Department of Economics
University of Southampton
Southampton SO17 1BJ
UK

Discussion Papers in Economics and Econometrics

1999

COOPERATION AND NON-HALTING STRATEGIES

LUCA ANDERLINI
(University of Southampton)

HAMID SABOURIAN
(King's College, Cambridge)

October 1999

ABSTRACT. This note is a response to an unpublished paper by Evans and Thomas (1998) of which we have recently become aware.

Evans and Thomas (1998) take issue with a paper that we published some years back on 'Cooperation and Effective Computability' in repeated games (Anderlini and Sabourian 1995). In that paper we showed that it is only the cooperative equilibria of an infinitely repeated two-player common-interest game with no discounting that survive both the restriction that players' strategies must be computable, and appropriately computable trembles.

Evans and Thomas (1998) assert that our results are seemingly not robust to changes in the set of computable strategies at the disposal of each player. In particular, they claim that our equilibrium selection result does not extend to the case in which players are allowed to choose strategies that halt on certain histories but do not halt on others.

The purpose of this note is to show that the claim in Evans and Thomas (1998) is misleading. We present a modification of the set-up of our earlier paper in which the cooperative equilibria are selected when strategies that halt on certain histories and do not halt on others are allowed.

Although extensive modifications are required, the proof of this extension of our earlier result runs along the same general line of argument as the original proof.

JEL CLASSIFICATION: C72, C73, C79.

KEYWORDS: Computability, Repeated Games, Cooperation, Non-Halting Strategies.

ADDRESS FOR CORRESPONDENCE: Luca Anderlini, Department of Economics, Georgetown University, 37th and O Streets, North West, Washington DC-20057-1036, U.S.A.

E-MAIL: LA2@GUNET.GEORGETOWN.EDU

URL: WWW.ECON.CAM.AC.UK/FACULTY/LA13/LA13.HTM

1. Introduction

1.1. Anderlini and Sabourian (1995)

In Anderlini and Sabourian (1995) (henceforth A-S) we prove the following result. Consider an infinitely repeated, two-player, finite, strategic-form common-interest game with no discounting. Assume that the players are constrained to use supergame strategies that are computable by a Turing machine. Assume further that the only allowable Turing machines are those that either halt on all possible histories of play or those that never halt. This yields a normal form (machine) game in which each player's strategy set is the set of allowable Turing machines and the payoffs are given by the long-run average payoffs associated with each pair of computable strategies. Now consider the Trembling-Hand Perfect equilibria of this machine game where the trembles are assumed to be themselves appropriately computable. Then, provided that the trembles have sufficiently large support, the unique long-run payoff vector that survives is the Pareto-efficient one in which the players cooperate in the underlying common-interest stage game.

The intuition behind the cooperation result in A-S is relatively easy to explain. The assumption of computable strategies guarantees that the strategy sets of both players in the supergame are countable. The assumption of computable trembles as well as computable strategies guarantees that the following construction is well defined. The set of strategies that are in the support of the perturbation *and* that have the property that after cooperating once they *do not* cooperate for ever after can be *enumerated* in a computable way (call this set $\overline{\mathcal{Q}}$). It now follows that there exists a computable strategy, call it $\overline{x}$, that enumerates 'sufficiently many' strategies in $\overline{\mathcal{Q}}$, so that the probability 'tail' of strategies that have not been enumerated is 'sufficiently small'. Strategy $\overline{x}$ now uses the first, say, $\overline{t}$ periods of play to ensure that its behaviour is *different* from all the strategies in $\overline{\mathcal{Q}}$ that have been enumerated,[1] and after $\overline{t}$ always cooperates (regardless of the previous history of play). Crucially, to be able to differentiate itself from the enumerated strategies in $\overline{\mathcal{Q}}$, strategy $\overline{x}$ must *simulate* what their actions would have been, given a certain history of play.

[1] A common way to describe this way of proceeding in the mathematical literature is to say that $\overline{x}$ 'diagonalizes' the machines that have been enumerated.

We can then appeal to a corollary of the so-called Recursion Theorem[2] to close the construction as follows. The probability of the tail mentioned above can be made arbitrarily small relative to the probability of $\overline{x}$ *itself.* Once this is done, it is clear that the updated probability of facing a cooperative strategy conditional on the history of play in the first $\bar{t}$ periods can be made arbitrarily close to one. Given that the stage game is one of common interest, any maximizing opponent will then have to reciprocate by cooperating after date $\bar{t}$. In A-S we call a strategy like $\overline{x}$ a 'smart machine'. We shall do the same here.

A smart machine as above can be constructed for any given computable perturbation. Since we restrict attention to computable perturbations, we can now conclude that the set of all possible smart machines as above is itself countable. Let this set of machines be denoted by be $\mathcal{R}$. Theorem 1 in A-S asserts that all equilibria of the repeated game that are robust to perturbations that include $\mathcal{R}$ in their support must be cooperative.

1.2. Non-Halting Machines

In A-S, non-halting machines are allowed in the support of the perturbations of players' strategies. Non-halting strategies seem to be a necessary feature for results like the one in A-S.[3] The intuitive reason is that the simulation step that we mentioned in Subsection 1.1 is not guaranteed to halt. This is a version of a well-known fact called the *halting problem.*[4]

In A-S, to keep matters simple we specify the model in a way that allows us to consider only two types of machines: the ones that halt and produce a 'legal' action in the stage game on *all* possible histories of play, and machines that do not halt on *any* history of play. In that paper we also discuss the presence of non-halting machines, and we defend the way in which the players' payoffs are extended to deal with these machines. The same arguments apply to the present context. In effect, a non-halting machine is treated like a machine that makes an 'illegal' move in the

[2] See A-S, Theorem A.7, or Cutland (1980), Theorem 11.1.1 and Corollary 11.1.4. The Recursion Theorem is a (pseudo) fixed-point theorem in the space of Turing machines.

[3] See also Anderlini (1999) for an application of the same techniques to one-shot common-interest game with pre-play communication. Non-halting strategies feature in that paper as well.

[4] See, for instance, Cutland (1980).

game. A chess player who overturns the board instead of making a legal move loses the game. Assumption 3 below stipulates that the payoffs to a non-halting machine are strictly dominated by the payoffs to some allowable halting machine. Moreover, according to Assumption 3 below any machine that plays cooperatively is a best response to any non-halting machine.[5] This is consistent with the idea that any legal move in chess is a best response to a player who overturns the board. Therefore, if we accept the idea that non-halting machines are present in the game and that the payoff functions must be extended to include these machines, then the payoffs stipulated in Assumption 3 below are easy to justify on a primitive, intuitive level.

The fact that there are actual payoffs associated with the non-halting machines in this paper (and in A-S) also needs justification. How does a machine that does not halt come to earn its payoff? Since its computation does not halt, and no referee can ascertain this in advance, when is the payoff awarded? Again, an interpretation that is appealing on a primitive, intuitive, level is not difficult to outline.

The results in this paper (and in A-S) should be interpreted as the result of the following limit operation. Begin with a model in which the computations of all allowable Turing machines are only made to run for a maximum of *s steps.* After s steps the computations are *truncated.* If a given machine's computation has halted, and has yielded a legal action in the stage game in a number of steps less than or equal to s, then that machine's action is taken to be the result of its computation. On the other hand, if the given machine does not halt within s steps then its computation is truncated, and its output is treated as undefined. In this case the machine earns the bad payoffs exactly as described in Assumption 3 below. Moreover, any completed computation yielding a legal action in the stage game is a best response to any strategy profile that contains a truncated computation, again exactly as in Assumption 3 below. For each positive integer s, let $G^\infty(s)$ denote the infinitely repeated machine game in which all computations are truncated at s steps as we have just described.

The results that we present in this paper (and those in A-S) can now be interpreted as applying to the limit game obtained from $G^\infty(s)$, as s becomes unboundedly large.[6]

[5] See also footnote 14 below.

[6] Note that the same interpretation also applies to the results in Anderlini and Sabourian (1998) and Anderlini (1999).

In this sense, we are dealing here with a *limit case of bounded rationality.* As s becomes larger and larger the computational resources of each machine become larger and larger. Either the time allowed for each computation expands without bound, or the steps are executed at an ever increasing speed. In the limit, all that matters is that the computation halts and yields a legal action in the stage game G. If it does not, then the extended payoffs described in Assumption 3 below apply.

Our assumptions of computability, together with the presence of non-halting machines and associated payoffs, can therefore be interpreted as restricting attention to the following world. Strategies and perturbations must be capable of being computed by some imaginable finite device in a *finite* number of steps. However, in this world, the number of steps is not limited in any way by time or other resource constraints.

1.3. Evans and Thomas (1998)

Evans and Thomas (1998) (henceforth E-T) examine a two-player infinitely-repeated common-interest game with no discounting with perturbations that give positive probability to a countable set of strategies.

They show that, under certain conditions, a *necessary* condition to select the cooperative equilibria of the repeated game is that the perturbations give positive probability to a certain set of 'draconian' strategies. A strategy is draconian if, it is prepared to 'minmax' the opposing player 'almost all' the time.

