) actors always cooperate, whatever their reputation, and impatient ($\delta < \hat{\delta}_1^b(\theta)$) actors never cooperate—the threshold discount factor being given by:

$$\hat{\delta}_1^b(\theta) \equiv \frac{c_1}{p_1 q[r - \theta c_1]} \tag{5.2}$$

Proof. The proof is analogous to Proposition 3.1. Using the fact that the actor's strategy is stationary, we similarly partition $(0,1)$ into two intervals I_{D_1} and I_{C_1}, over which the actor respectively always cheats or always cooperates. We can then calculate the reputational benefit R_δ^b in this baseline equilibrium in the same manner than in Lemma 3.2, by plugging in $\beta = \gamma = \pi_1 = 0$ (since the actor never contributes). We obtain::

$$R_\delta = \begin{cases} \frac{qr}{1+q\delta p_1 \theta} & \text{if } \delta \in I_{D_1} \\ \frac{q(r-c_1)}{1-q\delta p_1(1-\theta)} & \text{if } \delta \in I_{C_1} \end{cases} \tag{5.3}$$

Using the same reasoning as in Proposition 3.1, we deduce the proposed value for $\hat{\delta}_1^b(\theta)$ □

Similarly to before, we refer to $\hat{\delta}_1^b(\theta)$ as the difficulty of cooperation in the baseline equilibrium for θ. We also call $\hat{\delta}^b(0) = \max\{\hat{\delta}_1^b(\theta), \theta \in [0,1]\}$ the **intrinsic difficulty of cooperation**, or simply the difficulty of cooperation, and note it δ^b. $\delta^b = (c_1/p_1 qr)$ is a function of our parameters, and characterizes the repeated game as a whole; in contrast to before, it is not defined in relation chooser (equilibrium) strategy. We will use δ^b in our graphical representations, in section 6 (y-axis of each Density Plot).

5.3 Actor payoffs, long-run reputation and level of cooperation

Lemma 5.1: Normalized actor payoff

Assume that the chooser plays $\sigma_{ch}^{base,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{base,\theta}, \sigma_{act}) \in \mathcal{S}^b$. For any $\delta \in (0,1)$, we have:

$$\overline{U}_\delta = \theta \times R_\delta \tag{5.4}$$

Proof. Since $p_2 = 0$, we have: $U_\delta^B = \delta \times U_\delta^\emptyset$ (these quantities are defined analogously to before): when in bad standing, the actor always achieves null standing in the next round—she is either distrusted in the trust game, or faces the institution game and is not observed.

Using $U_\delta^\emptyset - U_\delta^B = \theta R_\delta$ (analogous to condition 3.3), we deduce:

$$\begin{aligned} U_\delta^\theta - \theta R_\delta^b &= \delta \times U_\delta^\theta \\ (1-\delta) \times U_\delta^\theta &= \theta R_\delta \\ \overline{U}_\delta &= \theta R_\delta \end{aligned} \tag{5.4}$$

□

In the baseline equilibrium, the actor's reputation again follows a Markov process—with two cases of interest (high and low patience) this time instead of three.

Lemma 5.2: Long-run reputation for a high patience actor

Assume that the chooser plays $\sigma_{ch}^{base,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{base,\theta}, \sigma_{act}) \in \mathcal{S}^b$.
For any $\delta \geq \hat{\delta}_1^b(\theta)$, the actor's reputation follows a Markov process with two states: $\mathcal{C}_1$ and $\emptyset$. The steady state $\pi^H \equiv (\pi_{\mathcal{C}_1}^H, \pi_\emptyset^H)$ of the actor's reputation is:

$$\pi_{\mathcal{C}_1}^H \equiv \frac{q\theta p_1}{1 - q(1-\theta)p_1} \tag{5.5}$$

$$\pi_\emptyset^H \equiv 1 - \frac{q\theta p_1}{1 - q(1-\theta)p_1} \tag{5.6}$$

Proof. Take the strategy profile $\sigma = (\sigma_{ch}^{base,\theta}, \sigma_{act}) \in \mathcal{S}$. Consider a discount factor $\delta \geq \hat{\delta}_1^b(\theta)$, assuming such a value exists.

In any given round, the actor either plays C_1, or does not play. Her reputation can thus take two values during her life: $\mathcal{C}_1$ and $\emptyset$. For every $t \geq 0$, we note $\pi_{\mathcal{C}_1}^H(t)$ and $\pi_{\emptyset}^H(t)$ the probability that this high patience actor is in state $\mathcal{C}_1$ and $\emptyset$, respectively. By assumption, $\pi_{\mathcal{C}_1}^H(0) = \varepsilon$, and $\pi_{\emptyset}^H(0) = 1 - \varepsilon$.

Entering into any round $t \geq 1$, the actor has reputation $\mathcal{C}_1$ if and only if: in the previous round $t-1$, (i) the trust game was drawn, with probability q, (ii) she was trusted, with probability 1 if her reputation was $\mathcal{C}_1$, and probability θ if her reputation was $\emptyset$, and (iii) she was observed, with probability p_1.

It follows that, for any $t \geq 1$:

$$\pi_{\mathcal{C}_1}^H(t) = q \times [1 \times \pi_{\mathcal{C}_1}^H(t-1) + \theta \times \pi_{\emptyset}^H(t-1)] \times p_1$$

Using this equation, the steady state probabilities $\pi_{\mathcal{C}_1}^H$ and $\pi_{\emptyset}^H$ must verify:

$$\begin{aligned} \pi_{\mathcal{C}_1}^H &= q(\pi_{\mathcal{C}_1}^H + \theta \pi_{\emptyset}^H) p_1 \\ \pi_{\emptyset}^H &= 1 - \pi_{\mathcal{C}_1}^H \end{aligned}$$

Replacing in the first equation, we deduce:

$$\begin{aligned} \pi_{\mathcal{C}_1}^H &= q(\pi_{\mathcal{C}_1}^H + \theta(1 - \pi_{\mathcal{C}_1}^H)) p_1 \\ \pi_{\mathcal{C}_1}^H (1 - q(1-\theta)p_1) &= q\theta p_1 \\ \pi_{\mathcal{C}_1}^H &= \frac{q\theta p_1}{1 - q(1-\theta)p_1} \end{aligned}$$

