Journal of Logic, Language, and Information **6:** 171–190, 1997.

Approximate Common Knowledge and Co-ordination: Recent Lessons from Game Theory*

STEPHEN MORRIS
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.
E-mail: smorris@econ.sas.upenn.edu

HYUN SONG SHIN
Nuffield College, Oxford OX1 1NF, U.K.
E-mail: hyun.shin@economics.ox.ac.uk

(Received 20 June 1996; in final form 20 August 1996)

Abstract. The importance of the notion of common knowledge in sustaining cooperative outcomes in strategic situations is well appreciated. However, the systematic analysis of the extent to which small departures from common knowledge affect equilibrium in games has only recently been attempted.

We review the main themes in this literature, in particular, the notion of common p-belief. We outline both the analytical issues raised, and the potential applicability of such ideas to game theory, computer science and the philosophy of language.

Key words: Common knowledge, common belief, coordination, game theory, protocols, language

1. Introduction

Philosophers, computer scientists and game theorists are all interested in the problem of co-ordination. When and how can the behaviour of a collection of agents be successfully coordinated? In all three fields, researchers have found that in answering these questions, it is useful to introduce formal ways of discussing what agents know. In particular, something is said to be *common knowledge* among a group of agents if all know it, all know that all know it, and so on. Common knowledge turns out to be necessary for perfect co-ordination. This conclusion is true whether the "agents" be processors in a distributed system which must jointly execute a complex computer protocol (Fagin et al., 1995); or whether "agents" are people who must agree on how to use language (Lewis, 1969).

Given the importance of the implications of whether or not common knowledge exists, it is worth starting with the basic question of how a set of individuals might achieve common knowledge of a particular state of affairs. In some limited contexts, the attainment of common knowledge rests on firm foundations. For instance, if

* We are grateful to Johan van Benthem, Joe Halpern and Yoram Moses for valuable comments on an earlier draft. This paper grew out of a talk by the first author at the March 1996 conference on Theoretical Aspects of Rationality and Knowledge (TARK VI) in Renesse, the Netherlands. He is grateful to Yoav Shoham for organizing a session on common knowledge in game theory and computer science.

players are logically competent – in the sense that they are capable of making logical deductions – and share the same state space, and hence share a common "theory" of the world, then common knowledge of their environment, as described by those statements which are true in every state of he world, follows in a straightforward way from the following type of reasoning. A statement ϕ which is true at every state is known by all individuals at every state. Hence, the statement "everyone knows ϕ" is true at every state, implying that everyone knows *this* statement at every state. Proceeding in this way, any statement of the form "everyone knows that everyone knows that ... everyone knows ϕ" is true at every state, so that ϕ is common knowledge.* This type of reasoning is also involved in the so-called "fixed point" characterization of common knowledge (Clark and Marshall, 1981; Barwise, 1988).

However, such a view of knowledge may be of limited use in many fields in which beliefs are derived from empirical sources, or at least, sources which are not infallible. We may often have good evidence that a proposition is true, but usually there cannot be an iron-clad guarantee. If coordination requires common knowledge concerning such empirical facts about the world, then common knowledge seems to be an excessively strong requirement that will rarely be achieved in practice. Indeed, the classic coordinated attack problem (discussed below) shows that if communication between agents is not perfectly reliable, common knowledge cannot be achieved. Of course, this does not mean that distributed systems do not achieve any co-ordination and languages do not exploit any common understandings. It does mean that often the most we can hope for is imperfect co-ordination. Thus one natural question for researchers in all the above-mentioned fields to ask is: what form of approximate common knowledge is sufficient to achieve a reasonable level of co-ordination?

As a benchmark, consider one answer to this question provided by game theory. Say that something is p-believed if everyone believes it with probability at least p. It is common p-belief if it is p-believed, it is p-believed that it is p-believed, and so on. Common p-belief can also be given a fixed point characterization. Now consider a game between a finite number of players, each with a finite choice of actions. Game theoretic results surveyed below show that if co-ordination is achievable in such a game *when there is common knowledge of the structure of the game*, then approximate co-ordination is also achievable if there is only common p-belief, for some p sufficiently close to one. Common p-belief is not only sufficient but also necessary for such approximate co-ordination.

* This reasoning has a logical counterpart in the "rule of necessitation" in modal logic, which states that if ϕ is a theorem of the logic, then so is the statement "individual i knows ϕ". By iteration, any theorem ϕ of that logic is common knowledge. See Hughes and Cresswell (1968) or Chellas (1980). The formalization of knowledge in terms of an epistemic logic is usually attributed to Hintikka (1962), who built on developments in modal logic (for instance, Kripke, 1959). Epistemic logics in multi-person contexts have been discussed by Halpern and Moses (1990, 1992), Aumann (1992), Shin (1993), Bonanno (1996) and Lismont and Mongin (1995).