The smart machines that drive the main result in A-S are obviously not draconian. In fact they always cooperate after a certain date. In their paper, E-T ask what drives the different results in the two set-ups. They conclude that the difference is due to the way the non-halting machines are treated in A-S. In particular, they claim that the main result in A-S requires that "*the only machines allowable in the supports of the perturbations are those that always halt and those that never halt*" (Evans and Thomas 1998, page 11). Thus, according to E-T, the main result in A-S is non-robust in quite a strong sense.

The claim in E-T that we have quoted above is simply false. In this note we show that the main theorem in A-S can be extended to accommodate strategies that halt up to a given time t and do not halt after that. This extension of the main result in A-S requires a number of modifications of the original proof. However, the main line of reasoning is close to the argument presented in A-S. In Section 3 below we highlight the main differences between the proof of the main result in this note and in A-S.

It should be noted that we are not claiming that the results in A-S extend to *all* possible subsets of Turing machines being allowed in the perturbations of the machine game. For instance, we know that our results do not extend to the two possible extreme cases: the case in which no non-halting machines are allowed, and the case in which *all* Turing machines are allowed. The presence of non-halting machines is necessary because of the halting problem as we have mentioned above. The set of all Turing machines appears to be 'too rich' to be compatible with the type of assumptions that are needed for our results.[7]

It should be emphasized at this point that E-T do *not* take issue with the presence of non-halting machines per se,[8] but with what they claim is a crucial assumption about the types of machines that are allowed. In this note, we show that the results in A-S extend to the case in which a richer set of machines are allowed in the perturbations. We believe that the sets of allowable machines used here are quite appealing in their own right, and thus strengthen the results in A-S. The Turing machines that we allow here may halt on some histories and not on others. It seems a natural restriction to impose that if a machine does not halt at a certain date, then it will not halt thereafter. Once a machine begins a computation that goes on forever, it remains stuck in that state at all subsequent dates. The restriction that we have just described informally is precisely the one that we adopt in this note. In our view, the further techniques developed here demonstrate that the results in A-S are quite robust to changes in the model. The sets of allowable machines could be enriched in various other ways without affecting the basic viability of the main result.

E-T seem to be aware of the fact that extensions of the A-S results similar to the one are presented here may be possible. They proceed to dismiss their usefulness on the following grounds. If some sets of allowable machines can be found that make

[7] See also the discussion in Section 8 of A-S on 'Admissibility and Large Support'. Those remarks apply here virtually unchanged.

[8] In an earlier version of their paper (see Evans and Thomas (1997, page 2)) the authors state that "[...] *one can defend the assumption that non-halting machines may be present in the perturbations,* [...]".

viable some A-S type selection results, they must be 'too strong' as 'primitive' assumptions about allowable beliefs. The latter claim, they argue, follows from the fact that such restrictions on allowable strategies must make it impossible to consider the perturbations (containing draconian strategies) that they use to prove their results.

This, in our view, is simply the wrong way to judge whether an assumption is a good primitive restriction on beliefs. Surely, we should judge a primitive restriction on beliefs (a set of allowable machines in this case) on its own merits, and *not* on the type of theorems that it implies. As we stated above, we believe that the sets of allowable machines that we use in this note are quite natural and appealing in their own right, and hence acceptable as a primitive restriction on beliefs.

The results in A-S and in this note rely heavily on certain properties of computable functions.[9] In particular, the computability framework allows us to use the fact that Turing machines can 'simulate' the computations of other Turing machines,[10] a parameterization result known as the 's-m-n' theorem,[11], and, crucially, the Recursion Theorem[12] that guarantees the feasibility of the fixed-point argument that we have mentioned above. These properties of the computability framework are simply not available in a model that allows any countable set of strategies in the perturbations of a repeated game.

The strength of the results in A-S, in our view, is therefore measured by how appealing we think the computability restrictions on strategies and perturbations really is. In A-S we have defended the computability framework at some length, and we will not repeat those arguments in this note. Here we simply recall that the notion of computability that we use is widely accepted in the mathematical literature as embodying the widest possible notion of 'effective computability'. Any function that is effectively computable by any imaginable finite device in a finite number of steps is in fact computable by a Turing machine.[13] We do not know whether the

[9] See also Anderlini and Sabourian (1998) and Anderlini (1999) for two further related papers. The former contains an extension to N-player games of the main result in A-S

[10] We are referring to the existence of a so-called 'Universal Turing Machine'. See, for instance, Cutland (1980), Ch. 5.

[11] See, for instance Cutland (1980), Theorem 4.4.3.

[12] See footnote 2 above.

[13] This is a claim known as Church's thesis in the literature on recursive functions. See Cutland (1980) again for further details.

techniques developed in A-S can be extended to a framework that does not appeal explicitly to the notion of effective computability. It is of course possible that some other sets of restrictions on strategies and perturbations will yield equivalent results.

We conclude this section with the observation that the techniques developed in A-S and in this note are sufficiently powerful to yield selection results in repeated common-interest games with discounting and with finite horizon (Anderlini and Sabourian 1990) and to N-player repeated common-interest games (Anderlini and Sabourian 1998). By contrast, the use of draconian strategies as a device to select the cooperative equilibria of a repeated common-interest game developed in E-T does not appear to be robust to any of these changes in the model.

2. An Extension of Theorem 1 of Anderlini and Sabourian (1995)

2.1. Notation and Basics

We conform to the notation used in A-S whenever possible. We also refer extensively to that paper for some of the results that carry over from there in order to save space.

Using a standard technique known as Gödel numbering, Turing machines and their inputs and outputs can be put in a one-to-one (computable) correspondence with the natural numbers. Throughout the paper $\mathbb{N}$ denotes the set of natural numbers. The result of the computation of the Turing machine with Gödel number $x \in \mathbb{N}$ when applied to the input string coded by the Gödel number $y \in \mathbb{N}$ is denoted by $\varphi_x(y)$. The notation $\varphi_x(y) \uparrow$ and $\varphi_x(y) \downarrow$ respectively indicate that the computation $\varphi_x(y)$ does not halt (it 'loops'), and that it does halt.

Definition 1: *A function f from $\mathbb{N}^m$ to $\mathbb{N}$ is said to be computable if and only if*

$$\exists x \in \mathbb{N} \quad \textit{such that} \quad f(y_1, \cdots, y_m) \simeq \varphi_x(y_1, \cdots, y_m) \;\; \forall\, (y_1, \cdots, y_m) \; \in \mathbb{N}^m$$

The symbol '$\simeq$' used between two Turing machines, two computable functions or any combination of these means 'defined on the same set of inputs and equal whenever defined'.

The stage game of the repeated game we consider is denoted by $\widehat{G} = \{\widehat{\mathcal{A}}_i, \hat{\pi}_i\}_{i=1,2}$. We take $\widehat{G}$ to be a finite-action, two-player, strategic-form game. A generic player will be denoted by $i = 1, 2$, and unless otherwise stated j will denote i's opposing player. Player i's finite action set is denoted by $\widehat{\mathcal{A}}_i$, and $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_1 \times \widehat{\mathcal{A}}_2$ is the players' joint action set. Typical elements of $\widehat{\mathcal{A}}_i$ and $\widehat{\mathcal{A}}$ are denoted by $\hat{a}_i$ and $\hat{a}$ respectively. Following standard notation, $\hat{\pi}_i : \widehat{\mathcal{A}} \to \mathrm{I\!R}$ denotes player i's payoff function, while $\hat{\pi} : \widehat{\mathcal{A}} \to \mathrm{I\!R}^2$ yields a payoff vector given an action pair $\hat{a} \in \widehat{\mathcal{A}}$. Let $\widehat{V}$, with typical element $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2)$, be the payoff space of $\widehat{G}$. In other words, $\widehat{V} = \hat{\pi}(\widehat{\mathcal{A}})$.

Assumption 1: *The stage game $\widehat{G}$ is a common-interest game. In other words there exists a $\hat{\pi}^e \in \widehat{V}$ (which may be associated with more than one pair of strategies) that strongly Pareto-dominates all other elements of $\widehat{V}$. The action profile $a^e \in \widehat{\mathcal{A}}$ denotes one (arbitrarily fixed) pair of actions which yields payoff pair π^e to the players.*

For the sake of simplicity only we will focus attention on common-interest games in which each player has at least three pure strategies available. In A-S we discuss at some length why this property is not needed for our results. The same remarks apply here unchanged.

Assumption 2: *The cardinality of both $\widehat{\mathcal{A}}_1$ and $\widehat{\mathcal{A}}_2$ is at least three.*

In A-S we extend the payoffs of the infinitely repeated game to take into account the possibility of non-halting machines making two assumptions directly on the long-run payoffs of the two players. In the context of A-S this seems by far the simplest way to proceed.

In this note we consider machines that never halt, machines that always halt and machines that halt on histories of up to a given length, but do not halt thereafter. In this context, it seems more appropriate to extend the players' payoff functions at the stage game level. Our way of proceeding here is completely consistent with the way the corresponding assumption is formulated in A-S.