□

Lemma 5.3: Long-run reputation for a low patience actor

Assume that the chooser plays $\sigma_{ch}^{base,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{base,\theta}, \sigma_{act}) \in \mathcal{S}^b$.
For any $\delta < \hat{\delta}_1^b(\theta)$, the actor's reputation follows a Markov process with two states: $\mathcal{D}_1$ and $\emptyset$. The steady state $\pi^L \equiv (\pi_{\mathcal{D}_1}^L, \pi_{\emptyset}^L)$ of the actor's reputation is:

$$\pi_{\mathcal{D}_1}^L \equiv \frac{q\theta p_1}{1 + q\theta p_1} \tag{5.7}$$

$$\pi_{\emptyset}^L \equiv 1 - \frac{q\theta p_1}{1 + q\theta p_1} \tag{5.8}$$

Proof. Take the strategy profile $\sigma = (\sigma_{ch}^{base,\theta}, \sigma_{act}) \in \mathcal{S}^b$, and consider a discount factor $\delta < \hat{\delta}_1^b(\theta)$.

In any given round, the actor either plays D_1, or does not play. Her reputation can thus take two values during her life: $\mathcal{D}_1$ and $\emptyset$. For every $t \geq 0$, we note $\pi_{\mathcal{D}_1}^L(t)$ and $\pi_{\emptyset}^L(t)$ the probability that this low patience actor is in state $\mathcal{D}_1$ and $\emptyset$, respectively. By assumption, $\pi_{\mathcal{D}_1}^L(0) = \varepsilon$, and $\pi_{\emptyset}^L(0) = 1 - \varepsilon$.

Entering into any round $t \geq 1$, the actor has reputation $\mathcal{D}_1$ if and only if: in the previous round $t-1$, (i) the trust game was drawn, with probability q, (ii) she was trusted, with probability 0 if her reputation was $\mathcal{D}_1$, and probability θ if her reputation was $\emptyset$, and (iii) she was observed, with probability p_1.

It follows that, for any $t \geq 1$:

$$\pi_{\mathcal{D}_1}^L(t) = q \times [0 \times \pi_{\mathcal{D}_1}^L(t-1) + \theta \times \pi_{\emptyset}^L(t-1)] \times p_1$$

Using this equation, the steady state probabilities $\pi_{\mathcal{D}_1}^L$ and $\pi_{\emptyset}^L$ must verify:

$$\begin{aligned} \pi_{\mathcal{D}_1}^L &= q\theta \pi_{\emptyset}^L p_1 \\ \pi_{\emptyset}^L &= 1 - \pi_{\mathcal{D}_1}^L \end{aligned}$$

Replacing in the first equation, we deduce:

$$\begin{aligned} \pi_{\mathcal{D}_1}^L &= q\theta(1 - \pi_{\mathcal{D}_1}^L) p_1 \\ \pi_{\mathcal{D}_1}^L &= \frac{q\theta p_1}{1 + q\theta p_1} \end{aligned}$$

□

Lemma 5.4: Long-run level of cooperation

Assume that the chooser plays $\sigma_{ch}^{base,\theta}$, with $\theta \in [0,1]$, and that $\sigma = (\sigma_{ch}^{base,\theta}, \sigma_{act}) \in \mathcal{S}^b$. Then:

$$\overline{LC} = \mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta)) \times (\pi_{\mathcal{C}_1}^H + \theta \pi_{\emptyset}^H) \tag{5.9}$$

Proof. Immediate, and analogous to lemma 3.7. □

5.4 Chooser inferences and long-run payoff

As above, we note the resulting actor strategy $\sigma_{act}^{base,\theta}$, and the resulting strategy profile $\sigma^{b,\theta} \equiv (\sigma_{ch}^{base,\theta}, \sigma_{act}^{base,\theta})$.

Lemma 5.5: Long-run chooser inferences and payoff

Assume that both players play according to the strategy profile $\sigma^{b,\theta} = (\sigma_{ch}^{base,\theta}, \sigma_{act}^{base,\theta})$ for a certain $\theta \in [0,1]$, such that $\hat{\delta}_1^b(\theta) < 1$. The long-run probability of each reputation are obtained using the formulas defined in Lemmas 5.2-5.3:

$$\begin{aligned}
\mathbf{P}^\infty(\emptyset) &= \mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta)) \times \pi_{\emptyset}^H + \mathbf{P}(\hat{\delta}_1^b(\theta) > \delta) \times \pi_{\emptyset}^L \\
\mathbf{P}^\infty(\mathcal{C}_1) &= \mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta)) \times \pi_{\mathcal{C}_1}^H \\
\mathbf{P}^\infty(\mathcal{D}_1) &= \mathbf{P}(\hat{\delta}_1^b(\theta) > \delta) \times \pi_{\mathcal{D}_1}^L
\end{aligned}$$

We note $\mathbf{P}^\infty(C_1 \mid \emptyset) \equiv \mathbf{P}^\infty(\sigma_{act}^{base,\theta}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta)$ the long-run probability that the actor cooperates if trusted, given that her reputation is empty. Then:

$$\mathbf{P}^\infty(C_1 \mid \emptyset) = \frac{\mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta))\pi_{\emptyset}^H}{\mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta))\pi_{\emptyset}^H + \mathbf{P}(\hat{\delta}_1^b(\theta) > \delta)\pi_{\emptyset}^L} \tag{5.10}$$

The long-run payoff of the chooser u^∞, defined as the expected value of the chooser's payoff over every possible reputation in the steady state is equal to:

$$u^\infty \equiv \theta \times \mathbf{P}^\infty(\emptyset)(-k + \mathbf{P}^\infty(C_1 \mid \emptyset)b) + \mathbf{P}^\infty(\mathcal{C}_1)(-k + b) \tag{5.11}$$

Proof. The long-run probabilities of each reputational state are calculated similarly to before, by distinguishing between high and low patience actors, and using the formulas defined in Lemmas 5.2-5.3.

