The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge"

By ARIEL RUBINSTEIN*

The paper addresses a paradoxical game-theoretic example which is closely related to the coordinated attack problem. Two players have to play one of two possible coordination games. Only one of them receives information about the coordination game to be played. It is shown that the situation with "almost common knowledge" is very different from when the coordination game played is common knowledge.

A very basic assumption in all studies of game theory is that the game is "common knowledge." Following John Harsanyi (1967), situations without common knowledge are analyzed by a game with incomplete information. A player's information is characterized by his "type." Each player "knows" his own type and the prior distribution of the types is common knowledge. Jean-Francois Mertens and Samuel Zamir (1985) have shown that under quite general conditions one can find type spaces large enough to carry out Harsanyi's program and to transform a situation without common knowledge into a game with incomplete information in which the different types may have different states of knowledge. Harsanyi's method became the cornerstone of all modern analyses of strategic economic behavior in situations with asymmetric information (i.e., most of the theoretical Industrial Organization literature).

What does it mean that the game G is "common knowledge"? Following David Lewis (1969), Stephen Schiffer (1972), and Robert Aumann (1976), this concept has been studied thoroughly by relating it to concepts of "knowledge" and "probability" (for a recent presentation of this literature see Ken Binmore and Adam Brandenberger, 1987). Intuitively speaking it is common knowledge between two players 1 and 2 that the played game is G, if both know that the game is G, 1 knows that 2 knows that the game is G and 2 knows that 1 knows that the game is G, 1 knows that 2 knows that 1 knows that the game is G, and 2 knows that 1 knows that 2 knows that the game is G and so on and so on.

One of the main difficulties with this intuitive definition (and with the formal definitions which capture this perception) is that even "simple" sentences like "I do not know that you do not know that I know that you do not know that I know" are very difficult to visualize, thus making an assessment of their validity problematic. Therefore it would be interesting to understand whether a game-theoretic informational structure, referred to as "almost common knowledge," in which only a finite (but large) number of propositions of the type "1 knows that 2 knows that 1 knows...that the game is G" are true, is very different from the situation where the game G is common knowledge. In this short paper I will present a simple example of a situation with "almost common knowledge" of the game. The situation is analyzed using, as a tool, the idea of a game with incomplete information. It is shown that the game-theoretic "prediction" for the "almost common knowledge" situation is very different from the situation with common knowledge.

*Department of Economics, The Hebrew University, Jerusalem, and the London School of Economics, Houghton Street, London, England WO2A2AE. My thanks to Ken Binmore, Edi Dekel, John Geanakopolos, Avner Shaked, Chuck Wilson, Asher Wolinsky, and a referee of this journal for the very useful comments.

The example is similar to the "coordinated attack problem" which is well known in the distributed systems literature.[1] A description of the problem and a comparison with this paper analyzed appears in Section IV.

I. Coordination Through Electronic Mail

Two players, 1 and 2, are involved in a coordination problem. Each has to choose between two actions A and B. There are two possible states of nature, a and b. Each of the states is associated with a payoff matrix as follows:

The game G_a

	A	B
A	M, M	$0, -L$
B	$-L, 0$	$0, 0$

state a
probability $1-p$

The game G_b

	A	B
A	$0, 0$	$0, -L$
B	$-L, 0$	M, M

state b
probability p

In the state of nature a(b) the players get a positive payoff, M, if both choose the action $A(B)$. If they choose the same action but it is the "wrong" one they get 0. If they fail to coordinate, then the player who played B gets $-L$, where $L > M$. Thus, it is dangerous for a player to play B unless he is confident enough that his partner is going to play B as well. The state a is the more likely event; b appears with a priori probability of $p < 1/2$.

