COMMUNICATION, COMPUTABILITY AND COMMON INTEREST GAMES

Luca Anderlini

St. John’s College Cambridge
and Harvard University

October 1989

*Most of the research work for this paper was carried out while the author was visiting the Santa Fe Institute which specializes in the study of the sciences of complexity. The Institute’s generous hospitality is gratefully acknowledged. I should also like to thank Robert Aumann, Ken Binmore, Sergiu Hart, Andreu Mas-Colell, Eric Maskin, Hamid Sabourian, Hal Varian and John Vickers for their stimulating comments. The usual disclaimer, of course, applies.*
Abstract

This paper develops a theory of pre-play communication in one-shot normal-form games. The theory is founded on the postulate that players are Turing Machines or some equivalent computing devices.

The effectiveness of a pre-play stage of unlimited costless communication in conveying information about the players' actions at the second stage of the game is made possible precisely by the fact that players are taken to be computing devices. Messages reveal information about players' general reasoning and information-processing abilities and habits, and hence about their actions in the second stage of the game.

The theory of communication developed in the paper is shown to solve the problem of coordination (equilibrium selection) in any common interest game. When pre-play communication is allowed as modelled here, the only equilibrium of a common interest game which survives the appropriate version of the trembling hand perfection requirement is the unique pareto-efficient of the game.
1. **Introduction**

In this paper, I shall develop a theory of pre-play communication in games. The basic assumption will be that the set of possible players coincides in some appropriate sense with the set of Turing Machines or any equivalent model of computation.

The positive application of such theory of communication which I shall present is to a class of two-person one-shot strategic-form\(^1\) games known as common interest games. A game of common interest between, say, two players is a game in which a unique pair of pay-offs is available which strictly Pareto-dominates all other pay-offs available in the game.\(^2\) So called 'pure coordination' games are a subset of the class of common interest games.

The main result of the paper is that when pre-play communication is allowed and modelled as I shall do below the only equilibrium of a common interest game which survives the appropriate version of the 'trembling hand' perfection requirement is the pareto-efficient equilibrium of the game.

The model of pre-play communication which I develop below is quite radically different from the ones available in the literature. I will relate the present model to the pre-existing literature in the concluding section of the paper. In particular I will attempt to compare and contrast my results with those of the signalling literature (Cho and Kreps, (1987)), (Kreps, (1989)) and of the more specifically related 'cheap talk' literature such as Crawford and Sobel (1982) and especially Farrel (1983), Farrel (1987), Farrel (1988) and Farrel and Gibbons (1989).

An example seems to be the best way to introduce and motivate the analysis which follows.

Consider a pure coordination game like the following

---

\(^1\)By strategic-form it is meant what is often also called normal form.

\(^2\)Thus, it is possible that more than one pair of strategies leads to the Pareto-efficient pay-offs. The term Common Interest Game is borrowed from Aumann and Sorin (1989). The definition of Common Interest Game used here is exactly equivalent to theirs.
where player 1 chooses rows and player 2 columns. This game has two (pure-strategy) Nash equilibria: \((A_1; A_1)\) and \((B_1; B_2)\). Moreover, both equilibria survive just about any refinement of the Nash equilibrium concept which has been put forward in the literature.

There is, however, a wide consensus that the pareto-inferior equilibrium \((B_1; B_2)\) is in some sense 'unlikely' to prevail. An informal argument often put forward in order to discard the \((B_1; B_2)\) equilibrium runs roughly as follows. The pair \((A_1; A_2)\) makes both players better off relative to \((B_1; B_2)\), and it is self-enforcing\(^3\) (it is a Nash equilibrium). Hence 'rational' players will find a way to coordinate on \((A_1; A_2)\) rather than on \((B_1; B_2)\). Very often, implicit in lines of reasoning like the one above is that the obvious 'way to coordinate' is pre-play communication. The main aim of this essay is to give a formal account of how such pre-play communication may work to eliminate the inefficient \((B_1; B_2)\) equilibrium in the above game.

Although previous literature has often informally invoked pre-play communication to eliminate equilibria like \((B_1; B_2)\) above, a formal account of such communication process has not, to my knowledge, been given before.

In the existing literature, one approach has been simply to assume that 'collective rationality' guarantees that only equilibria which are not pareto-dominated by other equilibria will be selected. This is, very roughly speaking, the postulate on which most of the 'renegotiation' literature is based. This approach yields interesting problems and insights when a framework which is sequential in nature (for instance a repeated game, as

---

\(^3\)Here, I am using a simple-minded notion of self-enforcing which is equivalent to Nash Equilibrium. Subtle issues are involved in deciding whether this is indeed the correct notion of self-enforcing, as Aumann (1989) points out. I shall reform to this point in Section 7 below.
in Farrel and Maskin (1989)) is adopted. It does, however, bypass completely the modelling of the communication (or renegotiation) process by simply assuming that only equilibria which satisfy certain collective rationality axioms are viable in the first place. Thus, as far as the scope of this essay is concerned, the renegotiation literature disposes of the interesting question by just assuming it away.

An interesting approach to incorporating the effects of pre-play communication into a refinement of the Nash equilibrium concept is provided by the cheap-talk literature. The refinement proposed in Farrel (1983) and Farrel (1988) can, very broadly speaking, be justified in terms of the following intuitive story. Before the game is played, one player can make a suggestion to the other player on how to play the game. Only suggestions which are consistent (self-enforcing) can be expected to be followed. In addition one imagines that if a suggestion is both consistent and to the advantage of both players, then it will necessarily be followed. It is then relatively easy to conclude that in the game above the only viable equilibrium is in fact $(A_1; A_2)$.

One crucial point to note about the cheap-talk literature is that the appeal to a communication process is only an intuitive and suggestive way to justify a particular refinement of the Nash equilibrium concept. Pre-play communication is, again, not explicitly modelled, but rather its effects embodied in a set of axioms which are then used to single out a subset of the Nash equilibria as being the viable ones.

The usual trouble with modelling pre-play communication in traditional game theory is an obvious one. Imagine a two-stage set-up in which players are first allowed to 'communicate' and then are asked to play the coordination game above. If it is the case that pre-play communication does not affect players' pay-offs, then there is always an equilibrium of the two-stage game (pre-play communication and play) in which players simply ignore whatever happened at the communication stage and play the 'wrong' equilibrium $(B_1; B_2)$.

Such a failure of communication to affect the set of equilibrium pay-offs can be
intuitively explained as follows. It is well known that to justify the very idea that players will indeed play any Nash equilibrium, recourse to common knowledge assumptions is inevitable (cf. Brandenburger and Dekel (1989) for a survey). The standard justification for the certainty that players will play a Nash equilibrium of the game is common knowledge of rationality and of beliefs. A simple, but compelling, view of the reasons of the ineffectiveness of pure communication as exemplified above is then the following. With postulated common knowledge of rationality and of beliefs, what is there left of communicate? With common knowledge of rationality and of beliefs, the communication stage is ineffective simply because it has no content.

Consider again the two-stage set-up described above and imagine now that the entire two-stage set-up be 'perturbed' in the following sense. A probability distribution is placed on players' 'types'. A type is now a pair specifying what such type does during the communication stage and a map from the outcome of the communication stage into actions taken at the play stage of the game.

There is now evidently room for real communication in the model. Imagine, say, player 1 observing a message $C_2$ being uttered by player 2 at the communication stage. Player 1 can now update his beliefs about player 2 in Bayesian fashion on the basis of $C_2$. Of course, one can imagine prior distributions over types which will still render the communication stage totally ineffective. One could, for instance, imagine that the prior distribution over the types form player 2 is such that, whatever the message $C_2$ observed by player 1, the Bayesian posterior over player 2 types is exactly the same as the prior probabilities. The main point here is that by perturbing the two-stage set-up as described above one is, at least in principle, 'making room' for a communication stage which has effects on the equilibrium pay-offs of the game.

The obvious first difficulty which the perturbation approach informally described above is how to find a natural way to place probability distributions over pairs of communication strings and maps from such strings into actions. As noted above, allowing
any such distributions results in a communication stage which at least in some cases will have no effect on equilibrium pay–offs.

One possible approach to the problem is to imagine that players are 'programs' or 'computing devices'. The same computing device has to first produce a communication string and then map the outcome of the communication stage into actions of the coordination game above. What was called a 'type' above is now a computing device. Placing a probability distribution over types is now the same as (and as easy as) placing a probability distribution over a class of computing devices. This is precisely the approach taken in this paper.

An obvious worry of the computing devices approach just outlined is the following. Is it not the case that by restricting attention to computing devices one is introducing crucial constraints of bounded rationality? Are then any results obtained to be interpreted as driven by bounded rationality or communication, or both? The short answer to these questions is that one can identify and consider a most (in some appropriate sense) powerful class of computing devices. This is what is done in the sequel. The class of computing devices considered is the class of Turing machines. Hence none of the results can be deemed to depend on the fact that the class of computing devices considered is too narrow. Moreover, the appropriateness identifying what Turing Machines can compute with what human decision makers can compute is defensible in very general terms. This I briefly do in the next section.

Not too surprisingly, communication will not be always effective in eliminating the \((B_1;B_2)\) equilibrium of the two–stage set–up described above when any probability distributions over Turing machines is allowed. It turns out however that communication is effective in eliminating the pareto–inferior equilibrium when a surprisingly wide class of probability distributions over Turing Machines is allowed. In particular, the prior distributions will have to be sufficiently 'diffuse' and themselves (approximately) computable by a Turing machine.
The requirement of diffuseness mentioned above will, intuitively speaking, provide a rich enough 'vocabulary' for communication to be effective in eliminating the \((B_1;B_2)\) equilibrium. The computability requirement has an elegant interpretation. If one takes seriously the idea that players are computing devices then their prior beliefs have to capable of being computed (Megiddo (1989)) by such computing devices. If one wants to be able to interpret the Turing Machines used in the analysis below as acting in a Bayesian fashion on the basis of the communication string observed, the restriction that the prior probability distributions over players be computable by a Turing Machine is a completely natural one.