The long-run probability that the actor cooperates and has an empty reputation is equal to $\mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta)) \times \pi_{\emptyset}^H$—by dividing by $\mathbf{P}^\infty(\emptyset)$, we deduce condition (5.10).

We immediately deduce the long-run payoff of the chooser. □

5.5 Equilibrium value of θ

In contrast to before, we always obtain a unique equilibrium value $\theta^{*,b}$, by using the fact that $\mathbf{P}^\infty(C_1 \mid \emptyset)$ is a decreasing function of θ. (This is the unique value that θ must take in a subgame perfect equilibrium; we derive below the conditions under which we do obtain a baseline equilibrium, given that $\theta = \theta^{*,b}$).

Proposition 5.2: Equilibrium value of θ

Assume that both players play according to the strategy profile $\sigma^\theta = (\sigma_{ch}^{base,\theta}, \sigma_{act}^{base,\theta})$ for a certain $\theta \in [0,1]$. If $\sigma^\theta \in \mathcal{S}$ is subgame perfect, then θ is given by:

$$\theta = \theta^{*,b} \equiv \begin{cases} 0 & \text{if } -k + b \times \mathbf{P}^\infty(\sigma_{act}^{base,0}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) \leq 0 \\ 1 & \text{if } -k + b \times \mathbf{P}^\infty(\sigma_{act}^{base,1}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) \geq 0 \\ t & \text{such that } -k + b \times \mathbf{P}^\infty(\sigma_{act}^{base,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) = 0 \end{cases} \tag{5.12}$$

Proof. Using the same steps as before, we deduce that in a subgame perfect equilibrium, we must have:

$$\theta = \theta^{*,b} \equiv \begin{cases} 0 & \text{if } \forall t \in [0,1], -k + b \times \mathbf{P}^{\infty}(\sigma_{act}^{base,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) \leq 0 \\ 1 & \text{if } \forall t \in [0,1], -k + b \times \mathbf{P}^{\infty}(\sigma_{act}^{base,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) \geq 0 \\ t & \text{such that } -k + b \times \mathbf{P}^{\infty}(\sigma_{act}^{base,t}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta) = 0 \end{cases}$$

In addition, we note that $\mathbf{P}^{\infty}(C_1 \mid \emptyset) = \mathbf{P}^{\infty}(\sigma_{act}^{base,\theta}(\mathcal{T}, \emptyset, \delta) = C_1 \mid \emptyset, \delta \sim \Delta)$ is a strictly decreasing function of θ. Indeed, in a subgame perfect equilibrium, $\hat{\delta}_1^b(\theta)$ must be smaller than 1, and this function is then equal to:

$$\mathbf{P}^{\infty}(C_1 \mid \emptyset) = \frac{1}{1 + \frac{\mathbf{P}(\hat{\delta}_1^b(\theta) > \delta)\pi_{\emptyset}^L}{\mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta))\pi_{\emptyset}^H}}$$

As a function of θ, $\mathbf{P}^{\infty}(C_1 \mid \emptyset)$ varies like $(\frac{\mathbf{P}(\hat{\delta}_1^b(\theta) > \delta)\pi_{\emptyset}^L}{\mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta))\pi_{\emptyset}^H})^{-1} = \frac{\mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta))\pi_{\emptyset}^H}{\mathbf{P}(\hat{\delta}_1^b(\theta) > \delta)\pi_{\emptyset}^L}$.

We conclude by noting first that the proportion of trustworthy actors $\mathbf{P}(\delta \geq \hat{\delta}_1^b(\theta))$ is a strictly decreasing function of θ—the higher the baseline chance of being accepted, the lower the incentive to cooperate—and that the proportion of untrustworthy actors $\mathbf{P}(\hat{\delta}_1^b(\theta) > \delta)$ therefore strictly increases with θ.

We note second that:

$$\begin{aligned} \frac{\pi_{\emptyset}^H}{\pi_{\emptyset}^L} &= \frac{1 + q\theta p_1}{1 - q(1-\theta)p_1} \\ &= \frac{1 + q\theta p_1}{1 + q\theta p_1 - q p_1} \\ &= \frac{1}{1 - \frac{q p_1}{1 + q\theta p_1}} \end{aligned}$$

This fraction varies like $\frac{qp_1}{1+q\theta p_1}$ (since this fraction is on the denominator, but preceded by a minus sign): it is a strictly decreasing function of θ.

$\mathbf{P}^{\infty}(C_1 \mid \theta)$ is therefore the product of two strictly decreasing functions of θ. We deduce that the algorithm for finding $\theta^{*,b}$ can be simplified, by first comparing the maximum value of $-k + b \times \mathbf{P}^{\infty}(C_1 \mid \theta)$—obtained when $\theta = 0$—with 0: if this expression is negative even in this case, then the chooser always stand to lose from trusting given $\emptyset$, whatever the value of θ—and the only possibility in equilibrium is that he does not trust, i.e. $\theta^{*,b} = 0$. Similarly, if the minimum value—obtained when $\theta = 1$—is positive, then the only possibility $\theta^{*,b} = 1$. Otherwise, following the intermediate value theorem, we deduce a unique value for $\theta^{*,b} = 1$ inside the interval $(0, 1)$. □

5.6 Domain of existence

Proposition 5.3: Domain of existence of the baseline equilibrium

The strategy profile $(\sigma_{ch}^{base,\theta^{*,b}}, \sigma_{act}^{base,\theta^{*,b}})$, where $\theta^{*,b}$ is defined as in Proposition 5.2, is a subgame perfect equilibrium if and only if:

$$\hat{\delta}_1(\theta^{*,b}) < 1 \tag{5.13}$$

$$\theta^{*,b} > 0 \tag{5.14}$$

We obtain the baseline equilibrium as long as cooperation occurs with positive probability, and choosers trust given empty reputation with positive probability.