Some extra notation is required. Let $\mathcal{A}_i = \widehat{\mathcal{A}}_i \cup \uparrow$, so that $\mathcal{A}_i$ is payer i's strategy set in the stage game $\widehat{G}$ plus the non-halting action, denoted by $\uparrow$. Let also $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$. We can now define the extended payoff function $\pi : \mathcal{A} \to \mathrm{I\!R}^2$. This yields a pair of stage game payoffs for every pair of actions $a \in \mathcal{A}$. Our next step is to assume that π is such that not halting is a strictly dominated strategy for either player, and that playing the cooperative action a_i^e is a best response to an opposing player that does not halt.[14] For every action pair in the original set $\widehat{\mathcal{A}}$, the extended payoff function π, of course, is the same as the original payoff function $\hat{\pi}$.

Assumption 3: *The extended payoff function π satisfies the following properties. First of all for every $a \in \widehat{\mathcal{A}}$, $\pi(a) = \hat{\pi}(a)$. Moreover, for $i = 1, 2$ we have that*

$$\exists\, \hat{a}_i \in \widehat{\mathcal{A}}_i \text{ suchthat } \pi_i(\hat{a}_i, a_j) \, > \, \pi_i(\uparrow, a_j) \,\forall\, a_j \in \mathcal{A}_j \qquad \text{(Dominance)}$$

$$\pi_i(a_i^e, \uparrow) \, \geq \, \pi_i(a_i, \uparrow) \,\forall\, a_i \in \mathcal{A}_i \qquad \text{(Best Response)}$$

The stage game with joint action set $\mathcal{A}$ and payoff functions π is denoted by G. The payoff space of G is denoted by V.

The undiscounted infinitely repeated game obtained from G is denoted by G^∞. Let $a_{it} \in \mathcal{A}_i$ be player i's action at time $t = 0, 1, 2 \cdots$, and $a_t \in \mathcal{A}$ the players' joint action at t. Let $\mathcal{H}_t = \mathcal{A}^t$ be the set of all possible *finite* histories of play of length t, with typical element $h_t = (a_0, \cdots, a_{t-1})$ (define h_0 to be the empty set). The set of all possible finite histories of play, regardless of length, is denoted by $\mathcal{H} \equiv \bigcup_{t=0}^{\infty} \mathcal{H}_t$. A strategy for player i in G^∞ is a map $\sigma_i : \mathcal{H} \to \mathcal{A}_i$. The joint action that players take at time t along the outcome path induced by σ will be indicated by $\mathrm{a}_t(\sigma) = (\mathrm{a}_{1t}(\sigma), \mathrm{a}_{2t}(\sigma))$. The history of length t generated by a pair of supergame strategies σ is denoted by $\mathrm{h}_t(\sigma) = [\mathrm{a}_0(\sigma), \cdots, \mathrm{a}_{t-1}(\sigma)]$. The long-run undiscounted payoff to player i is $\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \pi_i[\mathrm{a}_t(\sigma)]$.

Since G is a *finite-action* game and we do not consider mixed strategies within the stage game, we can use the standard techniques above to code (in a computable way) any element of $\mathcal{H}$ into an element of $\mathbb{N}$. The elements of $\mathcal{A}_i$ can also obviously be assigned codes in $\mathbb{N}$. It follows that a strategy in G^∞ can always be viewed as a function from $\mathbb{N}$ to $\mathbb{N}$. Since this does not cause any ambiguity, now and throughout the rest of the paper, we shall use the same symbol for $h_t \in \mathcal{H}$ and a_i, and for their 'codes' in $\mathbb{N}$.

[14] Evidently, this includes as a special case the possibility that *any* halting choice is a best response to a non-halting opponent.

2.2. The Set of Allowable Turing Machines

As we mentioned above, we are forced to consider some Turing machines that do not always halt and produce a legal action in G^{∞}. The crucial point here is that we are able to consider a set of machines that is richer than that allowed in A-S. In particular, we are able to include machines that halt on all histories up to a certain length and do not halt thereafter. This motivates out next definition.

Let $\mathcal{S}_i^{\emptyset} = \{x_i \in \mathbb{N}$ such that $\varphi_{x_i}(h_t) \uparrow$ for all $h_t \in \mathcal{H}\}$. For every $t = 0, 1, 2, \ldots$ define also $\mathcal{S}_i^t = \{x_i \in \mathbb{N}$ such that $\varphi_{x_i}(h_\tau) \downarrow \in \mathcal{A}_i$ for every h_τ with $\tau \leq t$, and $\varphi_{x_i}(h_\tau) \uparrow$ for every h_τ with $\tau > t\}$. Lastly, let $\mathcal{S}_i^{\infty} = \{x_i \in \mathbb{N}$ such that $\varphi_{x_i}(h_t) \downarrow \in \mathcal{A}_i$ for every $h_t \in \mathcal{H}\}$. In other words, $\mathcal{S}^{\emptyset}$ is the set of machines that do not halt on any history, while $\mathcal{S}^t$ is the set of machines that halt on all histories of length less than or equal to t and do not halt on any history of length greater than t , and finally $\mathcal{S}^{\infty}$ is the set of machines that halt on all histories of play, regardless of length.

Definition 2: *The set of allowable Turing machines for player i in G^{∞} is the set of all machines that halt on all histories up to a given length and do not halt thereafter, together with machines that never halt and with machines that always halt. In other words, the set of allowable machines for player i is defined as*

$$\mathcal{S}_i \equiv \left\{ \bigcup_{t=0}^{\infty} \mathcal{S}_i^t \cup \mathcal{S}_i^{\emptyset} \cup \mathcal{S}_i^{\infty} \right\} \tag{1}$$

Notice that in A-S machines of the type contained in the sets $\mathcal{S}_i^t$ are not allowed. In this sense the sets of allowable machines are richer in this note that in A-S. The purpose of this note is to show that the results in A-S extend to this case.

Given a pair of Turing machines (x_1, x_2) in $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$, we can clearly define an outcome path in G^{∞} in the standard recursive way as above. The joint action at time t is denoted by $\mathrm{a}_t(x_1, x_2) \in \mathcal{A}$. The history of length t generated by (x_1, x_2) is denoted by $\mathrm{h}_t(x_1, x_2) \equiv \{\mathrm{a}_0(x_1, x_2), \cdots, \mathrm{a}_{t-1}(x_1, x_2)\} \in \mathcal{A}^t$. Given a pair $(x_1, x_2) \in \mathcal{S}$, the long-run payoff to player i is denoted by $\Pi_i(x_1, x_2) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \pi_i[\mathrm{a}_t(x_1, x_2)]$. The long-run continuation payoff to player i conditional on history h_t is denoted by $\Pi_i(x_1, x_2 | h_t)$.

2.3. Admissible Trembles

As in A-S we need the perturbations of the machine game that we have defined in Subsection 2.1 to satisfy two main computability restrictions. It is convenient to state the first one in terms of an abstract probability distribution over $\mathbb{N}$ before using it to define the admissible perturbations of the machine game. The following is identical to Definition 6 in A-S.

Definition 3: *A Probability distribution* $P = \{P(1), P(2), \cdots, P(x), \cdots\}$ *over* $\mathbb{N}$ *is said to be 'computable' if and only if there exists a Turing machine that computes (at least) all non-zero values of* P *as a function of* x*. Formally, let* Δ^∞ *represent the unit simplex in* $\mathbb{R}^\infty$ *and* $\text{supp}(P) = \{x \in \mathbb{N} \mid P(x) > 0\}$*, then* $P \in \Delta^\infty$ *is said to be computable if and only if* $\exists\, p \in \mathbb{N}$ *such that* $x \in \text{supp}(P)$ *implies*

$$\varphi_p(x) = P(x)$$

and $\varphi_p(x) \downarrow \Rightarrow \varphi_p(x) = P(x)$*.*

The second restriction on computable trembles concerns the ability to enumerate and compute the probability of the tail of the set $\overline{\mathcal{Q}}$ of non-cooperative machines as we described in the introduction. Since the set of allowable machines in this note is different from what we considered in A-S, the definition of the set $\overline{\mathcal{Q}}$ needs to be modified. The difference between what we do here and what is carried out in A-S is driven by the fact that we need to avoid the possibility that the smart machine described in Subsection 1.1 may have to simulate some machines that do not halt in order to distinguish itself from the set $\overline{\mathcal{Q}}$. Of course, at the same time we need to ensure that convincing the opposing player that the strategy employed is not in $\overline{\mathcal{Q}}$ is still sufficient to trigger cooperation in the long-run.

We start by defining the set of Turing machines that after cooperating once are guaranteed either to play the cooperative action or to not halt. Given any history of play h_t, if another history of play $h_{t'}$ is a continuation of history h_t, we write $h_{t'} \succ h_t$. Any history $h_{t'}$ that is either equal to h_t or is a continuation of h_t will be indicated as $h_{t'} \succeq h_t$.