Finally, one should clarify the equilibrium concept being used here, and its possible justifications. Consider again the two-stage set-up described above. One simple way to view the equilibrium concepts formally defined below is the following. Meta-players have strategy sets which consist of (all possible) Turing Machines. A choice of strategies by the two meta-players will then result in two Turing Machines first going through the communication stage and then picking actions in the play stage of the game on the basis of the outcome of the first stage. Such game between meta-players is certainly well defined. What is considered below are the trembling-hand perfect equilibria of such game where the probability distributions over strategies (Turing Machines) are restricted to be themselves computable by a Turing Machine.

The interpretation of the equilibrium concept used below which I have just outlined is certainly useful in sharpening one's intuition about the analysis which follows. It is not the only possible one, however. It is not the author's favourite option either.

The main problem with the meta-player interpretation is simple to outline. Suppose that one were to take seriously the idea that players should not be assumed to perform tasks which are not computable by a Turing Machine. (This, one should ineed do, I will try to argue in the next section.) The task of the meta-players of choosing optimal
Turing Machines can be easily shown to be not always computable by a Turing machine.\footnote{Gilboa (1988) analyses the complexity of the task of meta-players choosing finite automata to play a repeated game.}

The obvious alternative interpretation of the equilibrium concept outlined above is an evolutionary one. This seems to be by far the most appealing option available. Successive generations of Turing Machines are matched with given probabilities. The probabilities themselves evolve through time according to the score of a particular Turing Machine relative to, say, the average in that generation.

Unfortunately no general convergence results for evolutionary systems of this type are available. Fudenberg and Kreps (1989) and Friedman (1987) report on useful results, of a local nature, for systems in which a finite number of strategies are allowed. In what follows the fact that an infinite number of possible Turing Machines is taken into account is crucial, however. This limits the applicability of any existing results to the present framework.

The analysis of an evolutionary dynamical system generated by the present model seems to pose a number of interesting independent, technical and non-technical, questions. Although part of the same broad research agenda, this is clearly beyond the scope of this already overgrown essay.

The next section of the paper is devoted to a further defense of the modelling approach taken in this paper and to some mathematical preliminaries. Section 3 describes the model. Section 4 is brief and simply defines common interest games. In section 5, I state and prove a 'Communication Lemma' which in some sense drives all the results in the paper. In section 6, I apply the Communication Lemma to common interest games with pre-play communication. The results contained in this section state that, when the appropriate version of trembling-hand perfection is imposed, the only equilibrium pay-offs which survive are the pareto-efficient ones. Pre-play communication is enough to solve the problem of coordination in common interest games. Section 7 informally discusses the
interpretation of the main result of this paper when the Common Interest game at hand has a different structure from the pure coordination example above. The effects of pre-play communication as modelled in this paper to some games outside the class of Common Interest games like battle of the sexes, are also informally discussed in this section. Section 8 briefly and informally presents some results concerning the implication of the theory of communication developed in this paper for the theory of infinitely repeated games with no discounting. A formal account of these results can be found in a separate paper (Anderlini and Sabourian (1989)). Finally, section 9 further relates the present work to some of the pre-existing literature and presents some concluding remarks.

2. Church's Thesis, Turing Machines and Godel Numbers

Ken Binmore (1987) in a less-than-deservedly well known recent paper proposed the idea to model players using one of the models of computation available in the mathematical literature. The simple, but fundamental, idea behind this program of research runs as follows. A well developed branch of mathematics concerns itself with what functions are or are not effectively computable in the most general sense possible. Effectively computable roughly means the class of functions for which could possibly be computed by any imaginable finite device. If one takes this branch of mathematics seriously at all, it then follows that one should not assume that players perform tasks which involve evaluating functions which are not computable in the above sense. The consequences of this basic stance are, I believe, far-reaching. Some of its implications are analyzed in Binmore (1987), Anderlini (1988), Anderlini (1989), Canning (1988), Howard (1988) and McAfee (1984).

There are at least two convincing reasons for upholding the theory of computability or recursive function theory as appropriate for the purpose mentioned above. The first is
the 'empirical evidence' for Church's Thesis. Historically, many different models of computation have been proposed in order to explore the notion of effective computability. All of them give rise (at most) to the same class of functions — the class of (general) recursive functions. The second reason is that the models of computation which give rise to the class of (general) recursive functions are generally accepted in mathematics as being able to encompass the whole of mathematical logic. It seems that even just one quantum of modesty alone should be enough to accept the proposition that economic agents should not overall be smarter than the whole of mathematical logic!

The stance that economic agents should not be able to compute noncomputable functions naturally leads one to say that it is correct to think of economic players as being computing devices. The idea that players are Turing Machines (or some equivalent computational device), combined with the limitations on their abilities to compute which this entails, is what lies at the centre of the theory of communication which I will develop in this paper.

A Turing machine is a finite computing device. It can be shown that any (general) recursive function can be computed by a Turing machine such as the one which I define in Appendix 1 for the sake of completeness only. Many alternative definitions could have been given. What matters here is the class of functions computed by the set of computing devices considered. For a primitive definition of general recursive functions the reader is referred to Cutland (1980). Cutland (1980) also contains a useful discussion of the fact that the class of (general) recursive functions coincides with the widest possible intuitive notion of computability.

Any Turing Machine is essentially identified by its 'program' (cf. Appendix 1), which is a finite string of symbols drawn from a given alphabet. Using a technique known

---

5 The hypothesis that the class of (general) recursive functions coincides with the correct intuitive notion of 'effective computability' is known in mathematics as Church's Thesis.
6 More detailed accounts of Church's Thesis can be found in Davis (1958) Rogers (1967) or Cutland (1980). A brief presentation can also be found in Anderlini (1989).
as Gödel numbering it can then be shown that (cf. Cutland (1980) or Davis (1958) or, for a brief exposition, Anderlini (1989)) there is an effective procedure for coding and decoding each Turing machine into a natural number and vice versa. The technique used to show this fact allows one to code and decode into the naturals and vice versa any string of characters drawn from any arbitrary fixed alphabet. It follows that the possible inputs of computations of Turing machines as well as their outputs, can also all be coded and decoded into the naturals and vice versa. In other words, this framework implies that there is never any loss of generality in considering only functions from the natural numbers into the natural numbers.

The set of natural numbers will be denoted by \( \mathbb{N} \) throughout the paper. The result of the computation of Turing machine with code number \( X \in \mathbb{N} \) on the input with code number \( Y \in \mathbb{N} \) will be denoted by \( \{X\}(Y) \). Whenever different parts of a computation's input need to be treated as separate variables, these will be separated by semi-colons. Of course, the result of the computation of machine \( X \) on input \( Y \) need not be defined. The reason is that the computation of machine \( X \) on input \( Y \) may not halt (cf. Appendix 1). Also the result of the computation of machine \( X \) on input \( Y \) may not be defined simply because the number \( X \in \mathbb{N} \) does not encode a string of symbols which is admissible as a machine. In either case the fact that the computation of \( X \) on input \( Y \) is not defined will be indicated by \( \{X\}(Y) \). The fact that the result of the computation of machine \( X \) on input \( Y \) is defined will be denoted by \( \{X\}(Y) \).

One last remark about notation which simply stresses again a point already made above is that since one only needs to refer to codes in \( \mathbb{N} \) (or Gödel numbers) for both machines and their inputs, the notation \( \{X\}(X) \) although strange at first sight is perfectly admissible in this context.

---

7This immediately implies that there are some functions which are not computable. There are \( 2^{\aleph_0} \) functions from the naturals into the naturals, but only countably many are computable.
3. The Model

The set-up which I will consider is extremely simple. Two players (Turing Machines) $X_{\in N}$ and $Y_{\in N}$ 'meet'. First they are given as input a special symbol, $\cdot$. Such special symbol is interpreted as signalling to the players their opportunity to 'talk'.

The outputs of machines $X$ and $Y$ on the special symbol $\cdot$ are taken to be the Gödel numbers of the two machines' 'communication strings'. Let, if both computations halt,

$$\{X\}(\cdot) = C_x \quad \text{and} \quad \{Y\}(\cdot) = C_y$$

Notice that, as with machines and their inputs, I will be considering only the Gödel numbers of $C_x$ and $C_y$. Hence no restrictions are placed on the alphabet from which the original strings are drawn, nor on their length. All that is required is that the alphabet be fixed throughout.

After the communication stage has been carried out, and assuming both $\{X\}(\cdot)$ and $\{Y\}(\cdot)$, the communication strings are 'exchanged'. Each machine now receives as input a pair consisting of its own message and its opponent's message. The outputs of the two players on such pairs are taken to be the players' actions in a strategic form game $G$ to be defined shortly. Formally, let

---

8The fact that players receive as inputs the pairs of their own message and the opponent's message is completely inessential to the analysis that follows. All the results stated below still hold if a player's input after the communication stage consists only of its opponent's message. Moreover, some results in the preceding literature on automata playing repeated games (like, for instance, Rubinstein (1986)) seem to hinge on the use of 'reactive' strategies — that is strategies which allow the conditioning of players' actions on the opponent's moves only. The present formulation serves the purpose to clarify that this is not the case in the model discussed here. Finally, in the mathematical literature it is usual to consider a given Turing Machine as either a device computing a function of one input or of two (or of any finite number of distinct input variables). In the model described above, say, $X$ is first asked to use a single symbol $\cdot$ as input, and then the pair $(C_x; C_y)$. This is inessential, however. Everything that follows would be completely unaffected if one considered machines which take as inputs pairs of variables only. All that is needed is to modify the initial input of the machines from $\cdot$ to the pair $(\cdot; \cdot)$. 