Similar questions can and have been asked in other literatures. In this paper, we wish to highlight two issues which are raised by the game theoretic analysis. First, there is an important difference between *strategic* problems, where agents act in their own interests, and *non-strategic* problems, where agents can be relied upon to follow rules given to them. We will argue that – at least from a Bayesian perspective – approximate co-ordination is relatively easy to achieve in non-strategic problems. In particular, common p-belief is not required. This suggests that it will typically be easier to achieve co-ordination in protocols in computer science (where processors are typically not strategic) than it will be to achieve co-ordination in economic problems (where individuals almost invariably have some differences in objectives).

Second, co-ordination becomes much harder to achieve when there are many possible outcomes of co-ordination. Consider a game where each agent must choose one of a continuum of actions. Each agent wants his action to be as close as possible to the actions of the others, i.e., they want to co-ordinate on some *convention*. However, some conventions may be more socially desirable than others. If there is common knowledge of the social desirability of alternative conventions, then there is an equilibrium where each agent always chooses the socially optimal convention. On the other hand, we will show that even if there is common p-belief, with p close to 1, of the social desirability, the only equilibrium is a "simple convention" where the same action is taken whatever the social desirability. This is the case however close p is to 1. Philosophers' primary concern in this area has been with conventions such as *languages*. Our analysis suggests that co-ordination will be especially hard to achieve in such settings.

Two caveats are required. In providing a unified but informal discussion of a set of issues from the game theory literature, we do not attempt to provide a comprehensive survey. In particular, we provide our own interpretation of the game theory results without explaining the exact connection to the original results; and we make no attempt to cover related work in other fields. Beyond the few articles mentioned in the text, relevant sources would include Geanakoplos (1994) and Dekel and Gul (1996) for economics and game theory; Fagin et al. (1995) for computer science; and van Benthem and ter Meulen (1996) for the philosophy of conventions.

In this paper we attempt to describe a large number of game theoretic results using examples. Precision is valued in game theory and game theorists have very specific ideas in mind when they discuss games, strategies, rationality, equilibrium and other such concepts. It is unfortunately beyond the scope of this paper to provide an introduction to game theory. Therefore it should be understood that all the examples, arguments and results reported are intended to illustrate examples, arguments and results which can be made precise in the language of game theory. The textbook of Osborne and Rubinstein (1994) provides a meticulous introduction to game theory; Chapter 5 of their book describes how economists model knowledge and common knowledge, and their applications to game theory.

2. The Co-ordinated Attack Problem: A Common Knowledge Paradox

The following is a slightly altered version of the co-ordinated attack problem described by Halpern and Moses (1990).

> Two divisions of an army, each commanded by a general, are camped on two hilltops overlooking a valley. In the valley awaits the enemy. The commanding general of the first division has received a highly accurate intelligence report informing him of the state of readiness of the enemy. It is clear that if the enemy is unprepared and both divisions attack the enemy simultaneously at dawn, they will win the battle, while if *either* the enemy is prepared *or* only one division attacks it will be defeated. If the first division general is informed that the enemy is unprepared, he will want to coordinate a simultaneous attack. But the generals can communicate only by means of messengers and, unfortunately, it is possible that a messenger will get lost or, worse yet, be captured by the enemy.

The crucial feature of this story is that it is necessary for at least one message to be delivered from the first division general to the second division general in order for an attack to occur. In this version, the reason is that the second division general must be informed that the enemy is unprepared. In the Halpern and Moses version, the reason is that they do not have a prior agreement on the *time* to attack. For the arguments in this section, this distinction is unimportant. But in later sections, it is useful to appeal to the former more concrete reason why at least one message is required.

Is co-ordinated attack possible in this environment? Let us first define co-ordinated attack. We would like to design both a *communication protocol*, specifying which general sends which message to the other in which circumstances, and an *action protocol* specifying which general attacks in which circumstances. These two protocols achieve co-ordinated attack if (1) it is *never* the case that an attack occurs when the enemy is prepared, (2) it is *never* the case that one division attacks alone, and (3) both divisions sometimes successfully co-ordinate an attack. But remarkably, it has been shown that:

- Co-ordinated attack is not possible under *any* communication protocol with unreliable communication.

We can illustrate why this is the case by considering the following "naive communication protocol". If the first division general hears that the enemy is unprepared, he sends a message to the second division general with the instruction "attack". If that first message arrives, the second division general sends a messenger with a confirmation that the first message was safely received. If the confirmation is delivered without mishap, the first division general sends another message to the second division general informing him of this fact. And so on.