The information about the state of nature is known initially only to player 1. Without transferring the information, the players cannot achieve an expected payoff higher than $(1-p)M$. If the information could become common knowledge they would be able to achieve the payoff M. However, imagine that the two players are located at two different sites and they communicate only by electronic mail signals. Due to "technical difficulties" there is a "small" probability $\varepsilon > 0$, that the message does not arrive at its destination. At the risk of creating discord, the electronic mail network is set up to send a confirmation *automatically* if any message is received, including not only the confirmation of the initial message but a confirmation of the confirmation; and so on. To be more precise, it is assumed that, when player 1 gets the information that the state of nature is b, his computer automatically sends a message (a blip) to player 2 and then player 2's computer confirms the message and then player 1's computer confirms the confirmation and so on. If a message does not arrive, then the communication stops. No message is sent if the state of nature is a. At the end of the communication phase the screen displays to the player the number of messages his machine has sent. Let T_i be a variable for the number of messages i's computer sent (the number on i's screen).

Notice that sending the messages is not a strategic decision by the players. It is an automatic device carried out by the computers. The designer of the system sets up the communication network between the players and they can only choose between A and B after the communication phase has ended.

If the two machines exchange an infinite number of messages, then we may say that the two players have common knowledge that the game is G_b. However, since only a finite number of messages are transferred, the players never have common knowledge that the game they play is G_b.

In choosing between A and B after the end of the communication phase, player 1 (and similarly player 2) faces uncertainty: given that he sent T_1 messages he does not know whether player 2 did not get the T_1th message, or whether player 2 got the T_1th message, but the T_1th confirmation has been lost. Any number on the screen corresponds to a state of knowledge not only about the state of nature but also about the other

[1] I should like to thank John Geanakopolos for referring me to the "coordinated attack problem."

player's knowledge. For example if player 1's computer sent two messages it means that:

$K_1(b)-$ 1 knows that b

$K_1K_2(b)-$ 1 knows that 2 knows that b (by the fact that he has received confirmation of his first message). However, it is not true that $K_1K_2K_1K_2(b)-1$ does not know that 2 knows that 1 knows that 2 knows that b. Player 1 assigns probability $z=\varepsilon/[\varepsilon+(1-\varepsilon)\varepsilon]$ to $T_2=1$ and $(1-z)$ to $T_2=2$. Therefore player 1 believes that:

with probability $1-z$ $K_2K_1K_2(b)$ and
with probability z that
2 believes that
with probability $1-z$ $K_1K_2(b)$ and
with probability z that
1 believes that
with probability z 2 believes that with probability $(1-p)/(1-p\varepsilon)$, a, and with probability $(1-z)$, 2 knows that b.

The statements of higher order are even more complicated. Notice that, under the model's assumption that player 1 gets accurate information about the state of nature, "x" and "$K_1(x)$" are two equivalent statements.

Similarly, any number on a player's screen at the end of the communication stage corresponds to a sequence of propositions describing the player's knowledge about the state of nature, about his opponent's belief about the state of nature, about his opponents's belief about his belief about the opponent's belief about the state of nature and so on. The larger is T_1, the more statements of the type $K_1K_2K_1\dots K_1K_2(b)$ are true, and the closer we are to the common knowledge situation.

How could we analyze the situation when the two players have the numbers T_1 and T_2 on their screens? To calculate his best action when $T_1=2$, for example, player 1 may have to form beliefs about player 2's actions when T_2 is 1 or 2. The optimality of these would have to be checked given player 1's behavior when $T_1=1$, 2, or 3, and so on. Harsanyi's method suggests that we analyze a situation given any pair of numbers on the screens, as part of a game of incomplete information which I will refer to as "the electronic mail game" (to distinguish from the coordination games). The set of types in the electronic mail game is the set of natural numbers and the distribution of the pairs of types is deduced from the electronic mail technology (namely, the probability of (T_1,T_2) being respectively $(0,0)$, $(n+1,n)$, and $(n+1,n+1)$ are $1-p$, $p\varepsilon(1-\varepsilon)^{2n}$, and $p\varepsilon(1-\varepsilon)^{2n+1}$, respectively). Define player i's strategy in the electronic mail game, S_i, to be a function from the set of natural numbers $0,1,2,\dots$ into the action space $\{A,B\}$. Then $S_i(t)$ is interpreted as i's action if his machine sent t messages.