---
where $a_x$ and $a_y$ are, again, natural numbers.

The finite action strategic form game to be played is defined in the usual manner, except that the 'labels' for players' strategies are constrained to be natural numbers as described below. $G$ is a two-player strategic-form game. I shall typically denote a particular type of player 1 by $X$ and a particular type of player 2 by $Y$. Formally $G$ consists of four elements $(\mathcal{S}_1; \mathcal{S}_2; \pi_1; \pi_2)$. $\mathcal{S}_1 \subseteq \mathbb{N}$ represents player 1's finite strategy set. In particular the number of elements in $\mathcal{S}_1$ is taken to be $N_1$. The strategies in $\mathcal{S}_1$ are then simply labelled with the natural numbers from 1 to $N_1$. Mutatis mutandis identical notation is valid for player 2. Finally $\pi_1$ and $\pi_2$ represent the two players' pay-off functions. So, for instance, if $1 \leq a_1 \leq N_1$ and $1 \leq a_2 \leq N_2$ then 1's pay-off is defined as $\pi_1(a_1; a_2)$ and 2's pay-off is defined as $\pi_2(a_1; a_2)$. For technical reasons and to simplify matters, all pay-offs of $G$ are taken to be rational numbers throughout the paper.\(^9\)

Because of the nature of the set-up, pay-offs to the two players have to be defined in some special cases. I do so below in a manner which is largely arbitrary. Simply, some of the assumptions seem to be 'natural' in this context.

The first set of special cases which needs to be considered are the ones in which one (or both) of the players' computations do not halt at the communication stage of the game. The pay-offs for such cases are assumed to be the same as in the case in which both players' computations halt at the communication stage of the game, but one (or both) of the players' computations do not halt at the second stage of the game. These are described below.

Given that $\{X\}(\bullet) \downarrow$ and $\{Y\}(\bullet) \downarrow$, either of the two players may end up not taking any action in the game for one of two conceivable reasons. As far as player 1 is concerned

\[^9\]This assumption is by far stronger than needed. It helps to keep matters simple in many respects. Canning (1988) explores alternatives to the simple short-cut adopted here.
it may be that \( \{X\}(C_x;C_y)\downarrow \) but at the same time \( \{X\}(C_x;C_y)\notin S_1 \), or simply \( \{X\}(C_x;C_y)\uparrow \). The same possibilities are true for player 2. What Assumption 1 below is designed to capture the fact that offering no output at all or offering an inadmissible output is a strictly dominated strategy for either player, and that if a player's opponent is offering no output at all or an illegal output then all actions in that player's strategy set are indeed equally good responses to the opponent's inadmissible\(^{10}\) behaviour.

With a slight abuse of notation I will state Assumption 1 by extending the domains of \( \pi_1(\cdot) \) and \( \pi_2(\cdot) \) to all possible combinations of strategies and two special symbols, \( \uparrow \) and \( \notin \). So, for instance, if \( \{X\}(C_x;C_y) = a_x \in S_1 \) but \( \{Y\}(C_y;C_x)\notin \), then X's pay-off is \( \pi_1(a_x;\notin) \). If \( \{X\}(C_x;C_y) = a_x \in S_1 \) while \( \{Y\}(C_y;C_x)\notin \), then \( \{Y\}(C_y;C_x)\uparrow \) but \( \{Y\}(C_y;C_x)\notin \), then the pay-off to X is \( \pi_1(a_x;\notin) \). If \( \{X\}(C_x;C_y)\uparrow \) and \( \{Y\}(C_y;C_x)\notin \), then the pay-off to X is \( \pi_1(\notin;\notin) \). If \( \{X\}(C_x;C_y)\notin \) with \( \{X\}(C_x;C_y)\notin \) but \( \{Y\}(C_y;C_x)\notin \) then the pay-off to X is \( \pi_1(\notin;\notin) \).

\(^{10}\)A good example (which I owe to Ken Binmore) of a situation in which these assumptions are realistic is a game of chess. At any point when it is his turn to move a player may simply decide to hit the board and knock over all the pieces on it. This is clearly a possibility which is open to him and which at the same time is not included in the set of 'legal moves'. It is a natural convention in chess that a player behaving as described will lose the game. Thus in some sense such inadmissible behaviour is a dominated strategy. Moreover it is clear that any strategy only involving legal moves during the play is an equally good response to such inadmissible behaviour. The question of not producing an output at all is more delicate. The problem here arises not in justifying the fact that not producing an output should be a dominated strategy, since a similar justification as in the case of illegal output seems plausible. In short, the problem is, how does one know when to award the (bad) pay off for not producing an output? One, admittedly contrived, but consistent interpretation runs as follows. Assume, as it seems natural in this context that the players do not discount the future at all. Assume also, for simplicity, that the pay-off to a player when he offers no output is the same, whatever the move of the opponent. Let such pay-off be \( k \). One can then imagine that when the computations start a stream of pay offs is awarded to the player as follows. \( \frac{1}{2}k \) after one unit of time has elapsed, \( \frac{1}{4}k \) after two units have elapsed and so on (\( \frac{1}{2^n}k \) after \( n \) units of time have elapsed). If at any point the player's computation halts, the player receives the payoff entailed by his move minus the fraction of \( k \) awarded through time so far. Clearly, if the player's computation does not halt, he ends up with a stream of payoffs worth exactly \( k \). Also, if the player's computation does halt, he will end up with a stream of payoffs worth exactly the payoff prescribed by the extended game.
behavior just described, then the pay-off to X is $\pi_1(\uparrow, \emptyset)$. Finally, if $\{X\}(C_x;C_y) \downarrow = n \notin S_1$ and $\{Y\}(C_y;C_x) \downarrow = n \notin S_2$ then the pay-off to X is $\pi_1(\emptyset, \emptyset)$. Totally symmetric notation holds as regards the pay-offs to player 2 — Y — in the situations outlined above.

Let $S_i^* = S_i \cup \emptyset (i = 1; 2)$. Enough notation has been established to state Assumption 1 formally.

**Assumption 1** The extended functions $\pi_i(\cdot, \cdot) (i = 1, 2)$, satisfy the following properties:

a) **(Dominance)** $\forall i = 1; 2$ and $j \neq i$ $\exists a_i \in S_i$ such that $\pi_i(\uparrow; a_j) < \pi_i(a_i; a_j) \forall a_j \in S_j^*$.  
   $\forall i = 1; 2$ and $j \neq i$ $\exists a_j \in S_j$ such that $\pi_i(\emptyset; a_j) < \pi_i(a_i; a_j) \forall a_i \in S_i^*$.

b) **(Indifference)** $\forall i = 1; 2$ $\pi_i(a_i; \uparrow) = \pi_i(a_i; \emptyset) \forall a_i \in S_i$ and $\pi_i(a_i; \emptyset) = \pi_i(a_i; \emptyset) \forall a_i \in S_i$.

Some brief final comments about Assumption 1 are in order. Property a) stipulates that not producing output at any stage of the game or producing an inadmissible output is a strictly dominated strategy for both players. Property b) requires that if the opposing player is not producing an output at any stage of the game then all actions in $S_i$ yield the same pay-off for player i. It also requires that the same applies to a player whose opponent is producing an illegal output at the play stage of the game.

Lastly, in the sequel I shall refer to a game obtained from $G$ extending the domain of the pay-off functions as above as the extended game $G^*$.

4. **Common Interest Games**

In this section I briefly review the class of games known as Common Interest games and establish some notational conventions.

**Definition 1** A two—person strategic form game $G=(S_1; S_2; \pi_1; \pi_2)$ is said to be a Common Interest game iff there exists a pair of pay—offs (which may be associated with more than
one pair of strategies) which strictly Pareto–dominates all other pairs of pay-offs available in \( G \). In symbols:

\[
\exists \ h_1 \text{ and } h_2 \text{ such that }
\begin{align*}
& a) \ h_1 = \pi_1(a_1; a_2) \text{ and } h_2 = \pi_2(a_1; a_2) \text{ for some pair } (a_1; a_2) \in S_1 \times S_2 \\
& b) \ \forall (a_1; a_2) \in S_1 \times S_2, \ (\pi_1(a_1; a_2); \pi_2(a_1; a_2)) \neq (h_1; h_2) \text{ implies } h_1 > \pi_1(a_1; a_2) \\
& \text{and } h_2 > \pi_2(a_1; a_2).
\end{align*}
\]

Hence the definition of Common Interest game adopted here is equivalent to that of Aumann and Sorin (1989). The analysis that follows is somewhat complicated by the fact that one is allowing more than one pair of strategies to lead to the Pareto–efficient outcome. The extra complications seem worthwhile, however. Constructing a strategic form game from an extensive form one it is often hard to avoid multiple pairs of strategies leading to identical pay-offs, and many applications of the theory of Games crucially rely on extensive forms.

I conclude this section by pointing out that all 'pure coordination' games (like the example in the introductory section of this paper) satisfy the definition of Common Interest game, and by establishing some notational conventions.