Let

$$\mathcal{Z}_i = \{x_i \in \mathcal{S}_i \mid \varphi_{x_i}(h_t) = a_i^e \Rightarrow \text{ either } \varphi_{x_i}(h_{t'}) = a_i^e \text{ or } \varphi_{x_i}(h_{t'}) \uparrow \ \forall\, h_{t'} \succ h_t\} \tag{2}$$

and let $\overline{\mathcal{Z}}_i$ denote the complement of $\mathcal{Z}_i$ in $\mathcal{S}_i$. Next, we define the set of machines that do not halt for some history of length less than or equal to the Gödel number of the machine itself. Let

$$\mathcal{X}_i = \{x_i \in \mathcal{S}_i \mid \exists\, t \leq x_i \ \exists\, h_t \in \mathcal{H}_t \text{ such that } \varphi_{x_i}(h_t) \uparrow\} \tag{3}$$

and let $\overline{\mathcal{X}}_i$ denote the complement of $\mathcal{X}_i$ in $\mathcal{S}_i$. The following is the equivalent of Definition 7 of A-S in the present context.

DEFINITION 4: *Let $\mathcal{Q}_i = \mathcal{Z}_i \cup \mathcal{X}_i$. Any machine in $\mathcal{Q}_i$ is called* quasi-cooperative. *Let $\overline{\mathcal{Q}}_i$ denote the complement of Q_i in $\mathcal{S}_i$. Thus $\overline{\mathcal{Q}}_i$ is the set of allowable machines that are* not *quasi-cooperative.*

Notice that, because of Assumption 3 signaling that one's strategy is quasi-cooperative is an effective way to trigger cooperation in the machine game. To see why we have included the machines in $\mathcal{X}_i$ in the set of quasi-cooperative machine observe the following fact. Since $\overline{\mathcal{Q}}_i$ is the complement of $\mathcal{Q}_i$, we clearly have that $\overline{\mathcal{Q}}_i = \overline{\mathcal{Z}}_i \cap \overline{\mathcal{X}}_i$. Therefore, any machine x_i in $\overline{\mathcal{Q}}_i$ has the property that it halts on all histories of play of length up to and including $t = x_i$. This fact enables us to avoid the need for a smart machine that attempts simulate a strategy that does not halt.[15]

We are interested in trembles that guarantee that the probability of $\overline{\mathcal{Q}}_i$ is computable. Since both $\mathcal{Q}_i$ and $\overline{\mathcal{Q}}_i$ are infinite sets, this is not a property which follows automatically from computability of the probability distribution in the sense of Definition 3. Our next Definition is identical (save for the fact that $\overline{\mathcal{Q}}_i$ is a different set) to Definition 8 of A-S.

DEFINITION 5: *A Probability distribution $P = \{P(1), P(2), \cdots, P(x), \cdots\}$ over the natural numbers is said to be $\overline{\mathcal{Q}}_i$-computable if and only if the probability that P*

[15] See the proof of Lemma B.4 and Section 3 below in which we highlight the main differences between the proof of the main result in A-S and the argument presented here.

assigns to $\overline{\mathcal{Q}}_i$ is a 'computable real number' in the sense that it can be approximated by a Turing machine up to any arbitrarily given degree of precision. Formally, let $P(\overline{\mathcal{Q}}_i) = \sum_{x \in \overline{\mathcal{Q}}_i} P(x)$ then $P \in \Delta^\infty$ is said to be $\overline{\mathcal{Q}}_i$-computable if and only if $\exists q \in \mathbb{N}$ such that

$$\mid \varphi_q(c) - P(\overline{\mathcal{Q}}_i) \mid < \frac{1}{c} \quad \forall\, c \in \mathbb{N}$$

Our next step is to define the probability distributions that will constitute the trembles of our machine game. This is the equivalent of Definition 9 in A-S.

Definition 6: *A Probability distribution $P = \{P(1), P(2), \cdots, P(x), \cdots\}$ over $\mathbb{N}$ is said to be admissible for player i if and only if a) it gives positive probability only to machines in $\mathcal{S}_i$, b) it is computable according to Definition 3, and c) it is $\overline{\mathcal{Q}}_i$-computable according to Definition 5. We denote by $\mathcal{P}_i$ the set of probability distributions that are admissible for player i.*

Lemma B.2 shows that, given any admissible probability distribution, the set $\overline{\mathcal{Q}}_i^P = \text{supp}(P) \cap \overline{\mathcal{Q}}_i$ is *recursively enumerable.*[16] Therefore, exactly as in A-S, there are three Turing machines associated with each $P_i \in \mathcal{P}_i$. One which computes the probabilities of individual machines, one which computes the probability of $\overline{\mathcal{Q}}_i$, and a third one which 'enumerates' the elements of $\overline{\mathcal{Q}}_i^P$. We will refer to such a triple of Turing machines as a 'basis' for P.

Definition 7: *A triple $(p, q, m) \in \mathbb{N}^3$ is said to be a 'basis' for an admissible $P \in \mathcal{P}_i$ if and only if $\varphi_p(\cdot)$ computes the values of P as in Definition 3, $\varphi_q(\cdot)$ computes (approximately) the value of $P(\overline{\mathcal{Q}}_i)$ as in Definition 5, and $\varphi_m(\cdot)$ 'enumerates' $\overline{\mathcal{Q}}_i^P$ 'without repetitions' (as in Theorems A.3 and A.4)* .[17]

2.4. Equilibrium

The concept of equilibrium that we use here is identical to the one used in A-S. There, we discuss at length its justification and possible interpretations. Here, for the sake of brevity, we only reproduce the formal definitions without comment.

[16] See Definition A.2.

[17] If $\overline{\mathcal{Q}}_i^P$ is empty, we require m to compute the function 'nowhere defined'.

Exactly as in A-S, we parameterize the possible perturbations by a 'lower bound' on their support. Given any $\mathcal{R}_i \subseteq \mathbb{N}$, let

$$\mathcal{P}_i(\mathcal{R}_i) \equiv \{\ P \in \Delta^\infty \quad | \quad \mathcal{R}_i \subseteq \mathrm{supp}(P) \quad \text{and} \quad P \quad \text{is admissible for} \quad i\ \}$$

Next, we establish some notation for the machines in $\mathcal{S}_i$ that are a best response to a given strategy and perturbation. Given any x_j in $\mathcal{S}_j$, P, and $\epsilon \in (0, 1)$, let

$$\mathcal{B}_i(x_j, P, \epsilon) \equiv \arg\max_{x_i \in \mathcal{S}_i} \left\{ (1-\epsilon)\Pi_i(x_i, x_j) + \epsilon \sum_{x \in \mathcal{S}_j} P(x)\Pi_i(x_i, x) \right\}$$

An $(\epsilon, \mathcal{R})$ Computable Trembling Hand Equilibrium (with $(\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2))$ is a pair of computable strategies and a pair of perturbations, with support at least $\mathcal{R}_i$ respectively, such that the given strategies are a best response to each other given the perturbations. The two definitions that follow are the same as Definitions 11 and 12 of A-S.

Definition 8: *An $(\epsilon, \mathcal{R})$ Computable Trembling Hand Equilibrium (abbreviated $(\epsilon, \mathcal{R})$-CTHE) is a quadruple (x_1^E, x_2^E, P_1, P_2) with $x_i^E \in \mathcal{S}_i$ and $P_i \in \mathcal{P}_i(\mathcal{R}_i)$ such that $\forall i = 1, 2$*

$$x_i^E \in \mathcal{B}_i(x_j^E, P_j, \epsilon)$$

The set of equilibrium quadruples for a given pair $(\epsilon, \mathcal{R})$ is denoted by $E(\epsilon, \mathcal{R})$, the set of equilibrium long-run payoff pairs is denoted by $\Pi^E(\epsilon, \mathcal{R})$.

Definition 9: *A $\mathcal{R}$-CTHE is the limit of any sequence of $\Pi^E(\epsilon, \mathcal{R})$ as ϵ vanishes. The set of $\mathcal{R}$-CTHE is denoted by $\Pi^E(\mathcal{R})$.*

2.5. Selecting Cooperative Equilibria in the Repeated Game

We are now ready to state our main result. The statement of the theorem that follows is identical to the main result in A-S. In the limit, as the noise vanishes, all Computable Trembling-Hand Equilibria of the machine game are cooperative, provided that the perturbations have 'sufficiently large' support.

Of course, the main difference between Theorem 1 below and Theorem 1 in A-S is that here we have machines that halt on some histories but not on others in the sets of allowable machines $\mathcal{S}_i$.

THEOREM 1: *There exists a set* $\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2) \subset \mathbb{N}^2$ *such that*

$$E(\epsilon, \mathcal{R}) \neq \emptyset \quad \forall\, \epsilon \in [0,1]$$

and

$$\Pi^E(\mathcal{R}) = \pi^e$$

PROOF: See Appendix B.

We conclude this Section with two observations. The first is the same as Remark 1 in A-S.