Now if the second division general never receives an instruction to attack, he cannot attack under any co-ordinated attack action protocol: it is possible that the first division general knows that the enemy is actually prepared and has not sent any messenger. Thus if the first division general never receives any confirmation of an instruction to attack, he will not attack: he thinks it possible that the second division general never received his message, and thus is not attacking. We can proceed in this manner and verify that no matter how many messages are successfully delivered, co-ordinated attack cannot occur.

The naive communication protocol was just an example of a communication system. The result is quite general: as long as all messages may be lost with some probability, and as long as perfect co-ordination is required, there is no co-ordinated attack. But the naive communication protocol illustrates the apparent fragility of common knowledge; if n messages have been successfully sent and received, one might say that they have approximate common knowledge of the fact that the enemy is unprepared. Yet their behaviour remains unco-ordinated.

3. The *Probabilistic* Co-ordinated Attack Problem: A Paradox Resolved

From a decision theoretic point of view, the above analysis seems intriguing but suffers from the flaw that *perfect* co-ordination is required. If we could design a protocol where co-ordinated attack almost always occurred when the enemy was unprepared and it was very unlikely that only one division attacked, decision theorists would be satisfied. In a Bayesian view of the world, bad outcomes do not matter if they occur at sufficiently small probability events. In order to analyse this question, we will present a probabilistic version of the co-ordinated attack problem.*

Suppose now that with probability $\delta > 0$ the enemy is prepared, while with probability $1 - \delta$ the enemy is unprepared. Recall that the first division general knows which of these two contingencies is true, while the second division general does not. We will consider the same communication protocol outlined above, except now we assume that each messenger gets lost with independent probability $\varepsilon > 0$. We will be focusing on the case where ε is small and in particular is smaller than δ. We will make the unrealistic assumption that there is no upper bound on the number of messages that might be successfully sent (although with probability one a message will be lost eventually). Now the environment can be described by a *state space* as in Table I.

Thus state (n, m) refers to a situation where the first division general has sent n messages (but does not know whether his last message arrived or not), while the second division general has sent m messages (and similarly does not know whether his last message arrived or not). Note that the infinite sum of the probabilities is naturally equal to one.

* Halpern and Tuttle (1993) discuss probabilistic versions of the co-ordinated attack problem.

Table I. The naive communication protocal state space.

State	Enemy's preparedness	Probability
$(0,0)$	Prepared	δ
$(1,0)$	Unprepared	$(1-\delta)\varepsilon$
$(1,1)$	Unprepared	$(1-\delta)(1-\varepsilon)\varepsilon$
$(2,1)$	Unprepared	$(1-\delta)(1-\varepsilon)^2\varepsilon$
...	...	...

To complete our analysis, we must specify payoffs for the different outcomes. Suppose that a successful attack has a payoff of 1 for the generals, while an attack that is unsuccessful (either because the enemy is prepared or only one division attacks) has a payoff of $-M$, where M is a very large number. Both generals not attacking has a payoff of 0. These payoffs capture the qualitative feature underlying the non-probabilistic co-ordinated attack problem that the cost of an unco-ordinated attack is much larger that the benefit from a successful attack (otherwise, why would we be interested in perfect co-ordination?).

Let us fix the naive communication protocol; say that an action protocol for the generals is *optimal*, given the communication protocol, if it maximizes the generals' expected payoff.

- If the communication system is sufficiently reliable, then the optimal action protocol has co-ordinated attack almost always occurring when the enemy is unprepared.

The optimal action protocol, given the naive communication protocol and ε sufficiently small, has the first division general attacking whenever the enemy is unprepared (even if he has not received any message confirmation from the second division general); while the second division general attacks even if he has received only one message from the first division general.* Thus co-ordinated attack occurs with probability $(1-\delta)(1-\varepsilon)$. In fact, this can be achieved under any communication protocol where at least one message is sent to the second division general informing him that the enemy is unprepared.

Thus continuity is restored and the paradox is resolved, at least for a Bayesian decision maker who is interested maximizing expected payoffs. If the communi-

* To illustrate this claim, consider three action protocols. (1) If the first division general attacks whenever the enemy is unprepared and the second division general always attacks, the expected payoff is $\delta(-M)+(1-\delta)(1)=1-(M+1)\delta$. (2) If the first division general attacks whenever the enemy is unprepared, and the second division general attacks if he has received at least one message, the expected payoff is $(1-\delta)\varepsilon(-M)+(1-\delta)(1-\varepsilon)(1)=(1-\delta)(1-(M+1)\varepsilon)$. (3) If each general attacks if he has received at least one message, the expected payoff is $(1-\delta)(1-\varepsilon)(1-(M+1)\varepsilon)$. Protocol (2) is better than protocol (3) if $\varepsilon<(1/M+1)$. Protocol (2) is better than protocol (1) if $\varepsilon<(\delta/1-\delta)(M/M+1)$. Thus protocol (2) is better than protocols (1) and (3) for all ε sufficiently small. Similar arguments show that protocol (2) is better than *all* alternative protocols.

cation system is sufficiently reliable (i.e., ε is sufficiently small), there exists a protocol which gives an expected payoff that is arbitrarily close to what it would be with perfect communication.