**Assumption 2** Let \( G \) be a strategic–form Common Interest game, and \((h_1; h_2)\) the pair of Pareto–efficient pay–offs associated with \( G \). Then

\[
h_1 = \pi_1(1; 1)
\]

and

\[
h_2 = \pi_2(1; 1)
\]

Assumption 2 is simply a re–labelling convention which will be useful in keeping notation

\[\text{I am grateful to Robert Aumann for alerting me to this point. In previous versions of this paper a more restrictive definition of Common Interest games was used.}\]
down in the sequel. The following will also prove handy below.

Definition 2 The best response sets of the two players to their opponent's choice of strategy 1 are denoted by $H_1$ and $H_2$ respectively. Formally

$$H_1 = \arg\max_{a_1 \in S_1} \pi_1(a_1;1)$$

$$H_2 = \arg\max_{a_2 \in S_2} \pi_2(1;a_2)$$

5. A Lemma on Communication

It is a property of all models of computation which generate the entire class of (general) recursive functions that there are always infinitely many machines which compute the same function (cf. Brainerd and Landweber (1974) and Cutland (1980)). Intuitively this is simply because 'dummy' lines like 'now multiply the result by 2 and then divide it by two' can always be added to any program.

Moreover, it is a property of all models of computation co-extensive with the one which I consider here, that any machine $X$ can be 'mimicked' by infinitely many other machines on chosen subsets of inputs $Q \subseteq \mathbb{N}$, with the mimicking machines producing the same output as $X$ on any $n \in Q \subseteq \mathbb{N}$ and a systematically different output on any $n \notin Q$.

Consider now an arbitrary probability distribution over the set of possible player 1 machines, $N$. Can one, in general, devise a communication string which 'identifies' a machine $X$ playing an arbitrary strategy that is a given map from pairs of messages $(C_X; C_Y)$ into actions allowed in $G$, that is $S_1$? To put the question more precisely; consider an arbitrary probability distribution over $N$. Consider now an arbitrary message string $C_X$. One can then define the updated probability distribution over machines, given $C_X$. The machines which do not output $C_X$ on input $\bullet$ have probability zero, and the probabilities of all machines which do output $C_X$ on input $\bullet$ are appropriately re-scaled.
according to Bayes’ rule. The question that is being asked is then the following. For an arbitrary probability distribution over \( N \) and an arbitrary fixed strategy in the action–choosing stage of the game (a map from pairs of messages into actions) can one always find a ‘smart’ message such that in the updated probability distribution the probability that the arbitrary fixed strategy is played is arbitrarily close to 1?

Intuitively, the answer to the above question would seem to be no. The reason, one could argue, is that for any arbitrary machine \( X \) and \( C_X = \{X\}(\cdot) \) infinitely many machines \( Y \) can be found which output the same message string as \( X \) but behave according to a different map from messages into actions at the action choosing stage of the game. Since the probability distribution is arbitrary it must be possible to construct one which makes the construction of ‘smart’ messages as described above impossible.

The intuition just outlined turns out to be wrong. Under surprisingly mild restrictions on the initial probability distribution, ‘smart’ messages can always be constructed. The reason for this is, very roughly speaking, that a surprising ‘fixed point’ theorem (Fact 2 below) over the space of machines can be applied to give what one could call ‘introspective’ machines.

Before stating the results formally, two more notational conventions need to be established. The notation \( f(X) \) has the standard meaning of a function, except that it is always taken to mean a map from \( N \) to \( N \). The symbol \( \sim \) used between two functions, two machines or any combination of these means ‘equal wherever defined’. A ‘computable’ function \( f(e) \) is simply a function \( N \to N \) such that \( \exists X \in N \) for which \( \{X\}(e) \sim f(e) \forall e \in N \). A total computable function is a computable function which is everywhere defined.

Obvious generalizations of the above terminology to the case of \( f(\cdot) \) being a function of many variables will also be used.

The first three results which I report are totally standard in the mathematical literature. I state them here for the sake of completeness only. The reader is referred to Cutland (1980) or Rogers (1967) for an easy and a more advanced treatment respectively.
Fact 1 (Existence of a Universal Turing Machine)

\[ \exists U \in \mathbb{N} \text{ such that } \{U\}(M;y) \sim \{M\}(y) \quad \forall M;y \in \mathbb{N} \]

Fact 2 (The Recursion Theorem) Let \( f : \mathbb{N} \to \mathbb{N} \) be a total computable function. Then \( f \) has a 'fixed point' in the following sense

\[ \exists X \in \mathbb{N} \text{ such that } \{X\}(e) \sim \{f(X)\}(e) \quad \forall e \in \mathbb{N} \]

Fact 3 (A simple consequence of the s--m--n Theorem). For any given \( m \geq 1 \) there exists a total computable function \( g : \mathbb{N}^m \to \mathbb{N} \) such that

\[ \{g(M;y_1;\ldots;y_{m-1})\}(y_m) \sim \{M\}(y_1;\ldots;y_m) \quad \forall M;y_1;\ldots;y_m \in \mathbb{N} \]

Some brief comments are in order. The existence of a Universal Turing Machine is not surprising if one thinks of it as a 'universal compiler' program which is able to run any program on any input. No limits on memory storage are placed since Turing Machines have an infinite memory tape.

The Recursion Theorem is not a simple result. Its (surprisingly short) proof relies on a 'diagonal' construction which I will not attempt to describe here (cf. Cutland (1980), Rogers (1967)). One point to note is that in Fact 2 above \( X \) is not quite a fixed point of \( f(\cdot) \) in the ordinary sense. In general, \( X \) will be a different number from \( f(X) \). All that the theorem states is that the two machines with Gödel numbers \( X \) and \( f(X) \) compute the same function.

Fact 3 is a simple version of a result which is usually stated in a more general form. Roughly speaking, what Fact 3 asserts is that a set of variables can always be 'encoded' in the program to give the same result on the remaining variable. The result should not
anyone acquainted with actual computers and computer programs.

One last technical preliminary result is needed before the Communication Lemma. The name which I give to the next Lemma is due to the fact that it shows the existence of a machine which computes an arbitrary computable function with its own Gödel numbers as an input. The line of reasoning in the proof is identical to Cutland (1980), Corollary 11.1.4.

**Lemma 1 (Introspective Machine).** Consider any computable $f : \mathbb{N}^2 \rightarrow \mathbb{N}$. Then there exist $X \in \mathbb{N}$ such that

\[
\{X\}(e) \preceq f(X; e) \quad \forall e \in \mathbb{N}
\]

**Proof:** Since $f$ is computable $\exists M \in \mathbb{N}$ such that $\{M\}(y; e) \preceq f(y; e) \quad \forall y, e \in \mathbb{N}$. By the s-m-n Theorem (Fact 3) there exists a total computable $g$ such that

\[
\{g(y)\}(e) \preceq f(y; e) \quad \forall y, e \in \mathbb{N}
\]

Since $g(\cdot)$ is total and computable, the Recursion Theorem applies. Hence, there exists $X \in \mathbb{N}$ such that

\[
\{X\}(e) \preceq \{g(X)\}(e) \preceq f(X; e)
\]

and this is enough to prove the Lemma.

I now come to the assumptions to be placed on the probability distributions over Turing Machines.

The Communication Lemma below relies on two sets of assumptions. The first set requires that the probability distributions considered are computable. These requirements
are embodied in Definition 3 of admissible probability distribution which follows. The second requires that the probability distributions considered are sufficiently 'diffuse' — have a sufficiently large support. This requirement is embodied in the statement of the Communication Lemma 2 itself. The justification, significance, and roles in the proof of the Lemma will all be discussed at some length after the result has been established formally.

**Definition 3** A probability distribution \( P = \{P(1), P(2), \ldots, P(n), \ldots\} \) over \( N \) is said to be admissible iff

a) \( \exists M \in \mathbb{N} \) such that for all \( n \in \mathbb{N} \) and for all rational numbers \( \epsilon > 0 \) one has
\[
\{M\}(n; \epsilon) = P(n; \epsilon),
\]
where the latter satisfies
\[
|P(n; \epsilon) - P(n)| < \epsilon.
\]

b) \( \exists H \) such that, for all rational numbers \( \epsilon > 0 \) one has \( \{H\}(\epsilon) = S(\epsilon) \) where the latter satisfies
\[
|S(\epsilon) - \sum_{i \in Q} P(i)| < \epsilon,
\]
and
\[
Q = \{i \in \mathbb{N} | \{i\}(\cdot) \models C_i \text{ and } \{i\}(C_i; \epsilon) \not\models 1 \text{ for some } \epsilon \in \mathbb{N}\}
\]

Part a) of Definition 2 stipulates that an admissible probability distribution must be, at least approximately, computable. This is, of course, consistent with some, or all, the probabilities being exactly computable.

Part b) of the definition of admissibility requires the approximate computability of the probability of the event \( i \in Q \). It should be noted that b) does not follow from a).

The set \( Q \) contains a countable infinity of elements. The set \( Q \) is the set of machines which do produce an output at the communication stage and at the same time do not always (for any message of the opponent) play action 1 (cf. Assumption 2) at the play stage of the game.

In what follows the notation \( \text{supp}(P) \) stands for the support of the probability
distribution over the natural numbers designated by \( P \).