II. The Analysis of the Electronic Mail Game

PROPOSITION 1: *There is only one Nash equilibrium in which player* 1 *plays A in the state of nature a. In this equilibrium the players play A independently of the number of messages sent.*

PROOF:

Let (S_1,S_2) be a Nash equilibrium such that $S_1(0)=A$. We will prove by induction that $S_1(t)=S_2(t)=A$ for all t. If $T_2=0$ then player 2 did not get a message. He knows that it might be because player 1 did not send him a message (this could occur with probability $1-p$) or because a message was sent but did not arrive (this happens with probability $p\varepsilon$). In the first case, player 1 plays A $(S_1(0)=A)$. If player 2 plays A, then, whatever $S_1(1)$ is, player 2's expected payoff is at least; $[(1-p)M+p\varepsilon 0]/[(1-p)+p\varepsilon]$ and if he plays B he gets at most $[-L(1-p)+p\varepsilon M]/[(1-p)+p\varepsilon]$. Therefore it is strictly optimal for 2 to play A, that is $S_2(0)=A$.

Assume now that we have shown that, for all $T_i<t$, players 1 and 2 play A in equilibrium. Assume $T_1=t$. Player 1 is uncertain whether $T_2=t$ (in the case where player 2 received the tth message but 2's tth message was lost) or $T_2=t-1$ (in the case where 2 did not receive the tth message). Given that he did not receive confirmation of his tth message, his conditional probability that $T_2=t-1$ is $z=\varepsilon/[\varepsilon+(1-\varepsilon)\varepsilon]>1/2$. Thus it is more likely that player 1's last message did not arrive than that player 2 got the message. (This fact is the key to our argument). By the inductive assumption, player 1 assesses that, if $T_2=t-1$, player 2 will play A.

If player 1 chooses B, player 1's expected payoff is at most $z(-L)+(1-z)M$. If he chooses A, then his utility is 0. Given that $L > M$ and since $z > 1/2$, his only best action must be A. Thus $S_1(t) = A$. Similarly we show that $S_2(t) = A$.

Thus even if both players know that the actual played coordination game is G_b and even if the noise in the network (the probability ε) is arbitrarily small, the players ignore the information and play A. The best expected payoff the players can obtain in any equilibrium is still $(1-p)M$, just as if no electronic mail system existed!

Remark 1: Consider the mechanism described above but with the addition that, after a commonly known fixed finite number of messages, T, the system stops, if it has not stopped before. If $\varepsilon(-L)+(1-\varepsilon)M > 0$ then there is an equilibrium in which each player plays B if he receives confirmations of all his messages. The expected payoffs of this equilibrium, conditional on the state b are: $(1-\varepsilon)^T M$ to the last player who is supposed to get a message and $(1-\varepsilon)^{T-1}$ $[\varepsilon(-L)+(1-\varepsilon)M]$ to the other player.

Notice that these two numbers are decreasing in T and therefore the only "efficient" schemes might be those with $T = 1$ and $T = 2$. The mechanism with $T = 1$ is a better scheme for player 2 and $T = 2$ is a better scheme for player 1. If the communication channel is so noisy that $\varepsilon(-L)+(1-\varepsilon)M < 0$ then the efficient equilibrium is the one where the messages are ignored (the argument is similar to the proof of the proposition).

III. The Coordinated Attack Problem

As was mentioned in the introduction the electronic mail game is strongly related to the coordinated attack problem known in the distributed systems folklore. The problem as described in Joseph Halpern (1986, p. 10) is the following:

> Two divisions of an army are camped on two hilltops overlooking a common valley. In the valley awaits the enemy. It is clear that if both divisions attack the enemy simultaneously they will win a battle, whereas if only one division attacks it will be defeated. The divisions do not initially have plans for launching an attack on the enemy, and the commanding general of the first division wishes to coordinate a simultaneous attack (at some time the next day). Neither general will decide to attack unless he is sure that the other will attack with him. The generals can only communicate by means of a messenger. Normally, it takes the messenger one hour to get from one encampment to the other. However, it is possible that he will get lost in the dark or, worst yet, be captured by the enemy. Fortunately, on this particular night, everything goes smoothly. How long it will take them to coordinate an attack?
>
> Suppose the messenger sent by general 1 makes it to general 2 with a message saying "Let's attack at dawn." Will general 2 attack? Of course not, since general 1 does not know he got the message, and thus may not attack. So general 2 sends the messenger back with an acknowledgment. Suppose the messenger makes it. Will general 1 attack? No, because now general 2 does not know he got the message, so he thinks general 1 may think that he (general 2) didn't get the original message, and thus not attack. So general 1 sends the messenger back with an acknowledgment. But of course, this is not enough either. I will leave it to the reader to convince himself that no amount of acknowledgments sent back and forth ever guarantee agreement. Note that this is true if the messenger succeeds in delivering the message every time.