REMARK 1: *Recall that, in our definition of equilibrium, the sets* $\mathcal{R}_i$ *are 'lower bounds' on the support of the perturbations. It follows that whenever* $\mathcal{R}_i \subseteq \mathcal{R}'_i$ $\forall i = 1, 2$, *we must have* $\Pi^E(\mathcal{R}') \subseteq \Pi^E(\mathcal{R})$. *Therefore Theorem 1 implies that all Computable Trembling Hand Equilibria with perturbations having supports larger than the sets* $\mathcal{R}_i$ *of Theorem 1 are cooperative.*

Our last remark concerns the absence draconian strategies from the sets $\mathcal{R}_i$ of Theorem 1.

REMARK 2: *As is apparent from the proof of Theorem 1 all the machines in the sets* $\mathcal{R}_i$ *are quasi-cooperative in the sense of Definition 4. Thus the sets* $\mathcal{R}_i$ *do not contain any machines that are draconian in the sense of E-T. This is in marked contrast with Theorem 1 in their paper. Their result asserts that the presence of draconian strategies in the perturbations is* necessary *to induce cooperation in an infinitely-repeated undiscounted two-player common-interest game.*

The stark difference between the two results stems from the computability restrictions on players' strategies and on the perturbations of the repeated game, and from the sets of allowable machines that are used here.

3. The Proof of Theorem 1: Comparison to A-S

The basic intuition behind Theorem 1 above and behind Theorem 1 in A-S is essentially the same. It revolves around the 'diagonalization' argument that we described in Subsection 1.1 above. Our aim here is just to highlight the two main differences in the way the formal argument is carried out here and in A-S.

The first difference lies in the way Lemma B.4 is stated and proved here, compared to its equivalent (Lemma 2) in A-S. In Lemma B.4 we show that it is in fact feasible for the smart machine $\overline{x}_i$ to distinguish itself from (most of) the machines in $\overline{\mathcal{Q}}_i$ by simulating the machines that are not in the tail of $\overline{\mathcal{Q}}_i$ and then setting its own action to be different from each of these machines as times goes from 0 to $\overline{t}$.

In A-S, this step can be carried out simply by enumerating an appropriately large subset of $\overline{\mathcal{Q}}_i$ (so that the tail that has not been enumerated has sufficiently small probability), then simulating the machines that have been enumerated, and finally setting the actions of $\overline{x}_i$ to be different from the actions that have been so computed. This can always be done in A-S in this simple way because the only machines from which $\overline{x}_i$ needs to distinguish itself are machines that always halt. In this note, we face a further difficulty in carrying out this step since some of the machines that have been enumerated may be machines that halt on all histories up to a certain length and do not halt after that. The solution to the difficulty lies in the fact that here we have included in the set $\mathcal{Q}_i$ all those machines x_i that are certain to stop halting on or before $t = x_i$. It follows that all the machines to be simulated in $\overline{\mathcal{Q}}_i$ are certain to halt on any history of length up to $t = x_i$. We can then rearrange the machines to be simulated in ascending order and simulate the first one on h_0, the second one on h_1 and so on up to $h_{\overline{t}}$. Since each machine is certain to halt up to time $t = x_i$, it is now clear that all these simulations are certain to halt as is required.

The second main difference between the proof of the main result here and in A-S is in the way that Lemma B.6 is proved here, compared with the equivalent Lemma 4 in A-S. Lemma B.6 states that (unless the equilibrium is cooperative already) the long-run payoff to the smart machine $\overline{x}_i$ against the equilibrium machine x_j^E must be π_j^e, in the limit as the degree of precision with which $\overline{x}_i$ reveals itself to be quasi-cooperative increases without bound.

Observe that in A-S once the smart machine has halted *once*, it has revealed itself to be a machine that halts on all histories of play regardless of length. This is precisely because in A-S only machines that either always halt or never halt are allowed. Therefore in A-S we are able to prove Lemma 4 using the following relatively simple argument. Once the signaling phase of the smart machine $\overline{x}_i$ has ended, a machine x_j that always cooperates will achieve an expected continuation payoff that is arbitrarily close to π^e_j. Therefore, since x^E_j must be optimal in expected terms given any history of play that takes place with positive probability, the expected continuation payoff to x^E_j must also be arbitrarily close to π^e_j. Since G is a common-interest game this now implies that the payoff to the smart machine $\overline{x}_i$ against x^E_j must be arbitrarily close to π^e_i as required.

In this note, because we are allowing machines that only halt on histories up to a certain length and do not halt after that, and since we need to modify the definition of the set $\mathcal{Q}_i$ accordingly, the argument above is not viable anymore. In particular, as before, once the signaling phase of the smart machine has ended at $\overline{t}$, the opposing player knows that he is facing a machine in $\overline{\mathcal{Q}}_i$ with probability arbitrarily close to one. However, the set $\mathcal{Q}_i$ contains machines that halt up to $\overline{t}$ but stop halting at dates after $\overline{t}$. Therefore, it is no longer true that a machine x_j that cooperates for ever will achieve an expected continuation payoff arbitrarily close to π^e_j.

We circumvent this difficulty using Assumption 3 (Best Response).[18] This tells us that cooperating is a best response to machines that do not halt. Recall that, after the signaling phase of the smart machine $\overline{x}_i$ has ended, the opposing player faces machines that either cooperate or do not halt with probability arbitrarily close to one. Therefore, a machine that does not cooperate 'almost all the time' after the signaling phase has ended cannot be one that maximizes its expected continuation payoff. Therefore, the equilibrium machine x^E_j must cooperate almost all the time after the signaling phase of $\overline{x}_i$ is over. Therefore, we can now conclude that the payoff of $\overline{x}_i$ against x^E_j is arbitrarily close to π^e_i as required.

[18] It is interesting to notice that the best response assumption plays no role in the optimality of equilibrium statement in Theorem 1 of A-S. There, this assumption is only used to prove that the set of equilibria is not empty. By contrast, here it is needed to show that all equilibria are cooperative.

Appendix A: Preliminaries

We start with some basic results from the literature on recursive functions. For reasons of space, we refer extensively to A-S whenever this allows us to omit a proof altogether. The proofs that are given here are self-contained wherever possible. All results that are stated without proof or reference to A-S can be found, for instance, in Cutland (1980) or Rogers (1967).

Definition A.1: *A computable function $f : \mathbb{N}^m \to \mathbb{N}$ is called a total computable function if and only if $f(e_1, \cdots, e_m) \downarrow \ \forall (e_1, \cdots, e_m) \in \mathbb{N}^m$.*

Theorem A.1 [s-m-n]: *For each $m \geq 0$ and $n \geq 1$ there exists a total computable function of $m+1$ variables f such that $\forall\ e \in \mathbb{N}$ and $\forall\ (h_1, \cdots, h_m, h_{m+1}, \cdots, h_{m+n}) \in \mathbb{N}^{m+n}$ we have*

$$\varphi_e(h_1, \cdots, h_{m+n}) \simeq \varphi_{f(e,h_1,\cdots,h_m)}(h_{m+1}, \cdots, h_{m+n})$$

Theorem A.2 [Universal Turing Machine]: *Given any $m \geq 1$, there exists a number u, such that*

$$\varphi_u(n, e_1, \cdots, e_m) \simeq \varphi_n(e_1, \cdots, e_m) \ \ \forall\ (n, e_1, \cdots, e_m) \in \mathbb{N}^{m+1}$$

Definition A.2: *A set $S \subseteq \mathbb{N}$ is recursively enumerable (abbreviated r.e.) if and only if it is equal to the domain of a computable function. Formally, $S \subseteq \mathbb{N}$ is r.e. if and only if for some $n \in \mathbb{N}$ we have $\varphi_n(e) \downarrow \Leftrightarrow \ e \in S$. (The empty set is r.e. since the function 'nowhere defined' is computable.)*

Theorem A.3: *An infinite set $S \subseteq \mathbb{N}$ is r.e. if and only if it is the range of a one-to-one total computable function of one variable. Formally, given an infinite set $S \subseteq \mathbb{N}$, S is r.e. if and only if there exists a Turing machine n computing a total computable function such that $v \neq v' \Rightarrow \varphi_n(v) \neq \varphi_n(v')$ and*

$$e \in S \quad \Leftrightarrow \quad \exists v \text{ such that } \varphi_n(v) = e$$

The Turing machine n is said to enumerate S 'without repetitions'.

Theorem A.4: *Any finite set is r.e. Any finite set can be enumerated 'without repetitions' in the following way. Let $x_0, \ldots, x_S$ be the elements of S. Then there exists a Turing machine n computing a total computable function such that $\varphi_n(s) = x_s$ for $s = 0, \ldots, S$, and $\varphi_n(s) = x_S$ for all $s > S$.*

Theorem A.5: *The intersection of two r.e. sets is r.e. The union of two r.e. sets is r.e.*

Theorem A.6 [Pseudo-Fixed Point]: *For any computable function f of $m+1$ variables, there exists $\overline{x} \in \mathbb{N}$ such that*

$$\varphi_{\overline{x}}(e_1, \cdots, e_m) \simeq f(\overline{x}, e_1, \cdots, e_m) \ \ \forall\ (e_1, \cdots, e_m) \in \mathbb{N}^m$$

Appendix B: Proof of Theorem 1

Lemma B.1: *The set* supp(P) *is r.e. for any $P \in \Delta^\infty$ which is computable in the sense of Definition 3. It follows that the same statement is true for any $P \in \Delta^\infty$ which is admissible in the sense of Definition 6.*

Proof: See Lemma A.1 in A-S.