**Lemma 2 (Effective Communication).** There exists a set \( R \subseteq \mathbb{N} \) such that for any admissible probability distribution \( P \) such that \( R \subseteq \text{supp}(P) \) and any \( 1 > \ell > 0 \), one can find \( x \in \mathbb{N} \) such that

\[
\begin{align*}
a) & \quad \{X\}(\bullet) \downarrow = C_x^- \\
b) & \quad \{X\}(C_x^-; C_y) \downarrow = 1 \quad \forall C_y \in \mathbb{N} \\
c) & \quad \ell \cdot P(X) > \sum_{i \in Q_{c_x^-}} p_i
\end{align*}
\]

where

\[
Q_{c_x^-} = \{ i \in \mathbb{N} | \{i\}(\bullet) \downarrow = C_x^- \text{ and } \{i\}(C_x^-; C_y) \downarrow \neq 1 \text{ for some } C_y \in \mathbb{N} \}
\]

**Proof:** In this argument I shall make use of a widely accepted method of proof in this area of mathematics. The claim implicit in the methodology is that whenever a 'clear procedure' can be established for computing a function then it follows that a Turing Machine (or equivalent) exists which will carry out the computation. Extreme care, of course, has to be taken with the definition of a 'clear procedure'. This way of proceeding is known as proof by Church's Thesis. Without it, even the simplest statements would require staggering amounts of constructive detail. The reader unconvinced by this brief discussion is, as usual, referred to Cutland (1980) or Rogers (1967).

The procedures (Turing Machines) which I consider below are ones which on a particular input, first output their own Gödel number, and then compute a particular function of such number. Clearly by Fact 2 (Introspective Machines) this is allowed as long as the function of its own code which the machine is required to compute is computable.

The proof is rather lengthy and hence it is divided into three steps.
Step 1. A set $L \subseteq N$ is said to be recursively enumerable (r.e.) iff there exists a machine
$H \in N$ such that $\{H\}(n) \downarrow \forall n \in N, n \neq n' \Rightarrow \{H\}(n) \neq \{H\}(n')$ and $\forall m \in L \exists n \in N$ such that
$\{H\}(n) = m$, while $m \not\in L \Rightarrow \{H\}(n) \neq m \forall n \in N$. The output of $\{H\}(n)$ is sometimes called the
$n$–th element in the (H) enumeration of $L$.

Consider the set $Q$ of Definition 3. By standard results this set is r.e. Let $q_n$ be
the $n$–th element in its enumeration, and $E_n = \bigcup_{i=1}^{n} q_i$.

Step 2. Consider now the following procedure (Turing Machine with Gödel number $\hat{X}$). On
any quadruple of inputs $(m; h; \ell; e) \in N^4$, $\hat{X}$ performs the following steps:

1) Compute own Gödel number $\hat{X}$
2) Compute $\{K\}(\hat{X}; m; h; \ell)$ such that
   $\forall e \in N \ \{\{K\}(\hat{X}; m; h; \ell)\}(e) \equiv \{\hat{X}\}(m; h; \ell; e)$
3) Set $n = 1$ (counter)
4) Compute $z(n) = \sum_{i=1}^{n} \{m\}(q_i; n)$
5) Compute $p(n) = \{m\}\{\{K\}(\hat{X}; m; h; \ell); \frac{1}{n}\}$
6) Compute $s(n) = \{h\}(\frac{1}{n})$
7) Check whether the following inequality holds
   $\ell \cdot p(n) - \frac{1}{n} > s(n) - z(n) + \frac{2}{n}$
8) If the answer to 7) is NO set $n := n + 1$ and go to 4)
9) If the answer to 7) is YES, let the current value of $n$ be denoted by $\bar{n}$ and
   proceed to 10)
10) Check whether $e = \ast$
11) If the answer to 10) is NO, output 1 and halt.
12) If the answer to 10) is YES proceed to 13)
13) Compute $C = \{q_i\}(\ast) \ \forall i = 1...\bar{n}$ with $i \neq \hat{X}$

\footnote{For more detail the reader is referred to Cutland (1980) or Rogers (1967).}
14) Output $C_{\hat{X}}$ such that $C_{\hat{X}} \neq C_{q_1}$ for all $i \neq \hat{X}$

Notice now that procedure $\hat{X}$ is well defined, but does not necessarily produce an output. However, it is easily checked that, provided that $m$ and $h$ are such that $p(n), s(n)$ and $z(n)$ are defined for all $n \in \mathbb{N}$ and

$$\lim_{n \to \infty} p(n) > 0; \lim_{n \to \infty} s(n) - z(n) = 0$$

then $\{X\}(m; h; \ell; e) \ni$ for all $e$, and $\{\hat{X}\}(m; h; \ell; e) \ni 1$ if $e \neq \bullet$, while $\{\hat{X}\}(m; h; \ell; \bullet) \ni C_{\hat{X}}$

Step 3 In this step I shall argue that the statement of the Lemma is correct when one sets $R$ equal to the range of the function computed by $K$ in step 2 over $m$, $h$ and $\ell$.

Formally, let

$$R = \{n \in \mathbb{N} | \exists (m; h; \ell) \in \mathbb{N}^3 \text{ such that } \{K\}(\hat{X}; h; m; \ell) = n\}$$

Consider now an admissible probability distribution $P$ such that $R \subseteq \text{supp}(P)$. By admissibility there exist a pair of natural numbers $M$ and $H$ with the following properties (cf. a) and b) of Definition 2) $\{M\}(n; \epsilon) = P(n; \epsilon) \forall n \in \mathbb{N}$ and any rational $\epsilon > 0$, and $\{H\}(\epsilon) = S(\epsilon)$ for all rational $\epsilon > 0$.

Consider now that machine $\hat{X} = \{K\}(X; M; H; \ell)$ where $K$ and $\hat{X}$ are as in Step 2. By the fact that $\{K\}(X; M; H; \ell) \in R$ and by the choice of $M$ and $H$, it is clear that in the computation of $\{X\}(M; H; \ell)$ described in Step 2 one has $\lim_{n \to \infty} p(n) > 0$ and

$$\lim_{n \to \infty} s(n) - z(n) = 0.$$ 

Hence, since $\{\{K\}(\hat{X}; M; H; \ell)\}(e) \ni \{X\}(M; H; \ell; e) \forall \ell \in \mathbb{N}$ one has that $\epsilon \neq \bullet$ implies that $\{\{K\}(\hat{X}; M; H; \ell)\}(e) \ni 1$, and $\epsilon = \bullet$ implies $\{\{K\}(\hat{X}; M; H; \ell)\}(e) \ni C_{\hat{X}}$

Hence, it only remains to check that property c) of the statement of the Lemma is satisfied.
To confirm that this is the case, notice that by the construction in 13) and 14) of Step 2 it must be that if \( X \in Q_{c_x} \) then \( X = q_i \) with \( i > \bar{n} \). However, simple algebra shows that the inequality in 7) of Step 2 guarantees that

\[
el \cdot P(X) > \sum_{i \in Q} P(i) - \sum_{i=1}^{\bar{n}} P(q_i)
\]

and hence the Lemma is proved.

Several comments are in order about the Communication Lemma and the hypotheses on which it rests.

Firstly, the statement of the Lemma demands that the support of the probability distribution in question be a sufficiently 'large' set. The role of such requirement in the proof of the Lemma is quite clear. An intuitive way to look at the procedure defined in Step 2 of the proof is as follows. The machine \( X \) satisfying the required properties first enumerates that machines which could potentially output the same message as \( X \). When sufficiently many machines have been enumerated, so that the remaining probability weight is 'small' relative to the probability of \( X \), a message \( C_{c_x} \) is constructed so as to differ from all the messages of the machines which have been enumerated. Hence, property c) of the statement of the Lemma can be satisfied. A necessary condition for such procedure to be successful in constructing such a 'smart' message is that the probability of \( X \) itself be strictly positive. This is precisely what is guaranteed by the stipulation that \( R \subseteq \text{supp}(P) \) in the statement of the Lemma.

The second set of remarks concerns the role of the assumption of admissibility of \( P \) in the proof of the Lemma. Firstly, without a) of Definition 2, it is intuitively clear how that argument in the proof would fail. The machine \( X \) of the statement of the Lemma, has to make sure that the weight remaining on the machines in \( Q \) which have not yet
been enumerated is sufficiently small, relative to $P(X)$ itself. The ability to compute its own Gödel number guaranteed by the recursion theorem would not be enough to guarantee the feasibility of the operation, unless a procedure for computing (at least approximately) $P(X)$ were not available.

Part b) of the definition 2 of admissibility also plays a crucial role in making the procedure $X$ well defined. As machine $X$ is enumerating the set $Q$, an 'upper bound' is needed on the probability mass on the whole of $Q$ in order to compute the weight of the machines in $Q$ which have not yet been enumerated. The computability of such appropriate upper bound is guaranteed by b) of Definition 2.

The third set of remarks concerns possible alternative set of hypotheses for the Communication Lemma. From the discussion so far, one may wonder why it was not simply assumed that $P$ assigns positive probability to all $n \in \mathbb{N}$. This indeed would seem to be compatible with the statement of the Lemma. Contrary to this prime facie intuition, it is not hard to show that any probability distribution placing strictly positive probability to all $n \in \mathbb{N}$ is not admissible in the sense of Definition 2. Technically, this is because admissibility according to Definition 2 can be shown to imply that the set $L = \text{supp}(P) \cap \overline{Q}$ (where $\overline{Q}$ is the $\mathbb{N}$ complement of $Q$) is a recursively enumerable set. By standard results, however, the set $\overline{Q}$ itself is not recursively enumerable. Hence setting $\text{supp}(P) = \mathbb{N}$ yields a contradiction. This would not be the case of b) of Definition 2 were not imposed. Thus b) of Definition 2 is in some sense a stronger requirement than may appear at first sight.