Proof. The proof is analogous, and simpler, to the proof of Proposition 4.2. In this section, we have derived necessary conditions to obtain a cooperative equilibrium, given that $p_2 = 0$ (as we assume throughout this section). We have shown that when the chooser trusts given $\mathcal{C}_1$, distrusts given $\mathcal{D}_1$, and trusts with probability θ given $\emptyset$, then the actor's strategy is fully determined, as a function of θ (Proposition 5.1). We have also shown that there are no beneficial deviations for choosers given empty reputation as long as $\theta = \theta^{*,b}$ (Proposition 5.2).

All that is left to look for are beneficial chooser deviations given $\mathcal{C}_1$ or $\mathcal{D}_1$. If $\hat{\delta}_1(\theta^{*,b}) \geq 1$, then there are no trustworthy actors: the actor's δ is above the threshold with probability 0, and she never reciprocates if trusted. It is then beneficial for the chooser to distrust even given $\mathcal{C}_1$; this proves that the above condition is necessary.

Conversely, if $\hat{\delta}_1(\theta^{*,b}) < 1$, there are two cases. First case: $\theta^{*,b} > 0$. In this case, things are similar to the institution equilibrium: with positive probability, the actor's threshold is above or under the threshold, and, the actor's reputation is equal to $\mathcal{C}_1$ or $\mathcal{D}_1$ in any given round and in the steady state—as can be seen using the formulas derives in the proofs of Lemmas 5.2-5.3. Since the actor's strategy is stationary, the chooser always strictly benefits from trusting given $\mathcal{C}_1$ and distrusting given $\mathcal{D}_1$: we have a subgame perfect equilibrium.

Second case: $\theta^{*,b} = 0$. In this case, things are different than before (we did not obtain this case before, because the institution game with $p_2 > 0$ allowed non-empty information to be retained throughout the repeated game). Using the formulas derived above, a patient actor's reputation is $\mathcal{C}_1$ in a given round t with probability $(qp_1)^t \times \varepsilon$, and equal to 0 in the steady state. An impatient actor's reputation is $\mathcal{D}_1$ is equal to 0 from round 1, and equal to 0 in the steady state.

In the steady state, the actor's reputation is empty with certainty. Since $\theta^{ast,b} = 0$, this means that $-k+b\times\mathbf{P}(\delta \geq \hat{\delta}_1^b(0) \leq 0$, using the definition in Proposition 5.2. The chooser thus benefits from not trusting an individual at random, in the absence of any specific information.

In any given round $t > 0$, the chooser then benefits from trusting given $\mathcal{C}_1$ (because this event is non-null and perfectly predictive of the actor's trustworthiness), and also from not trusting given $\mathcal{D}_1$ (because the chooser benefits from not trusting at random, and therefore also from not trusting given that $\mathcal{C}_1$ is false).

In the steady state however, the same argument shows that the chooser benefits from deviation to playing $\neg T$ given $\mathcal{C}_1$, which is then the null event. Since we rely on the steady state to define the relevant chooser payoffs, this strategy profile cannot be subgame perfect. □

As the above proof highlights, we do not obtain a subgame perfect equilibrium when $\theta^{*,b} = 0$ for, arguably, technical reasons—because we do not make any assumptions on out-of-equilibrium beliefs (section 2.1.6), and because we look at chooser payoffs in the steady state even for $\mathcal{C}_1$ (section 2.1.5). (We have written this demonstration to show the specific point where this assumption matters.)

With less conservative assumptions, we could have obtained an equilibrium when $\theta^{*,b} = 0$, but $\hat{\delta}^b(\theta^{*,b}) < 1$. However, this concern proves to be moot: all of the outputs of our model—the level of cooperation, the chooser's long-run payoff, and the actor's normalized payoff—are equal to 0 in this case. Whether or not we take this conservative view of the domain of existence of the baseline equilibrium does not affect the numerical values we obtain below, and use to plot our mathematical results.

6 Implementation into Mathematica

The results obtained in the previous sections are under-specified. They notably depend on the specific distribution of time preferences, on the specific allocation of incentives performed by the institution (as captured by the relative weights β, γ, and π_1), and on its effectiveness (as captured by ρ).
In this section, we outline the algorithm that we use to compute our numerical results using the software Mathematica, and compare our baseline results with four different types of institution.

6.1 Motivation and general algorithm

The results obtained in the previous sections are under-specified. In both the baseline and institution equilibria, the probability that the actor will reciprocate if trusted—or the fraction of trustworthy actors with an infinite population model—depend on the distribution of time preferences, and the equilibrium value of θ; which itself depends on the fraction of trustworthy actors, and therefore the distribution of time preferences.

Using Mathematica, we resolve our model numerically as detailed in the file called `Institutions.nb`. For both of our equilibria, we calculate the threshold(s) defining actor strategy ($\hat{\delta}^b(\theta)$ or $\hat{\delta}_1(\theta)$ and $\hat{\delta}_2(\theta)$), as well as the actor's normalized payoff, the chooser's long-run payoff, the steady state probabilities, and the level of cooperation in the steady state—each as a function of θ and other relevant variables, including the fraction of trustworthy actors, and the fraction of actors that contribute to the institution when given the opportunity. The formulas defining each of these values are given in sections 3-5.

We consider a specific type of distribution of time preferences, namely a truncated normal distribution of mode μ, and standard deviation σ. We call μ the **patience of the population**. We fix $\sigma = 0.25$ and vary μ between 0 and 1 below, to represent the variation of our outputs with the patience of the population—μ varies on the x-axis of each graph below.

We fix $q = 0.5$, $r = 2$, thus normalizing the maximum payoff of one interaction for the actor, and take $p_1 = 0.25$. The difficulty of cooperation is then $\delta^b = c_1/(p_1 qr) = 4 \times c_1$.

Below, in each graph, we vary δ^b between 0 and 4—on the y-axis each time. This leaves plenty of room for improvement with the institution, since, following Proposing 5.3, the baseline equilibrium is impossible when $\delta^b \geq 1$ (since $\delta^b \geq \hat{\delta}^b(\theta), \forall \theta \in [0,1]$). By varying δ^b between 0 and 1, we vary c_1 between 0 and $1 = q \times r$.