Lemma B.2: *Let P be a probability distribution that is admissible in the sense of Definition 6. Then the set $\overline{\mathcal{Q}}_i^P \equiv \mathrm{supp}(P) \cap \overline{\mathcal{Q}}_i$ is r.e.*

Proof: Consider the sets

$$\widehat{\mathcal{Z}}_i \;=\; \{x_i \in \mathbb{N} \,|\, \exists\, h_t\ \exists h_{t'} \text{ such that } h_{t'} \succ h_t,\ \varphi_{x_i}(h_t) = a_i^e \text{ and } \varphi_{x_i}(h_{t'}) \downarrow\neq a_i^e\} \tag{B.1}$$

and

$$\widehat{\mathcal{X}}_i \;=\; \{x_i \in \mathbb{N} \mid \varphi_{x_i}(h_t) \downarrow\ \forall\, h_t \in \mathcal{H}_t \text{ with } t \leq x_i\} \tag{B.2}$$

Observe that $\widehat{\mathcal{Z}}_i$ and $\widehat{\mathcal{X}}_i$ are respectively the same as $\overline{\mathcal{Z}}_i$ and $\overline{\mathcal{X}}_i$ but with x_i ranging in $\mathbb{N}$ rather than in $\mathcal{S}_i$.

By Church's thesis, it is easy to establish that the functions

$$f_{\widehat{\mathcal{Z}}_i}(x_i) = \begin{cases} 1 & \text{if } x_i \in \widehat{\mathcal{Z}}_i \\ \uparrow & \text{otherwise} \end{cases} \qquad \text{and} \qquad f_{\widehat{\mathcal{X}}_i}(x_i) = \begin{cases} 1 & \text{if } x_i \in \widehat{\mathcal{X}}_i \\ \uparrow & \text{otherwise} \end{cases} \tag{B.3}$$

are computable. Therefore, by Definition A.2 we can conclude that the sets $\widehat{\mathcal{Z}}_i$ and $\widehat{\mathcal{X}}_i$ are r.e. Notice next that since P is admissible and therefore supp$(P) \in \mathcal{S}_i$ we must have that

$$\overline{\mathcal{Q}}_i^P \;=\; \mathrm{supp}(P) \cap \overline{\mathcal{Q}}_i \;=\; \mathrm{supp}(P) \cap \overline{\mathcal{Z}}_i \cap \overline{X}_i \;=\; \mathrm{supp}(P) \cap \widehat{\mathcal{Z}}_i \cap \widehat{X}_i \tag{B.4}$$

Lastly, recall that by Lemma B.1 we know that supp(P) is r.e. Therefore, using the right-hand side of (B.4), we know that $\overline{\mathcal{Q}}_i^P$ is the intersection of a three r.e. sets. By Lemma A.5 this is clearly enough to prove the claim. ■

Lemma B.3: *There exists a computable function d_i from $\mathbb{N}^5$ to $\mathbb{N}$ such that $\forall (x,p,q,m,k) \in \mathbb{N}$, whenever (p,q,m) is a basis (as in Definition 7) for an admissible probability distribution $P \in \mathcal{P}_i$, and $P(x) > 0$ we have $d_i(x,p,q,m,k) = \tilde{t}$, where $\tilde{t}$ satisfies*

$$\frac{1}{k}\varphi_p(x) > P(\overline{\mathcal{Q}}_i) - \sum_{\tau=0}^{\tilde{t}-1} \varphi_p(\varphi_m(\tau)) \tag{B.5}$$

PROOF: Identical to the proof of Lemma 1 of A-S.

LEMMA B.4: *There exists a computable function g_i from $\mathbb{N}^6$ to $\mathbb{N}$ such that $\forall (x, p, q, m, k, h_t) \in \mathbb{N}^6$ whenever (p, q, m) forms a basis for an admissible probability distribution $P \in \mathcal{P}_i$, and $P(x) > 0$, we have*

$$g_i(x, p, q, m, k, h_t) \equiv \begin{cases} a_i^e \in \mathcal{A}_i & \text{if } h_t \text{ has } t \geq \tilde{t} \\ a_i \in \mathcal{A}_i \;\; \text{with } a_i \neq a_i^e \text{ and } a_i \neq \varphi_{\varphi_r(m,t,\tilde{t})}(h_t) & \text{if } h_t \text{ has } t < \tilde{t} \end{cases} \tag{B.6}$$

where $\tilde{t}$ is as in Lemma B.3, and $\varphi_r(m, t, \tilde{t})$ is the t-th element in ascending order of the set $(\varphi_m(0), \ldots, \varphi_m(\tilde{t} - 1))$. Moreover, for any $(x, p, q, m, k) \in \mathbb{N}^5$, either $g_i(x, p, q, m, k, h_t) \uparrow$ for all h_t or $g_i(x, p, q, m, k, h_t) \downarrow$ for all h_t.

PROOF: A Turing machine d that computes g_i can be constructed as follows. Start by computing the value of $\tilde{t}$ as in Lemma B.3. If this computation does not halt, leave the output of d undefined. If this computation halts, proceed further as follows.

Compute all the elements of the set $(\varphi_m(0), \ldots, \varphi_m(\tilde{t} - 1))$. Observe that if (p, q, m) is a basis for an admissible probability distribution, then m computes a total computable function and hence all these computations halt. If any of the computations $\varphi_m(t)$ for $t = 0, \ldots, \tilde{t} - 1$ do not halt, leave the output of d undefined. If all these computations halt, proceed further as follows.

Rearrange the elements of the set $(\varphi_m(0), \ldots, \varphi_m(\tilde{t} - 1))$ in ascending order. In other words, compute $\varphi_r(m, t, \tilde{t})$ for every $t = 0, \ldots \tilde{t} - 1$. Notice that this step is clearly feasible by Church's thesis since it involves rearranging a finite set.

Compute the results of all the computations $\varphi_{\varphi_r(m,\tau,\tilde{t})}(h_\tau)$ for every $\tau = 0, \ldots \tilde{t} - 1$ and for every $h_\tau \in \mathcal{H}_\tau$. Observe that if (p, q, m) is a basis for an admissible probability distribution, then all these computations halt. This is because, by construction (the elements of $(\varphi_r(m, 0, \tilde{t}), \ldots, (\varphi_r(m, \tilde{t} - 1, \tilde{t}))$ are in ascending order), we know that $\varphi_r(m, \tau, \tilde{t}) \geq \tau$. This coupled with the fact that $\varphi_r(m, \tau, \tilde{t}) \in \overline{\mathcal{X}}_i$ implies that the result of the computation $\varphi_{\varphi_r(m,\tau,\tilde{t})}(h_\tau)$ is defined for every $h_\tau \in \mathcal{H}_\tau$. If any of the computations $\varphi_{\varphi_r(m,\tau,\tilde{t})}(h_\tau)$ do not halt, leave the output of d undefined. Otherwise, proceed further as follows.

Check whether h_t is such that $t \geq \tilde{t}$. If this is the case, simply set the output of d to be equal to a_i^e. If h_t is such that $t < \tilde{t}$ then set the output of d to be different from both a_i^e and $\varphi_{\varphi_r(m,t,\tilde{t})}(h_t)$ (the result of the latter computation was in fact computed in the previous step) . This step is always feasible since, by Assumption 2, $\mathcal{A}_i$ has at least three elements.

Clearly the procedure we have outlined defines a d that computes a function g_i as required. Clearly if $P(x) > 0$ and (p, q, m) is a basis for an admissible probability distribution then $\tilde{t}$ as in Lemma B.3 and the value of g_i is defined for every h_t. Finally, it is clear that, by construction, for any $(x, p, q, m, k) \in \mathbb{N}^5$, either $g_i(x, p, q, m, k, h_t) \uparrow$ for all h_t or $g_i(x, p, q, m, k, h_t) \downarrow$ for all h_t. ■

Definition B.1: *Given a probability distribution $P \in \Delta^\infty$, the symbol $P^{h_t} \in \Delta^\infty$ stands for the distribution P updated on the basis of history h_t using Bayes' rule.*

Definition B.2: *A Turing machine $x_i \in \mathcal{S}_i$ is said to be consistent with history h_t if and only if there exists an $x_j \in \mathcal{S}_j$ such that $\mathrm{h}_t(x_i, x_j) = h_t$. The set of machines x_i consistent with h_t is denoted by $\mathcal{D}_i(h_t)$.*

Definition B.3: *Given any h_t, the set of Turing machines that are quasi-cooperative and are consistent with h_t is denoted by $\mathcal{Q}_i(h_t) = Q_i \cap \mathcal{D}_i(h_t)$.*