The last set of remarks about the hypotheses of the Communication Lemma concerns their defensibility on general grounds. In particular, how 'restrictive' is the assumption of admissibility? One could argue that it is very restrictive in the following sense. How many admissible probability distributions over $\mathbb{N}$ are possible before the computability restriction is placed upon them? The answer is at least $2^{\aleph_0}$. On the other hand, only countably many admissible probability distributions are possible. Such
derogatory adjectives as non-generic, non-robust etc. seem to apply to the case at hand.

A rather more sympathetic view can be taken, however. Consider a probability distribution over $N$ which is not computable in the sense of a) of Definition 2. The non-computability of $P$ can be taken as meaning that there is no procedure for approximating the probabilities. This would not seem natural if one accepts the restrictions that players are Turing Machines and interprets (as is implicit in one interpretation of the equilibrium concepts employed below) such probability distribution as the beliefs\(^{13}\) of the players prior to the communication stage. If the probability distribution is not computable in the sense of Definition 1 and if the players are Turing Machines then when interrogated about their prior belief of, say, the opposing player having Gödel number $n$, players may not be able to give even an approximate answer. This, in turn, would cast doubts on the possibility of interpreting any player's behaviour as Bayesian-rational in the second stage of the game.

Similar remarks apply to defend stipulation b) of the Definition 2 of admissibility. If b) of Definition 2 were not satisfied it would not be possible to elicit (even approximate) prior beliefs about the overall probability of machines which do not always play strategy 1 at the play stage of the game. A little more care should be excercised here than in the defense of a) of Definition 2. Indeed one even with admissible probability distributions one can construct 'events' the overall probability of which would not be even approximately computable. Then why single out precisely the set $Q$? This question can be answered in 2, not mutually exclusive, ways. The first is to say that admissibility according to Definition 2 is in same sense a minimal requirement about the computability of $P$ and of 'events' in $N$. If it is thought necessary that the overall probabilities of events other than $Q$ be computable, then extra restrictions should be imposed on $P$. This would clearly not affect the validity of the Communication Lemma. The second possible answer

\[^{13}\text{The issue of computable beliefs is explored in detail by Megiddo (1989).}\]
involves the use of Communication Lemma in the remainder of the paper, where it is shown that perturbations of the two stage game given by admissible probability distributions with a sufficiently 'large' support, guarantee that all equilibria of the two-stage game yield pareto-efficient pay-offs. It is a key insight of this paper that such 'cooperative' outcomes are guaranteed (among other things) by the ability of the players to at least approximate the probability that they are playing against an opponent which will not always play a particular 'cooperative' strategy.

6. Common Interest, Communication and Pareto-Efficiency

Consider a two-person strategic form game \( G \) and its extension \( G^* \) as defined in section 3. In this section I shall show that if \( G^* \) is an extended common interest game which satisfies Assumption 1 and pre-play communication is allowed as described in section 3, then a pair of machines which form a Nash equilibrium for the two-stage game which satisfies a 'computable' version of the trembling hand refinement must necessarily play a pair of strategies leading to the Pareto-efficient outcome.

I shall comment at length on the result and its proof once it has been presented formally. It suffices here to recall that without allowing a pre-play communication stage in the way I do here, the possible Pareto-inefficient equilibria of \( G \) (cf. the game described in the introduction) cannot, in general, be ruled out by refining the Nash equilibrium concept.

Some extra pieces of notation are necessary. Consider two players \( X \) and \( Y \). Given an extended game \( G^* \) such pair maps into a pair of pay-offs, one for each player. Let such pay-offs be \( \varphi_1(X;Y) \) and \( \varphi_2(Y;X) \) respectively. To exemplify, if 

\[
\{X\}(\cdot) = C_x \quad \text{and} \quad \{Y\}(\cdot) = C_y \quad \text{and} \quad \{X\}(C_x;C_y) = a_x \quad \text{with} \quad a_x \in S_1 \quad \text{and} \quad \{Y\}(C_y;C_x) = a_y \quad \text{with} \quad a_y \in S_2,
\]

then \( \varphi_1(X;Y) = \pi_1(a_x; a_y) \). If some of the computations do not halt or if some of the action-choosing outputs are invalid, then the extensions of the pay-off functions defined in Section 3 are used to define the functions \( \varphi_1 \) and \( \varphi_2 \). Given
that no symmetry assumption has been placed on $G^*$ a pair of probability distributions will need to be considered. The probability distributions on possible 'types' of player 1 will be denoted by $P_1$ and its elements by $P_1(i)$. Symmetric notation will be used for player 2 types.

A precise definition of the equilibrium concept employed is now possible.

**Definition 3:** A $\xi$ Computable Trembling Hand Communication Equilibrium (\$CTHE\$) for an extended game $G^*$ is a pair of Turing Machines $X^*Y^*$ and a pair of probability distributions $P_1$ and $P_2$ over $N$ such that

a) $P_1(X^*) \geq 1 - \xi$ and $P_2(Y^*) \leq 1 - \xi$

b) $P_1$ and $P_2$ are admissible.

c) $\sum_i P_2(i) \varphi_1(X^*;i) \geq \sum_i P_2(i) \varphi_1(X;i) \forall X \in N$

and

$\sum_i P_1(i) \varphi_2(Y^*;i) \geq \sum_i P_1(i) \varphi_2(Y;i) \forall Y \in N$

As anticipated, admissibility of $P_1$ and $P_2$ is not enough to guarantee that the communication stage can be used as suggested by the Communication Lemma. Therefore the following is needed

**Definition 4:** A $(\xi;R)CTHE$ is a $\xi CTHE$ such that $R \subseteq \text{supp}(P_1)$ and $R \subseteq \text{supp}(P_2)$.

The main result of the paper can now be stated.

**Theorem 1.** There exists an $R \subseteq N$ such that, for any extended common interest game $G^*$
satisfying Assumptions 1 and 2 there exists a $0 < \xi < 1$ such that $\xi < \xi$ implies

a) the set of $(\xi; R)CTHE$ is not empty

b) the equilibrium players $X^*$ and $Y^*$ are such that

$$\{X^*(\cdot) \downarrow = C_x; \{X^*(C_x; C_y) \downarrow = \bar{a}_i\}$$

$$\{Y^*(\cdot) \downarrow = C_y; \{Y^*(C_y; C_x) \downarrow = \bar{a}_j\}$$

with $\pi_1(\bar{a}_i; \bar{a}_j) = h_1$ and $\pi_2(\bar{a}_i; \bar{a}_j) = h_2$

A formal proof of Theorem 1 can be found in Appendix 2. It can be clearly divided into two parts. Existence, a) and optimality, b). The issues involved in the proof of these two statements are quite different in nature.

The intuition behind the optimality result is not hard to outline. Condition c) in the definition of a $\xi CTHE$ requires that $X^*$ be optimal against the distribution of 'types' $P_2$. In other words it should be the case that no $X \neq X^*$ scores higher than $X^*$ itself in expected terms against the distribution $P_2$.

Consider now player 1 after the pair of messages $C_x$ and $C_y$ have been produced and exchanged. As noted in the introduction, given $P_2$ and the observed $C_y$, a Bayesian posterior probability over player 2 types is well defined for any possible $C_y$.

On the basis of the posterior probability which it generates, each message $C_y$ then induces an 'updated best response' set. This is simply the set of actions which are optimal in $G^*$ for player 1 given the posterior probability distribution obtained updating $P_2$ on the basis of $C_y$.

It is obvious that the requirement of optimality c) which equilibrium machines must satisfy will make it necessarily the case that after $C_x^*$ and $C_y$ have been exchanged, $X^*$ will play an action in $G^*$ which belongs to the updated best response set for player 1 given $P_2$ and $C_y$.

The Communication Lemma of Section 4 guarantees the existence of a machine $Y$ with the following properties. At the action-choosing stage of the game $Y$ always picks...
action 1, whatever the message $C_x$ observed. At the communication stage of the game, $Y$ produces a message $C_{y}$ such that the updated probability of playing against machine $Y$ given the prior $P_2$, on the basis of $C_y$, is arbitrarily close to 1. Given the Common Interest assumption on $G$ this argument shows that the Communication Lemma guarantees that there exists a machine $Y$ such that $\{X\}(C_{x^*};C_y)\in H_1$ (cf. Definition 2) and $\{Y\}(C_{y};C_{x^*})=1$.

Since $X^*$ has weight $(1-\xi)$ close to 1 in $P_1$, it then follows that the expected score of $Y$ against $P_1$ is arbitrarily close to the pareto-efficient level.

By optimality of $Y^*$ it then clearly must be the case that the expected score achieved by $Y^*$ against $P_1$ is also arbitrarily close to the pareto-efficient level.

A totally symmetric argument shows that the expected score achieved by $X^*$ against $P_2$ is arbitrarily close to the pareto-efficient level.

Given that both $X^*$ and $Y^*$ score close to the pareto-efficient level against $P_2$ and $P_1$ respectively, and that $\xi$ is close to zero the optimality statement b) of Theorem 1 follows trivially.

Before turning to the existence statement of Theorem 1, it is worth stating formally the following consequence of the Theorem.

**Corollary 1.** Consider a fixed extended common interest game satisfying Assumption 1 and 2. Consider, for a fixed $R\in N$, the correspondence associating to each $\xi\in (0;1)$ the set of pay-offs obtained by equilibrium players at the second stage of the game in any $(\xi;R)CTHE$ of $G^*$. Then there exists on $R\in N$ such that the limit of such correspondene as $\xi\to 0$ is well defined equals the unique pareto-efficient pair $(h_1;h_2)$.

Proof: The proof is obvious from Theorem 1 and hence omitted.