We fix $b = 1$, and take $k = \delta^b/4$. By varying δ^b between 0 and 1, we vary k between 0 and $1 = b$.

We take $p_2 = 3 \times p_1 = 0.75$ and $c_2 = c_1/3 = \delta^b/12$. By default, second-order cooperation is three times as observable, and three times less costly, than first-order cooperation. Because we are interested in institutions for extending the scale of cooperation, we assume that, in both equilibria, even after accounting for the incentives produced by the institution, second-order cooperation remains less costly, and more observable than first-order cooperation—that is, we assume:

$$c_2 \leq \mathbf{c_1} = c_1 - (\beta + \gamma) \tag{6.1}$$

$$p_2 \geq \mathbf{p_1} = p_1 + \pi_1 \tag{6.2}$$

This guarantees $(c_2/p_2) \leq (\mathbf{c_1}/\mathbf{p_1})$, allowing us define $\hat{\delta}_2(\theta)$ using only the top line of condition (3.7).

Finally, we contrast results between an inefficient and efficient institution by taking $\rho = 1/3$ and $\rho = 3$ respectively, in each of the four cases of institution introduced below.

6.2 Algorithm for the baseline equilibrium

As we saw, in Proposition 5.2, the equilibrium value of θ is uniquely defined for any set of parameters. The algorithm for doing so is provided by condition (5.12): we calculate the payoff of a chooser that trusts given empty reputation in the steady state as a function of θ, using our general formula introduced above. This is a strictly decreasing function of θ. If this payoff is negative even for $\theta = 0$, then we deduce that $\theta^{*,b} = 0$. If this payoff is positive even for $\theta = 1$, then we deduce that $\theta^{*,b} = 1$. Otherwise, we find a unique $\theta^{*,b} \in (0,1)$ by using a bisection algorithm.

Once we have obtained the value of $\theta^{*,b}$ as a function of a set of parameter values, we can determine whether the baseline equilibrium exists for that set of parameter values (Proposition 5.3), and, if so, the value for the level of cooperation (Lemma 5.4) and the actor and chooser's payoffs (Lemmas 5.1 and 5.5).

6.3 Algorithm for the institution equilibrium

The institution equilibrium characterized in section 3 depends on the specific allocation of incentives performed by the institution, as well as the distribution of discount factors.

Using Mathematica, we consider four different allocation of incentives, as described just below. For instance, the graphs presented in the main document are determined by considering a punishing-monitoring institution, which equally allocates contributions to punishment of defectors and increasing the probability of observation. Such an institution is characterized by $\gamma = \frac{1}{2}(\rho f_2 c_2)\frac{q}{1-q}$, $\pi_1 = \frac{1}{2}(\rho f_2)\frac{q}{1-q}$ and $\beta = 0$—where f_2 is the fraction of actors that contribute to the institution when given the chance.

In each case, we calculate f_2 as a function of θ, and deduce the value of all relevant variables as a function of θ—for instance, in the case above, the payoff of a defector, $r - \gamma$, and the value of the likelihood of observation, $p_1 + \pi_1$.

We check that the payoff of a chooser that trusts given empty reputation in the state is a decreasing function of θ in our entire parameter region, and then apply the same algorithm as above to deduce the value of θ^* (see Proposition 4.1).

Once we have obtained the value of θ^* as a function of a set of parameter values, we can similarly determine whether the institution equilibrium exists for that set of parameter values (Proposition 4.2), and, if so, the value for the level of cooperation (Lemma 3.7) and the actor and chooser's payoffs (Lemmas 3.3 and 4.3).

6.4 Mathematica output

We illustrate our results in fives cases: one case is the baseline equilibrium (or the case of no institution), and the other four are the institution equilibrium, for four different types of institution. More precisely, we compute the institution equilibrium for a purely rewarding institution (where all multiplied contributions are affected to increasing the payoff of cooperators by β), for a purely punishing institution (invest solely in γ), for a purely monitoring institution (invest solely in π_1), and for a monitoring-punishing institution, which equally divides its resources between increasing the likelihood of observation and punishing defectors—this is the example considered in the main article, that this document supplements.

6.5 Level of cooperation

6.5.1 Baseline equilibrium

In Figure 2, we plot the level of cooperation obtained in the baseline equilibrium.

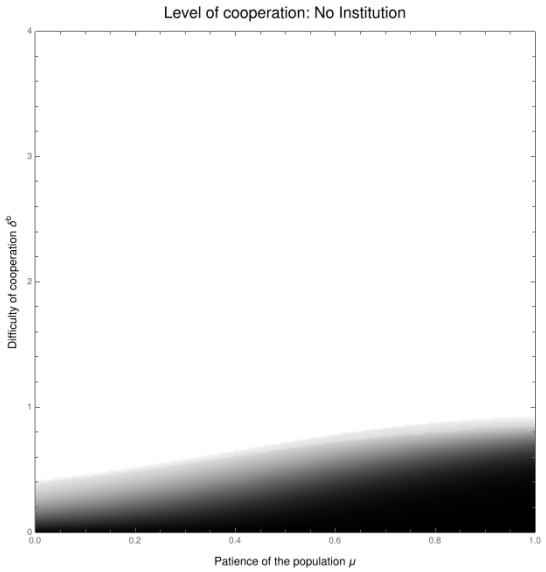

Figure 2: Level of cooperation in the baseline equilibrium, as a function of μ and δ^b.

In this graph—as in all graphs indicating the level of cooperation—the shade of gray indicates the level of cooperation at a given point: black indicates a level of cooperation of 1, and white indicates a level of cooperation of 0. We obtain this graph, as well as all other graphs presented in this supplementary document using using Mathematica's DensityPlot function.

We use the parameters defined above, and vary the patience of the population μ between 0 and 1 on the x-axis, and the difficulty of cooperation δ^b between 0 and 4 on the y-axis.