Definition B.4: *A Turing machine $x_i \in \mathcal{S}_i$ is said to be almost-cooperative after h_t if and only if it is consistent with h_t and is certain to either cooperate or not halt on any history that is a continuation of h_t. The set of Turing machines x_i that are almost-cooperative after h_t is denoted by $\mathcal{F}_i(h_t)$, and is defined as*

$$\mathcal{F}_i(h_t) \;=\; \{x_i \in \mathcal{Q}_i(h_t) \;\mid\; \text{either } \varphi_{x_i}(h_{t'}) = a_i^e \;\forall\; h_{t'} \succeq h_t \text{ or } \varphi_{x_i}(h_{t'}) \uparrow \; \text{for some } h_{t'} \succ h_t\} \qquad \text{(B.7)}$$

Lemma B.5 [Communication Lemma]: *There exists $\mathcal{R}_i \subset \mathcal{S}_i$ such that $\forall\, P \in \mathcal{P}_i(\mathcal{R}_i)\; \forall\, c \in \mathbb{N}$ there exist a machine $x^*_{ic} \in \mathbb{N}$ and a $\bar{t}$ for which the following conditions hold.*[19]

$$\frac{1}{c}\, P^{\mathrm{h}_t(x^*_{ic}, x_j)}(x^*_{ic}) \;>\; P^{\mathrm{h}_t(x^*_{ic}, x_j)}(\overline{\mathcal{Q}}_i) \qquad \forall\, t \geq \bar{t}\;\; \forall\, x_j \in \mathcal{S}_j \qquad \text{(B.8)}$$

$$\mathcal{F}_i(\mathrm{h}_t(x^*_{ic}, x_j)) \;=\; \mathcal{Q}_i(\mathrm{h}_t(x^*_{ic}, x_j)) \qquad \forall\, t \geq \bar{t}\;\; \forall\, x_j \in \mathcal{S}_j \qquad \text{(B.9)}$$

$$\varphi_{x^*_{ic}}(h_t) = a_i^e \qquad \forall\, t \geq \bar{t} - 1 \qquad \text{(B.10)}$$

Proof: The *s-m-n* Theorem A.1 guarantees that there exists a total computable function $s : \mathbb{N}^5 \to \mathbb{N}$ such that

$$\varphi_{s(x,p,q,m,k)}(h_t) \simeq \varphi_x(p,q,m,k,h_t) \qquad \forall\; (x,p,q,m,k,h_t) \in \mathbb{N}^6 \qquad \text{(B.11)}$$

By the existence of a universal machine (Theorem A.2) and by Church's thesis, f_i from $\mathbb{N}^6$ to $\mathbb{N}$ defined by

$$f_i(x,p,q,m,k,h_t) \equiv g_i(s(x,p,q,m,z),p,q,m,k,h_t) \qquad \text{(B.12)}$$

[19] The value of x^*_{ic} depends on P as well as on c, and the value of $\bar{t}$ depends on P, c and x^*_{ic} itself. We suppress this from the notation whenever there is no risk of ambiguity.

where g_i is as in Lemma B.4, is a computable function. By the pseudo-fixed point Theorem A.6 we then have that $\exists\ \overline{x}_i \in \mathbb{N}$ such that

$$\varphi_{\overline{x}_i}(p,q,m,k,h_t) \simeq f_i(\overline{x}_i,p,q,m,k,h_t) \quad \forall\ (p,q,m,k,h_t) \in \mathbb{N}^5 \tag{B.13}$$

Substituting (B.11) and (B.12) in (B.13) we finally obtain that

$$\varphi_{s(\overline{x}_i,p,q,m,k)}(h_t) \simeq g_i(s(\overline{x}_i,p,q,m,k),p,q,m,k,h_t) \quad \forall\ (p,q,m,k,h_t) \in \mathbb{N}^5 \tag{B.14}$$

Consider now a *fixed* $P \in \mathcal{P}_i$ and its basis $(p,q,m) \in \mathbb{N}^3$. Suppose that for such given basis and given k we have that

$$\varphi_p(s(\overline{x}_i,p,q,m,k)) > 0 \tag{B.15}$$

where $\overline{x}_i$ is the pseudo-fixed point of equation (B.13), and set

$$x^*_{ik} \equiv s(\overline{x}_i,p,q,m,k) \tag{B.16}$$

Now set $\overline{t} = \tilde{t} + 1$, where $\tilde{t}$ is as in (B.5) of Lemma B.3. From (B.14) and from the construction of g_i in Lemma B.4 we immediately know that x^*_{ic} satisfies condition (B.10) of the statement of the Lemma, as required.

To see that condition (B.9) is satisfied, simply notice that the claim follows immediately from the definitions of $\mathcal{F}_i(\cdot)$ and of Q_i, and the fact that $\varphi_{x^*_{ic}}(\mathrm{h}_{\overline{t}}(x^*_{ic},x_j)) = a^e_i$.

Since g_i is as in Lemma B.4, by construction we have that

$$\mathrm{h}_t(x^*_{ik},x_j) \neq \mathrm{h}_t(\varphi_r(m,t,\tilde{t}),x_j) \quad \forall\ x_j \in \mathcal{S}_j\ \ \forall\ t < \tilde{t} \tag{B.17}$$

Since $\tilde{t}$ is as in (B.5) of Lemma B.3, (B.17) implies

$$\frac{1}{k}\, P^{\mathrm{h}_t(x^*_{ic},x_j)}(x^*_{ic}) \ > \ P^{\mathrm{h}_t(x^*_{ic},x_j)}(\overline{Q}_i) \qquad \forall\, t \geq \overline{t}\ \ \forall\, x_j \in \mathcal{S}_j \tag{B.18}$$

and setting $k = c$ in (B.18) immediately gives condition (B.8), as required.

To close the argument we must now define $\mathcal{R}_i$ so as to ensure that (B.15) is satisfied $\forall\ P \in \mathcal{P}_i(\mathcal{R}_i)$, $\forall\ k \in \mathbb{N}$. It is clearly sufficient to set

$$\mathcal{R}_i \quad \equiv \quad \mathop{\text{Range}}\limits_{(p,q,m,k)\in\mathbb{N}^4} \quad s(\overline{x}_i,p,q,m,k) \tag{B.19}$$

and to notice that using Lemma B.4 it must be that $\mathcal{R}_i \subseteq \mathcal{S}_i$. Therefore, the proof of the Lemma is now complete. ■

LEMMA B.6: *Assume that $\mathcal{R}_i$ is as in (B.19) $\forall i = 1,2$. Fix an $\epsilon > 0$, and let an equilibrium quadruple $(x_1^E, x_2^E, P_1, P_2) \in E(\epsilon, \mathcal{R})$ be given. If it is not the case that $\Pi_i(x_i^E, x_j^E) = \pi_i^e \; \forall i = 1,2$, then for some i we have that*

$$\forall\, \delta > 0 \;\; \exists c \in \mathbb{N} \quad \textit{such that} \quad \left| \pi_i^e - \Pi_i(x_{ic}^*, x_j^E) \right| < \delta \tag{B.20}$$

where x_{ic}^* is the revealing strategy of the Communication Lemma corresponding to the level of precision $c \in \mathbb{N}$ (as in equation (B.8)) and the actual perturbation P_i of the equilibrium quadruple.

PROOF: Since it is not the case that $\Pi_i(x_i^E, x_j^E) = \pi_i^e \; \forall i = 1,2$, it must be that $a_{it}(x_i^E, x_j^E) \neq a_i^e$ for some i and for infinitely many t. Because x_{ic}^* satisfies condition (B.10) of the Communication Lemma, this implies that for some t it must be that $a_{it}(x_{ic}^*, x_j^E) \neq a_{it}(x_i^E, x_j^E)$. Let $\hat{t}$ be the minimum t for which this is the case. Also let $t^* = \max\{\hat{t}, \overline{t}\}$,where $\overline{t}$ is the date associated with x_{ic}^* as in the Communication Lemma. Therefore, by date t^*, x_{ic}^* has the following two properties: (i) it satisfies (B.8), and (ii) has revealed itself to be different from the equilibrium strategy x_i^E.

Our next step is to divide the expected continuation payoff of a generic machine $x_j \in \mathcal{S}_j$ after history $\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)$ into several components, in a way that will become useful later in the proof.

We start by defining those machines that are not only almost-cooperative after $\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)$, but are in fact certain to cooperate on input $\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)$ and on any history that follows $\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)$. Let,

$$\mathcal{C}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)) \;=\; \left\{ x_i \in \mathcal{F}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)) \mid \varphi_{x_i}(h_{t'}) = a_i^e \;\forall\, h_{t'} \succeq \mathrm{h}_{t^*}(x_{ic}^*, x_j^E) \right\} \tag{B.21}$$

Observe that clearly it is the case that $\mathcal{C}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)) \subseteq \mathcal{F}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E))$. Let also $\overline{\mathcal{C}}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E))$ be the complement of $\mathcal{C}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E))$ in $\mathcal{F}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E))$.