The question asked in the quoted paragraph is whether there is a common knowledge of the attack plan at the end of the information transmission stage. The above "communication protocol" cannot result in the players' having common knowledge about the time of the attack. However, the fact that the generals could not achieve common knowledge does not exclude the possibility that with positive probability they will both attack at dawn. This sounds plausible especially if the probability of a messenger failure is very small.

For this reason it is interesting to analyze the problem in the explicit form of a game. This is the minor contribution of this paper. In order to address the problem as a game, we need to add more structure to the problem and, in particular, we have to specify the probability conditions under which general 1 decides to initiate an attack at dawn. In terms of Section II, state b can be interpreted as the conditions which make an attack at dawn likely to succeed, while state a is the "status quo" state. Action B is "attack at dawn" and action A is the default action. The payoffs in Section I represent an assumption that, in case of an uncoordinated attack, only the general who attacks loses. If, alternatively, we assume that both generals' utilities are $-L$ if an uncoordinated attack is launched, then there is an equilibrium in which general 2 attacks as soon as he gets at least one message, provided that ε is small enough (less than $M/(M+L)$). This last fact emphasizes the importance of addressing the problem within a game-theoretic framework.

IV. Final Comments

A. *Is "Almost Common Knowledge" Close to "Common Knowledge"?*

It should be emphasized that the game about which knowledge is being hypothesized in the above is the coordination game and not the electronic mail game. One is concerned with what the two players do or do not know about the payoffs in the coordination game and with what the players do or do not know about the knowledge of their opponent. The story of the interchange of messages by electronic mail is intended only to provide a precise, albeit rather special, model of how knowledge on those questions may come to be shared by the players.

The main message of this paper is that players' strategic behavior under "almost common knowledge" may be very different from that under common knowledge. To emphasize, by "almost common knowledge" I refer to the case when the numbers on the screens are "very large." Then a "very large" number of statements of the type "player i knows that player j knows that...the coordination game is G_b" are correct. Still, the players will not coordinate on the action B whereas they are able to coordinate on the action B if it is common knowledge that the coordination game is G_b.

B. *The Electronic Mail Game as a Perturbed Game*

Selten's perfection definitions and the Kreps-Milgrom-Roberts-Wilson (1982) approach used small perturbations in a game in order to select an equilibrium in a game with multiplicity of equilibria and to create new equilibria in the absence of a reasonable equilibrium. If we think of ε as being small then the noisy electronic mail game is a perturbation of a non-noisy electronic mail game (the electronic mail game with $\varepsilon = 0$). The non-noisy game has several equilibria (since it is just a coordination problem) however the perturbation unfortunately excludes the more reasonable equilibria. Notice that the difference between a game and a perturbed version of the game has already been demonstrated many times in the past and I feel less paradoxical about this as compared to the paradoxical features of the present example.

C. *The Paradoxical Aspect of the Example*

What would *you* do if the number on your screen is 17? It is hard to imagine that when L is slightly above M and ε is small a player will not play B. The sharp contrast between our intuition and the game-theoretic analysis is what makes this example paradoxical.

The example joins a long list of games such as the finitely repeated Prisoner's Dilemma, the chain store paradox, and Rosenthal's game, in which it seems that the source of the discrepancy is rooted in the fact that in our formal analysis we use mathematical induction while human beings do not use mathematical induction when reasoning. Systematic explanation of our intuition that we will play B when the number on our screen is 17 (ignoring the inductive consideration contained within Proposition 1's proof) is definitely a most intriguing question.