6.5.2 Institution equilibrium

In Figure 3, we plot the level of cooperation obtained in the institution equilibrium for the first three cases, i.e. the rewarding (top), punishing (middle), and monitoring (bottom) institutions. In each case, we consider the inefficient and efficient variant of the institution, by fixing $\rho = 1/3$ (left column) and $\rho = 3$ (right column). To generate these six graphs (two per case; one case equals one row), we fix all parameters as above, and again vary μ between 0 and 1, and δ^b between 0 and 4.

Finally, we plot the level of cooperation obtained for a monitoring-punishing institution in Figure 4, again for $\rho = 1/3$ and $\rho = 3$.

6.6 Comparison between the monitoring-punishing institution and no institution

6.6.1 Increase in the level of cooperation

In each case, the level of cooperation is higher in the institution equilibrium than it is in the baseline equilibrium—and the difference is starker when the institution is more efficient (high ρ). To illustrate the effect of an institution on cooperation, we subtract the level of cooperation in the baseline equilibrium to the level of cooperation in the institution equilibrium in the case of a monitoring-punishing institution. We plot the resulting increase in the level of cooperation in Figure 5, for $\rho = 1/3$ (left) and $\rho = 3$ (right).

This time, the shade of gray indicates the increase in the level of cooperation at a given point: black indicates an increase of 1 (hence that the level of cooperation must be 1 in the institution equilibrium and 0 in the baseline equilibrium), and white indicates an increase of 0.

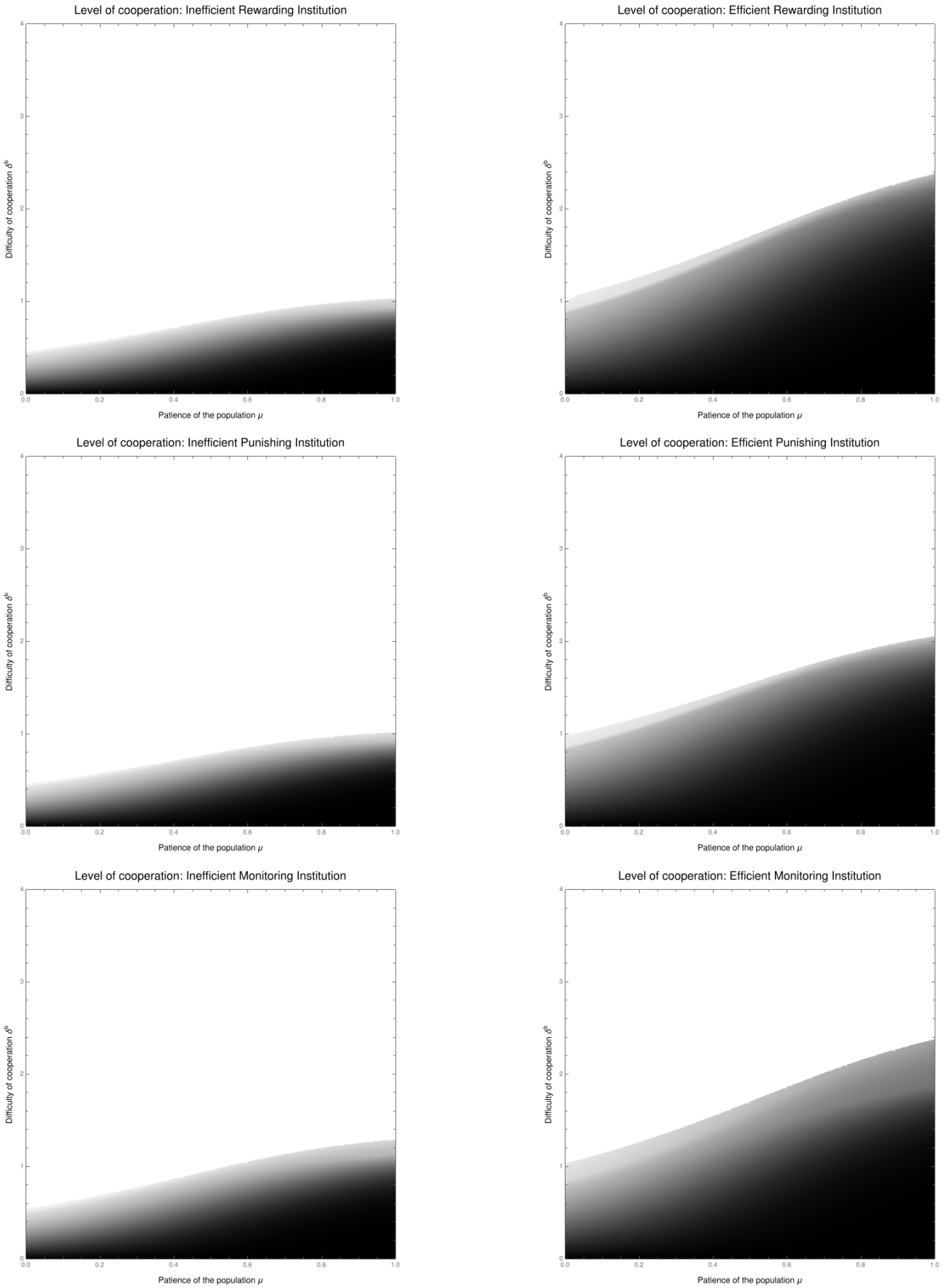

Figure 3: Level of cooperation in the institution equilibrium, for a purely rewarding institution (top row), a purely punishing institution (middle row), and a purely monitoring institution (bottom row). In each case, results are computed as a function of μ and δ^b, for $\rho = 1/3$ (inefficient institution, left column), and for $\rho = 3$ (efficient institution, right column).