Several issues arise in connection with the existence statement a) of Theorem 1.
As noted above, in equilibrium, each player will be choosing an action which is an optimal response to the probability distribution over opponents, correctly updated on the basis of the observed communication string. This can be an extremely complex operation, and the question naturally arises as to whether the task is always (that is for any arbitrary admissible initial probability distribution over opponents) computable. In other words one may ask whether the requirement of optimality of players against a probability distribution over opponents places restrictions over the probability distribution itself in addition to those explicitly entailed by its admissibility. This would clearly be the case if for some admissible initial probability distribution the choice of an action which is optimal against a correctly updated probability distribution were not computable. It turns out to be the case that such action choice maps from pairs of messages into actions may indeed be computable or not according to the initial probability distribution.

Any worries about the existence of a \( \xi \)CTHE are resolved by the proof of a) of Theorem 1 contained in Appendix 2. The proof does rely on the use of a pair of probability distributions with a very restricted support, however.

Characterizing exactly which class of probability distributions may constitute equilibria, and in particular their 'maximal' support seems to be an extremely complex problem. It has resisted the author's best efforts for some time now, and does not seem to follow from any standard techniques in this area of mathematics.

7. Discussion of the Results and Games Which Are Not Common Interest

Aumann (1989) discusses at length the following form game with two players

\[
\begin{array}{cc}
A_2 & B_2 \\
A_1 & 9;9 & 0;8 \\
B_1 & 8;0 & 7;7 \\
\end{array}
\]
to which, for convenience, I shall refer as the Aumann game.

The Aumann game is a Common Interest game. It is an example, however, of a case in which it is not necessarily obvious that the 'good' equilibrium \((A_1; A_2)\) will necessarily be selected over the 'bad' equilibrium \((B_1; B_2)\). First of all, the \((A_1; A_2)\) equilibrium is riskier than \((B_1; B_2)\).\(^\text{15}\)

Secondly, Aumann (1989) argues that the equilibrium \((A_1; A_2)\) is not 'self-enforcing' in the following sense. Suppose that player 1 proposes to player 2 to 'agree' to play \((A_1; A_2)\). Presumably, player 1 then wants to convince player 2 to play his part of such equilibrium, namely \(A_2\). Does this attempt to convince player 2 to play \(A_2\) 'reveal' in anyway that player 1 intends to play \(A_1\)?

The answer is no, as it is in player 1's best interest that player 2 should play \(A_2\), regardless of whether he (player 1) intends to play \(A_1\) or \(B_1\).

Hence \((A_1; A_2)\) is not self-enforcing in the sense that a non-binding agreement to play it does not make it anymore likely to be played than it was before the agreement.

Thus, pre-play communication, even in the sense of Farrel (1988) may be in trouble here, as Farrel (1988) itself acknowledges.

Yet, in the using the model of pre-play communication proposed here the only equilibrium which survives is \((A_1; A_2)\) since the Aumann game does satisfy the definition of a Common Interest game. The question then naturally arises of how the result can be reconciled with the apparent intuition outlined above that communication is likely to be worthless in the Aumann game?

The solution to the riddle is simple and useful to spell out since it points out salient features of the present model. Pre-play communication as modeled here is capable of revealing (up to any pre-specified degree of precision) what exactly is the map from messages into actions which a player is going to use in the play stage of the game. In very

\(^{15}\)(\(B_1; B_2\)) is risk-dominant in the sense of Harsanyi and Selten (1988).
intuitive terms it's as if by uttering a (perhaps very 'complex') communication string, a player were able to reveal his brain's internal design as an information processing unit. Once this is done it is clear how the bad equilibrium can be ruled out.

Thus pre-play communication here is not restricted to 'proposals' to play one equilibrium or another. It is powerful enough to reveal (up to any desired level of precision) all that is in some sense relevant for the subsequent stage of the game. This, despite the obvious differences in the set-up, is not unlike what drives the proof of optimality in Aumann and Sorin (1989).

Of course, the fact that some machine intending to reveal itself as wanting to play the good equilibrium of a common interest game can reveal itself, hinges on the precise assumptions of the Communication Lemma.

Before turning to other matters let me point out one intuitively appealing property of the procedure yielding the revealing messages used in the Communication Lemma. The length of such 'smart' messages, it is easily verified, is decreasing in the parameter $\ell$ of the Lemma. In other words the length of such revealing messages is increasing in the required level of precision with which one is requiring the machine to reveal itself. Simple inspection of the pay-offs of the Aumann game and of the game discussed in the introduction shows that in order to induce a response of $A$ from a Bayesian updating opponent such level of precision needs to be higher in the Aumann game than in the pure coordination game of the introduction. Hence longer messages will be required in the Aumann game$^{16}$ in order to 'destroy' any equilibrium which does not yield pareto-efficient pay-offs.

These claims do not, of course, imply that messages observed in equilibrium in the Aumann game will be longer than in the pure coordination example. Moreover, these

$^{16}$These remarks were originally prompted by some comments which Andreu Mas–Colell offered me on an earlier version of the paper. He is, of course, not responsible for the remarks themselves.
remarks apply to the 'smart' messages constructed as in the particular proof of the Communication Lemma above. One cannot be sure that they reflect some necessary features of the 'revealing' messages described. The intuition behind them seems nevertheless to be comforting in the sense that it points to the fact that it would take 'more' communication to achieve in the pareto-efficient outcome in the Aumann game than in the pure coordination example above.

I conclude this section with some remarks on the application of the theory of communication developed above to more general games (repeated and not) than the common interest games.

Consider the two-stage set-up above when the assumption that $G$ is a common interest game is violated in a rather strong way. In particular consider the case where $G$ is a version of the 'battle of the sexes' like the following

$$
\begin{array}{cc}
A_2 & B_2 \\
A_1 & 2,3 \\
B_1 & 0,0
\end{array}
$$

There is no clear intuition why in this case communication should be of particular help in selecting an equilibrium of the game. One equilibrium of the above game favours players 1 and the other favours player 2. This is reflected in the fact that if one considers the concept $(\xi;R)CTHE$ defined above then it is possible to show that no equilibrium exists for the two-stage set-up considered above. This is, again, not unlike what happens to the Aumann and Sorin (1989) cooperation result when a game like the above is considered. An intuitive explanation of this fact runs as follows.

In equilibrium, both players have to compute best response maps between message

---

17The set $R$ needed to support this statement is "larger" than the set used to prove the Communication Lemma above and hence Theorem 1.
pairs and actions in $G$. From an extension of the Communication Lemma it follows that if player 1 is an equilibrium player, there exists a communication string for player 2 to which player 1 will respond by taking the action $A_1$. Consider now a player 2 which outputs such communication string and takes action $A_2$, whatever the communication string received from 1. The pay-off to such player is clearly at least $(1-\xi) \cdot 3$. Hence in equilibrium it must be that player 2 receives at least $(1-\xi) \cdot 3$. By a totally symmetric argument one can conclude that in equilibrium player 1 must receive at least $(1-\xi) \cdot 3$. For $\xi$ close to 1 this is clearly a contradiction and therefore the set of equilibrium players must be empty.

The above result is not disturbing when two stages of pre-play communication and play are considered. There is quite a clear intuition for the fact that the problem of equilibrium selection in a 'battle of the sexes' game may be made even more acute by simple costless unlimited pre-play communication. The resolution of the conflict is in the realm of bargaining theory and sufficient assumptions have to be made so that the bargaining process yields an equilibrium. Costless unlimited pre-play communication is simply not enough.\textsuperscript{18}

8. Applications to the Theory of Repeated Games with No Discounting and Extensions

In an infinitely repeated game with no discounting any finite portion of the history of play is irrelevant as far as the players' long-run pay-offs are concerned. One question that naturally arises from the investigation of pre-play communication presented in the previous sections of this paper is then the following. In the context of an infinitely repeated game with no discounting, is it possible that the players use the 'early stages' of the game as a means of communicating to each other information about their internal design and hence about their long-run strategy in the remainder of the game? Suppose moreover that the stage game of the infinitely repeated interaction is a common interest

\textsuperscript{18}Farrel and Gibbons (1988) analyze the role of communication in bargaining games.
game as defined above. Can such use of the early stages of the game as a communication device resolve the coordination problem between the two players in the infinitely repeated game?

The answers to the above two questions is positive. Using solution concepts virtually identical to the ones employed in Section 6 above, the only equilibrium long-run pay-offs possible in an infinitely repeated common interest game are the pareto-efficient pay-offs of the stage game.

The statement of this result is indeed very close to that of a theorem of Aumann and Sorin (1989). The assumptions on which it rests are quite different when looked at in detail, however. A formal statement of the result is beyond the scope of this already over-grown essay. A companion paper (Anderlini and Sabourian (1989)) formally states and proves the claim above as well as relating it to the Aumann and Scrin (1989) result.

9. Concluding Remarks and Related Literature

This paper has developed a model of pre-play communication in games. A pre-play communication stage modelled as proposed here can solve coordination (equilibrium selection) problems in strategic-form games. In a common interest game, the only equilibrium which survives after pre-play communication is allowed is the pareto-efficient equilibrium of the game.