The continuation (expected) payoff to machine $x_j \in \mathcal{S}_i$ after $\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)$ can be written as

$$\sum_{x_i \in \mathcal{Q}_i \cup \overline{\mathcal{Q}}_i} P_i^{\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)}(x_i)\, \Pi_j(x_i, x_j \mid \mathrm{h}_{t^*}(x_{ic}^*, x_j^E)) \tag{B.22}$$

Because of (B.9) of the Communication Lemma, we can re-write (B.22) as

$$\sum_{x_i \in \mathcal{C}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)) \cup \overline{\mathcal{C}}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)) \cup \overline{\mathcal{Q}}_i} P_i^{\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)}(x_i) \Pi_j(x_i, x_j \mid \mathrm{h}_{t^*}(x_{ic}^*, x_j^E)) \tag{B.23}$$

Observe next that if $x_i \in \overline{\mathcal{C}}_i(\mathrm{h}_{t^*}(x_{ic}^*, x_j^E))$ then we know that for some τ it will be the case that $\varphi_{x_i}(h_t) \uparrow$ for all h_t with $t \geq \tau$. Now let $\overline{x}_j$ be any machine in $\mathcal{S}_j$ that plays the cooperative action a_j^e on $\mathrm{h}_{t^*}(x_{ic}^*, x_j^E)$ and on every continuation history that follows it. Then, using (B.23), we know that the (expected) continuation payoff to $\overline{x}_j$ after history $\mathrm{h}_{t^*}(x^*_{ic}, x^E_j)$ is greater or equal to

$$\pi^e_j P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\mathcal{C}_i(\mathrm{h}_{t^*}(x^*_{ic}, x^E_j))) + \pi_j(\uparrow, a^e_j)P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\overline{\mathcal{C}}_i(\mathrm{h}_{t^*}(x^*_{ic}, x^E_j)) + \pi^w_j P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\overline{\mathcal{Q}}_i) \tag{B.24}$$

where π^w_j is the lowest payoff that j can achieve in any outcome of the stage game G.

Similar reasoning, using (B.21), now shows that the (expected) continuation payoff to x^E_j, after the history of play $\mathrm{h}_{t^*}(x^*_{ic}, x^E_j)$ is less than or equal to

$$\begin{aligned}\Pi_j(x^*_{ic}, x^E_j|\mathrm{h}_{t^*}(x^*_{ic}, x^E_j))P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\mathcal{C}_i(\mathrm{h}_{t^*}(x^*_{ic}, x^E_j))+ \\ \pi_j(\uparrow)P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\overline{\mathcal{C}}_i(\mathrm{h}_{t^*}(x^*_{ic}, x^E_j)) + \pi^e_j P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\overline{\mathcal{Q}}_i)\end{aligned} \tag{B.25}$$

where $\pi(\uparrow) = \max_{a_j \in \mathcal{A}_j} \pi_j(\uparrow, a_j)$.

Our next step will be that of comparing the lower bound in (B.24) with the upper bound in (B.25). To this end, it is useful to establish a piece of notation for the 'fraction of the time' that a given machine x_j does not cooperate when playing against x^*_{ic} after history $\mathrm{h}_{t^*}(x^*_{ic}, x_j)$. For any $x_j \in \mathcal{S}_j$, let[20]

$$z_j(x_j, c) \;=\; \liminf_{T\to\infty} \frac{\|\{\, t \leq T \mid \mathrm{a}_{jt}(x^*_{ic}, x_j|\mathrm{h}_{t^*}(x^*_{ic}, x_j)) \neq a^e_j \,\}\|}{T} \tag{B.26}$$

It is also convenient to establish a piece of notation for the difference between the efficient payoff π^e_j and the second largest payoff that j can achieve in any outcome of G. Let $\pi^s_j = \max\{\pi_j \in \mathbb{R}$ such that $\pi_j \neq \pi^e_j$ and $\pi_j = \pi_j(a)$ for some $a \in \mathcal{A}\}$, and define

$$\Delta_j \;=\; \pi^e_j \;-\; \pi^s_j \tag{B.27}$$

From the lower bound in (B.24) and the upper bound in (B.25), and using (B.26), (B.27) and the best response part of Assumption 3, we can now compute a lower bound for the difference in expected continuation payoff to the cooperative strategy $\overline{x}_j$ and the equilibrium strategy x^E_j. In fact, the following must hold

$$\begin{aligned}\textstyle\sum_{x_i\in\mathcal{S}_i} P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(x_i)\Pi(x_i, \overline{x}_j|\mathrm{h}_{t^*}(x^*_{ic}, x^E_j)) - \\ \textstyle\sum_{x_i\in\mathcal{S}_i} P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(x_i)\Pi(x_i, x^E_j|\mathrm{h}_{t^*}(x^*_{ic}, x^E_j)) \;\geq \\ z_j(x^E_j, c)\Delta_j P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\mathcal{C}_i(\mathrm{h}_{t^*}(x^*_{ic}, x^E_j))) + (\pi^w_j - \pi^e_j)P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\overline{\mathcal{Q}}_i)\end{aligned} \tag{B.28}$$

Observe now that the left-hand side of (B.28) must in fact be less than or equal to zero. This is simply because equilibrium strategies are required to maximize expected payoffs from the beginning of G^∞ and hence must be optimal after any history that take place with positive probability.

[20] As is standard, the notation $\|\cdot\|$ denotes the cardinality of a set.

Moreover, note that since, by construction, $x^*_{ic} \in \mathcal{C}_i(\mathrm{h}_{t^*}(x^*_{ic}, x^E_j))$ we also know that

$$P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\mathcal{C}_i(\mathrm{h}_{t^*}(x^*_{ic}, x^E_j)) \;\geq\; P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(x^*_{ic}) \tag{B.29}$$

Using these two facts it is now immediate to see that (B.28) implies

$$0 \;\geq\; z_j(x^E_j, c)\Delta_j P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(x^*_{ic}) + (\pi^w_j - \pi^e_j)P_i^{\mathrm{h}_{t^*}(x^*_{ic},x^E_j)}(\overline{\mathcal{Q}}_i) \tag{B.30}$$

Using (B.8) of the Communication Lemma, and rearranging, (B.30) directly implies that

$$0 \;\geq\; c\, z_j(x^E_j, c)\,\Delta_j \;+\; \pi^w_j - \pi^e_j \tag{B.31}$$

Since (B.31) must hold for all c, and since we know that $\Delta_j > 0$, we can now conclude that

$$\lim_{c\to\infty} \; z_j(x^E_j, c) \;=\; 0 \tag{B.32}$$

Since (B.32) tells us that machine x^E_j plays the cooperative action 'almost all the time' against x^*_{ic} after history $\mathrm{h}_{t^*}(x^*_{ic}, x^E_j)$, we can also conclude that

$$\lim_{c\to\infty} \; \Pi_j(x^*_{ic}, x^E_j \,|\, \mathrm{h}_{t^*}(x^*_{ic}, x^E_j)) \;=\; \pi^e_j \tag{B.33}$$

Since G is a game of common interest, (B.33) implies that

$$\lim_{c\to\infty} \; \Pi_i(x^*_{ic}, x^E_j \,|\, \mathrm{h}_{t^*}(x^*_{ic}, x^E_j)) \;=\; \pi^e_i \tag{B.34}$$

Finally, since the infinitely repeated game G^∞ is undiscounted, (B.34) implies that

$$\lim_{c\to\infty} \; \Pi_i(x^*_{ic}, x^E_j) \;=\; \pi^e_i \tag{B.35}$$

as required by the statement of the Lemma. ■

Lemma B.7: *There exists $\mathcal{R}_i \subset \mathbb{N}$ $i = 1, 2$ such that (provided the equilibrium set is not empty)*

$$\Pi^E(\mathcal{R}) = \pi^e \tag{B.36}$$

Proof: Given Lemma B.6 the proof is identical to the proof of Lemma 5 in A-S. The details are omitted. ■

Lemma B.8: *Let $\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2)$ be as in (B.19) of the proof of the Communication Lemma B.5. Then $\Pi^E(\epsilon, \mathcal{R}) \neq \emptyset$ for any ϵ with $0 \leq \epsilon \leq 1$.*

Proof: The proof of this claim is constructive and is, mutatis mutandis, identical to the proof of Lemma 6 in A-S. The details are omitted. ■

Proof of Theorem 1: The claim is a direct consequence of Lemma B.7 and Lemma B.8. ■

References

Anderlini, L. (1999): "Communication, Computability and Common Interest Games," *Games and Economic Behavior*, 27, 1–37.

Anderlini, L., and H. Sabourian (1990): "Cooperation and Effective Computability," Economic Theory Discussion Paper 159, Department of Applied Economics, University of Cambridge.

——— (1995): "Cooperation and Effective Computability," *Econometrica*, 63, 1337–1369.

——— (1998): "Cooperation and Computability in N-Player Games," University of Cambridge, Faculty of Economics and Politics, mimeo.

Cutland, N. J. (1980): *Computability: An Introduction to Recursive Function Theory.* Cambridge: Cambridge University Press.

Evans, R. A., and J. Thomas (1997): "Cooperation and Punishment," University of Cambridge, mimeo.

——— (1998): "Cooperation and Punishment," University of Cambridge, mimeo.

Rogers, H. (1967): *Theory of Recursive Functions and Effective Computability.* London: McGraw-Hill Book Company.