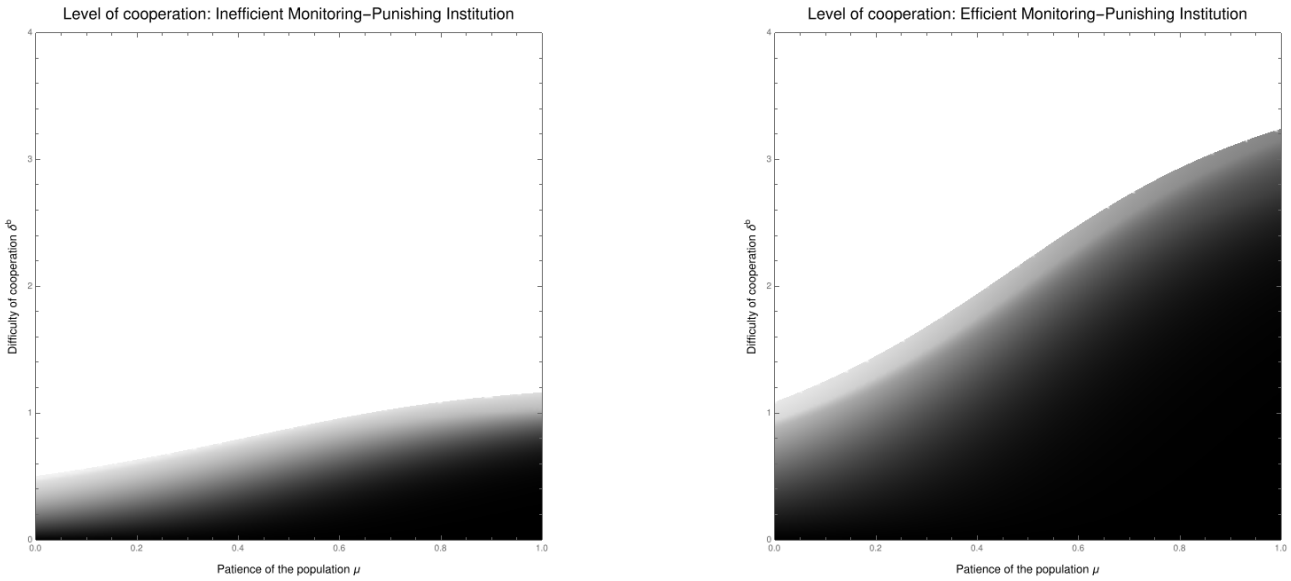

Figure 4: Level of cooperation in the institution equilibrium for a monitoring-punishing institution, as a function of μ and δ^b. Left: $\rho = 1/3$ (inefficient institution); right: $\rho = 3$ (efficient institution).

6.6.2 Change in chooser and actor payoffs (efficient institution)

Even an efficient institution (right of Figure 5) leads to only a marginal increase in the level of cooperation when δ^b is low and μ is high—in such a case, the level of cooperation is already high without an institution.

The institution may in fact lead to a decrease in actors' payoffs in a similar parameter space, as actors then pay the cost of second-order cooperation c_2 in a context where costly enforcement of first-order cooperation is largely unnecessary. To illustrate this, we subtract the actor expected payoff in the baseline equilibrium to the actor expected payoff in the institution equilibrium, for $\rho = 3$, in the case of the monitoring-punishing institution.

In contrast, the chooser's payoff is always higher in the institution equilibrium than in the baseline equilibrium even when δ^b is low and μ is high—because the costs of second-order cooperation are always payed by actors.

We compute the actor's payoff by taking the normalized lifetime payoff defined as a function of δ by condition 3.9 for the institution equilibrium, and condition 5.4 for the baseline equilibrium. For a given value of μ, we then compute the expected value of this payoff over the entire distribution, and normalize by dividing by $q \times r$. We subtract results from the baseline equilibrium to those obtained in the institution equilibrium for $\rho = 3$.

We proceed similarly for the chooser's long-run payoff, subtracting the value obtained in the baseline equilibrium (5.11) from the one obtained in the institution equilibrium (4.2), and normalizing, by dividing by b.

We plot our results in Figure 6, again as a function of μ and δ^b.

For both of these graphs, shades of blue indicate an increase in payoffs: dark blue indicates an increase of 1, and white an increase of 0. Shades of red indicate a decrease in actor expected payoff: dark red indicates an increase of 0.05 or more. Note that with our assumptions, the cost of cooperation $c_2 = c_1/3$ is very small when c_1 is small and μ high, i.e. in those points of the parameter space in which the institution appears unnecessary.

6.6.3 Change in total payoff (efficient institution)

Finally, we carry out the same computation for the expected payoff; that is, the expected payoff of a random individual, by averaging between the actor and chooser payoff used above. Note that we weigh the actor expected payoff by $1/(1+q)$ and the chooser long-run payoff by $q/(1+q)$ to capture the fact that the actor plays in both games and the chooser only in the trust game—weighing the actor and chooser equally would lead to a lower estimation of the red zone in which total payoffs decrease, since only actors pay the cost of second-order cooperation.

We plot the result in Figure 7. In regions in which the actor expected payoff decreases, this decrease is partially compensated by an increase in chooser long-run payoff. As a result, the maximum net decrease for the expected payoff is just above 0.006, or 0.6% of the maximum value. We have to even further shrink the red scale in order to see decreases—dark red now indicates a decrease of 0.06 or more. As before, dark blue indicates an increase of 1.