Communication of various forms has been considered before in a variety of economic models. Perhaps the first instance in which communication has been formally explicitly in economics is the incentive-compatibility literature of the late 70's (Hurvicz (1972), Dasgupta, Maskin and Hammond (1979)). The main difference in the set-up which this literature analyzes and the present paper is that the incentive-compatibility literature addressed the issue of game-design (with the strategy space of the game being messages) while this paper analyzes the impact of allowing pre-play communication on a given class of games.
A large body of literature on signalling games is available by now. A complete survey is provided by Cho and Kreps (1987). The central difference between a signalling problem and pre-play communication is straightforward. In general, in the signalling literature, players have at some stage of the game a choice among different signals which they can send to other players. Typically, the same signal will have a different cost for different types of players, and this is what drives the ability of the signal to convey information about a player's type. The difficult issues in this literature typically arise when out-of-equilibrium behavior (Kreps (1989)) is considered. Very roughly speaking, this is because of the break-down in Bayes' rule when probability zero events are considered. Subtle issues on what kind of beliefs are induced for other players when one player sends an out-of-equilibrium signal need to be considered and resolved.

By contrast, the theory of pre-play communication developed in this paper considers 'signals' which have no cost to any players. Crawford and Sobel (1982) and Farrel in a number of papers (Farrel (1983), Farrel (1988), Farrel and Gibbons (1989)) consider the problem of information transmission when messages do not have cost to the players (cheap talk). In particular, Farrel (1983) and Farrel (1988) analyse the effects of costless messages sent by one player to another by means of an equilibrium refinement.

Clearly, the analysis of 'cheap talk' is much closer in spirit to the present paper than any other case in which communication has been considered in the literature. This paper has provided a consistent and, I believe, compelling reason for 'cheap talk' to matter. It reveals something about the internal design or general information processing abilities and habits of the players. Under quite general circumstances costless pre-play communication is able to convey enough information to the problem of coordination (equilibrium selection) in any common interest game.
Appendix 1

For the sake of completeness I report here some basic definitions relative to one possible specification of a Turing machine. The material presented is more than standard in the mathematical literature. I draw freely on Davis (1958). The reader is referred also to Cutland (1980) for an introduction to the field, or to Rogers (1967) for a comprehensive and rigorous treatment.

A Turing machine is a device comprising a 'head' and a tape. The tape is assumed to be infinite on both sides.

The head scans one square of the tape only at any one time. The head can execute three operations on the tape according to a list of instruction which I shall describe shortly. These are, to move to the next square to the right, to move to the next square to the left, and to replace the symbol on the square which is being scanned with another symbol.

To a machine is associated a finite 'alphabet' of symbols $S_0, S_1, \ldots, S_n$. The first symbol is interpreted as a 'blank'. (There is actually no loss of generality in considering machines which use two symbols only, plus the blank.)

A finite list of states is also associated with the machine. Denote these by $q_0, q_1, q_2, \ldots, q_m$. The states are not specified any further, and they are simply meant to capture the possibility that the machine be in different 'internal configurations', however physically defined.

Inside the machine's head resides a 'program' consisting of a finite set of quadruples (representing 'instructions') each belonging to one of the following three types

1) $q_i S_j S_h q_m$
2) $q_i S_j R q_m$
3) $q_i S_j L q_m$

The interpretation of these is respectively as follows. 1) If the machines is in state
q_i and the symbol on the scanned square is S_j, then replace S_j with S_h and change the state of the machine to q_m. 2) If the state of the machine is q_i and the symbol on the scanned square is S_j, then move the head one square to the right and change the state of the machine to q_m. 3) If the state of the machine is q_i and the symbol on the scanned square is S_j, then move the head one square to the left and change the state of the machine to q_m.

To avoid ambiguous instructions at most one quadruple beginning with the same pair q_i S_j is allowed in any 'program'.

The operation of the machine can now be easily described. The machine starts in some designated 'initial state' q_0. All but a finite portion of its tape is blank at the beginning. The initial expression of the tape is the machine's 'input'. The machine 'looks' for a quadruple beginning with the current state (initially q_0) and the symbol on the scanned square. If it finds such quadruple, the instruction which the quadruple contains is executed. Instructions are thus executed in sequence, progressively modifying the expression of the tape. If the machine is in state q_i and the scanned symbol S_j, and there is no quadruple in its program beginning with q_i S_j, the computation halts. The expression on the tape when the machine halts is the output of the computation.

Although obvious, it should be remarked at this point that the above definitions contain a considerable degree of arbitrariness. Many different, but equivalent, specifications of the machines are possible. I have reported one possible set here for the sake of the completeness only.
Appendix 2

Proof of Theorem 1

To show that the existence statement a) is correct consider a pair of identical probability distributions constructed as follows.

Let \( R \cap N \) be the set described in Step 3 of the proof of the Communication Lemma. Since \( R \) is the range of a computable function, it is recursively enumerable by standard results (cf., for instance Cutland (1980) Ch. 7). Let \( \{q_1; \ldots; q_n; \ldots\} \) be an enumeration of \( R \) without repetitions which satisfies the following

\[
q_1 = \min_{X \in R} |\{X\}| \downarrow
\]

Let then

\[
P(X) = \begin{cases} 
0 & \text{if } X \notin R \\
1 - \xi & \text{if } X = q_1 \\
\frac{1}{2^i} & \text{if } X = q_i \quad \text{with } i \neq 1
\end{cases}
\]

It is easy to see that the probability distribution described above satisfies b) of Definition 3 of admissible probability distribution since \( \sum_{i \in Q} P(i) = 0 \). To see that it satisfies a) notice that one can construct a Turing Machine \( M \) which given any pair \((n; \epsilon)\) starts by enumerating \( R \) and keeping track of the cumulative probability obtained by adding up \( 1 - \xi \) and \( \frac{1}{2^i} \) defined above. If at any point \( n \) shows up in the enumeration of \( R \) as, say, the \( i \)-th element, the machine halts and outputs \( \frac{1}{2^i} \) (or \( 1 - \xi \) if \( i = 1 \)). If before \( n \) has showed up in the enumeration of \( R \), the cumulative probability exceeds \( 1 - \epsilon \) (which it will eventually do if \( n \notin R \)) the machine halts and outputs zero.

It is now easy to see that the quadruple \((X^*; Y^*; P_1; P_2)\) where \( X^* = Y^* = q_1 \) and \( P_1 = P_2 = P \) constitutes an \((\xi; R)\)CTHE for the given extended game \( G^* \).

Indeed all machines in \( R \) output a constant 1 at the action-choosing stage of the
game. Hence by using b) (Indifference) of Assumption 1, the claim is obvious.

To show that the optimality claim b) of Theorem 1 is correct one can follow closely the intuitive explanation of the result given in Section 6 and just fill in the details.

Let R be the set defined in Step 3 of the proof of the Communication Lemma and let \([X^*;Y^*;P_1;P_2]\) be any \((\xi;R)CTHE\) of an extended common interest game \(G^*\) satisfying assumptions 1 and 2.

By a) (Dominance) of Assumption 1 and must clearly have

\[\{X^*(\cdot)\} \subseteq \{X^*(C_x;\cdot)\} \in S_1\]

for any \(C_y\) such that for some \(Y\) one has \(\{Y(\cdot)\} = C_y\) with \(Y \in \text{supp}(P_2)\), and a symmetric statement holds for \(Y^*\).

Consider now a machine \(Y\) as follows. \(\{Y(\cdot)\} = C_{y^{-}}\), \(\{Y(C_{y^{-}};C_x)\} = 1 \forall C_x\) and \(\text{Prob}(Y|C_{y^{-}}) \geq k\) where \(\text{Prob}(Y|C_{y^{-}})\) is the updated probability of machine \(Y\) given message \(C_{y^{-}}\). The Communication Lemma guarantees the existence of such \(Y\) for any \(0 < k < 1\).

Consider now the pay-off to the equilibrium machine \(X^*\) conditional on having received the message \(C_{y^{-}}\). If it is the case that \(\{X^*(C_x^*;C_{y^{-}})\} \in H_1\) (cf. Definition 2) such conditional pay-off is at least

\[kh_1 + (1-k)m_1\]

where\(^{19}\) \(m_1 < h_1\) is the minimum pay-off which player 1 can achieve in any outcome of the game \(G\). If it is the case that \(\{X^*(C_x^*;C_{y^{-}})\} \notin H_1\) such conditional pay-off is at most

\[ke_1 + (1-k)h_1\]

where \(e_1 < h_1\) is the second highest pay-off which player 1 can achieve in any outcome of \(G\). It follows then from simple algebra that for \(k\) and \(\xi\) sufficiently close to 1 it must be

\(^{19}\)If \(m_1 = h_1\) the given game \(G\) is trivial and the optimality claim of Theorem 1 is true for completely obvious reasons.
that $\{X^*\}(C_x^*;C_y^-) \in \mathcal{H}_1$, otherwise $X^*$ could not satisfy condition c) of the Definition 3 of equilibrium.

Since $G^*$ is an extended common interest game it follows easily that if one lets $\{X^*\}(C_x^*;C_y^-) = i$ then it must be that $\pi_2(i;1) = h_2$. Since $X^*$ has weight $1-\xi$ in $P_2$ it therefore follows that

$$
\sum_{i \in \mathbb{N}} P_1(i) \varphi_2(i;Y) \geq (1-\xi)h_2 + \xi m_2
$$

where $m_2$ is the minimum pay-off which player 2 can achieve in any outcome of $G$. By optimality of $Y^*$ it then also must be the case that

$$
\sum_{i \in \mathbb{N}} P_1(i) \varphi_2(i;Y^*) \geq (1-\xi)h_2 + \xi m_2
$$

and since $X^*$ has weight $1-\xi$ in $P_1$, for $\xi$ small enough this can only be the case if $\varphi_2(X^*;Y^*) = h_2$. By the fact that $G^*$ is an extended common interest game, this can only be the case if it is also true that $\varphi_1(X^*;Y^*) = h_1$, and this concludes the proof of the Theorem.
References