D. *Games with Incomplete Information*

As mentioned earlier the situation without common knowledge is analyzed, à la Harsanyi, as a game with incomplete information. Notice that almost all the non-abstract literature uses the distinction between types to reflect differences in knowledge about payoff-relevant items. The current example is exceptional in that it demonstrates a family of natural game-theoretic scenarios in which the main difference between the types is in their knowledge about other players' knowledge.

E. *A Formal Presentation of the Type Spaces and the Information Partitions*[2]

Those readers who are familiar with Auman (1976), may found it helpful to have a formal statement of the type spaces and the information partitions in the electronic mail game. The type spaces of the two players are the sets which include $(a,0,0)$ and the triples (b,t,t') where $t>0$ and t' is either t or $t-1$. Array the set in the following order:

$$(a,0,0)(b,1,0)(b,1,1)(b,2,1)$$
$$(b,2,2)(b,3,2)(b,3,3)\dots.$$

Player 1's information partition is:

$$\{(a,0,0)\}\{(b,1,0),(b,1,1)\}$$
$$\{(b,2,1),(b,2,2)\}\{(b,3,2),(b,3,3)\}\dots$$

and player 2's information partition is:

$$\{(a,0,0),(b,1,0)\}\{(b,1,1),(b,2,1)\}$$
$$\{(b,2,2),(b,3,2)\}\{(b,3,3)\dots.$$

The meet of the two partitions is the trivial partition which contains only the entire type space. Thus the event "b" consists of the entire type space with the exception of $(a,0,0)$ and is never common knowledge. Notice that when $\varepsilon=0$, the feasible states are just $(a,0,0)$ and (b,∞,∞).

F. *Topology*

Two of the readers of the first version of this paper, both experts in the literature on common knowledge, raised objections to the way I use the term "almost common knowledge." They based their objection on the fact that when $\varepsilon\to 0$ the information partitions of the players do not converge to the information partitions when $\varepsilon=0$ (see this section, Part E). A referee suggested several topologies in which alternative concepts of "almost common knowledge" make sense.

Before reacting to this criticism let me emphasize again that I use the term "almost common knowledge" not for stating that the electronic mail game with ε close to 0 is almost the game with $\varepsilon=0$. What I am saying is that the situation with a high T_1 is close to the common knowledge situation. However, I would like to use this objection to spell out my opinion on the role that topology (in common with most other fields of "fancy mathematics") should play in economic theory. Topology should be used in one of two ways: (1) as a technical tool for phrasing a meta-claim about a family of models, or (2) as a substantial tool to formalize natural intuitions about "closeness." I envisage the high T_1 situation as being close to the common knowledge situation in the sense of (2). This may be unhelpful from a technical point of view and a conclusion from the example is indeed that the Nash equilibrium is not upper hemicontinuous in this convergence. However, lack of technical usefulness is not an argument against the perception that a situation with high T_1 is close to a situation with common knowledge. Obviously other definitions of convergence may be useful not only as technical methods but also for expressing other intuitions of closeness.

REFERENCES

Aumann, Robert J., "Agreeing to Disagree," *Annals of Statistics*, 1976, *4*, 1236–239.

[2] In this section I am closely following a referee's suggestion.

Binmore, Kenneth and Brandenberger, Adam, "Common Knowledge and Game Theory," Discussion Paper No. TE/88/167, STICERD, London School of Economics, 1987.

Halpern, Joseph Y., "Reasoning about Knowledge: An Overview," in *Reasoning about Knowledge*, J. Y. Halpern, ed., Morgan Kaufmann, 1986, 1–18.

Harsanyi, J. C., "Games with Incomplete Information Played by Bayesean Players," Parts I, II, III, *Management Science*, 1967, *14*, 159–82, 320–34, 486–502.

Kreps, D., Milgrom, P., Roberts, J. and Wilson, R., "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory*, August 1982, *27*, 245–52.

Lewis, David, *Convention, A Philosophical Study*, Cambridge: Harvard University Press, 1969.

Mertens, Jean-Francois and Zamir, Samuel, "Foundation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 1985, *14*, 1–29.

Schiffer, Stephen R., *Meaning*, Oxford: Oxford University Press, 1972.