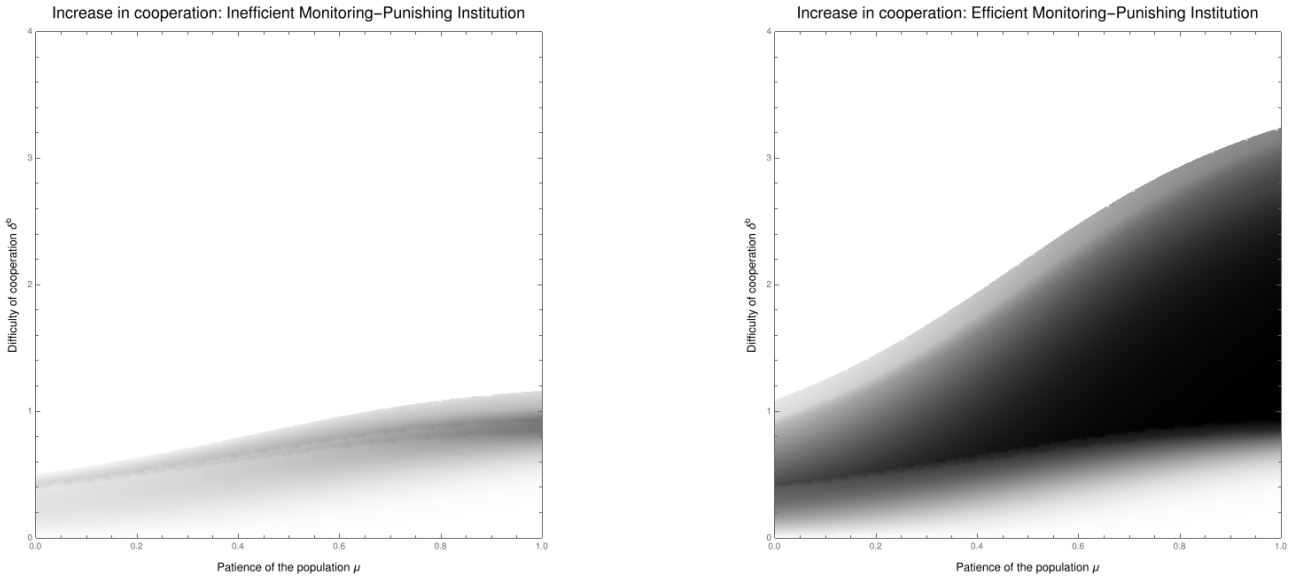

Figure 5: Increase in the level of cooperation due to the institution, as a function of μ and δ^b. To compute this value, we subtract results plotted in Figure 2 to those plotted in Figure 4 (monitoring-punishing institution), for $\rho = 1/3$ (left) and $\rho = 3$ (right).

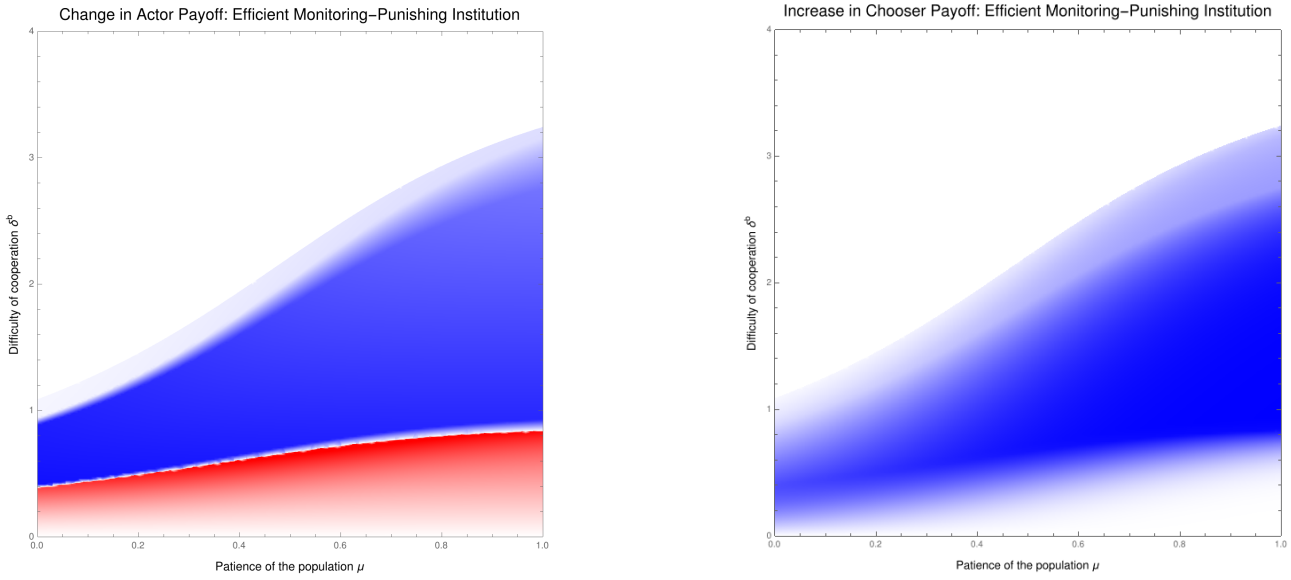

Figure 6: Change in expected actor payoff (left) and chooser payoff (right) as a function of μ and δ^b, in the case of the monitoring-punishing institution for $\rho = 3$. In both cases, shades of blue indicate an increase, on a scale of 0 to 1, i.e. 100% of the maximum value $qr = b = 1$. In the actor's case only, shades of red indicate a decrease, on a scale of 0 to 0.05, i.e. 5% of the maximum value qr.

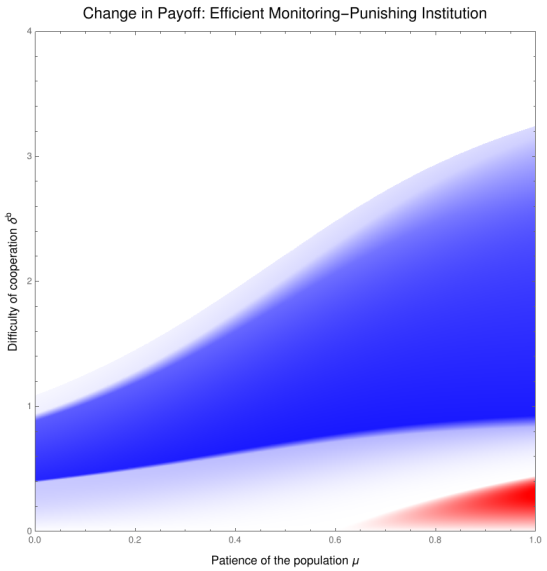

Figure 7: Change in expected payoff as a function of μ and δ^b, in the case of the monitoring-punishing institution for $\rho = 3$. Blue: increase, on a scale of 0 to 1, i.e. 100% of the maximum value. Red: decrease, on a scale of 0 to 0.006, i.e. 0.6% of the maximum value.

References

Selten, R. (1983). Evolutionary stability in extensive two-person games. Mathematical Social Sciences, 5(3), 269–363. https://doi.org/10.1016/0165-4896(83)90012-4