3.1. Implications for Computer Science

If it not possible to achieve perfect co-ordination in a computer protocol, it seems reasonable that the objective should be to maximize the probability of successful co-ordination, while minimizing the probability of an unsuccessful attempt to co-ordinate. The "optimal protocol" for the generals does just that in this example. The analysis of this section showed that if the objective is only to achieve "co-ordination with high probability" and agents are not strategic, then it is enough that the communication system is usually accurate.

Halpern and Tuttle (1993) discuss various alternative notions of "approximate common knowledge" that are sufficient to achieve various notions of "approximate co-ordination".* Our point is that – assuming computer science applications are non-strategic – co-ordination *with high probability* is achievable without strong notions of approximate common knowledge.** However, we will see that strong approximate common knowledge notions (with fixed point characterizations) *are* necessary for co-ordination with high probability in *strategic* environments.

4. The *Strategic* Co-ordinated Attack Problem: Paradox Redux

A probabilistic perspective seems to have made the co-ordinated attack problem go away. A remaining difficulty is that the optimal action protocol turns out to be sensitive to *strategic concerns*. The optimal action protocol described is optimal as long as the generals can be relied on to choose to follow it. Perhaps their battle orders instruct them to follow that protocol and generals always follow their battle orders. But suppose that the first division general knows that the enemy is unprepared, sends a message to the second division general, but receives no confirmation. He then believes with probability $1/(2-\varepsilon)$ that his own message never arrived, and thus that the second division general will not attack. For all ε, this probability is more than $1/2$. Perhaps he would be tempted not to commit his division to the battle in these circumstances. Anticipating this possibility, the second division general may hesitate to attack if he has not received a re-confirmation from the first division general. The unraveling argument may start all over again.

* See also Chapter 11 of Fagin et al. (1995).

** There are at least two reasons why computer scientists are nonetheless interested in the strong notions of approximate common knowledge. First, the objective of protocol design is typically not maximization of expected payoffs, but rather uniform lower bounds on performance. Second, computer scientists are concerned with the actual construction of protocols in environments where it is typically not possible to compute *optimal* protocols.

Table II. Payoffs if the enemy is prepared.

	Attack	Don't attack
Attack	$-M, -M$	$-M, 0$
Don't attack	$0, -M$	$0, 0$

Table III. Payoffs if the enemy is unprepared.

	Attack	Don't attack
Attack	$1, 1$	$-M, 0$
Don't attack	$0, -M$	$0, 0$

To understand this argument formally, we must treat the situation as an "incomplete information game" played between the two generals.* It is a game because each general, in seeking to maximize his expected payoff, must take into account the action of the other general. There is incomplete information because under the naive communication protocol, each general does not know the exact information held by the other general.

We have already described the state space capturing the relevant uncertainty, so now we must specify the generals' payoffs. Suppose that each general gets a payoff of 0 if his division does not attack. If his division participates in a successful attack, he gets a payoff of 1; if his division participates in an unsuccessful attack (either because the enemy is prepared or the other division does not attack), he gets a payoff of $-M$. Thus if the enemy is in fact prepared (i.e., the state is $(0, 0)$), the payoffs can be represented by the matrix as in Table II.

In this table, the row specifies the action of the first division general, the column specifies the action of the second division general. In each box, the first number specifies the payoff to the first division general; the second number specifies the payoff to the second division general.

If the enemy is unprepared (i.e., the state is anything other than $(0, 0)$), the payoffs are as given in Table III.

- In the strategic co-ordinated attack problem with the naive communication protocol, both generals *never* attack if the communication system is sufficiently reliable.

The argument is as follows. Clearly the first division general will never attack if he knows the enemy is prepared. Now suppose $\varepsilon < \delta$ and the second division general never receives a message. He believes that with probability $\delta/(\delta + (1 - \delta)\varepsilon) > 1/2$, the enemy is prepared. Whatever he believes the first division general will do if the enemy were unprepared, his optimal action must be not to attack:

* This section is based on Rubinstein (1989).

not attacking gives a payoff of 0, while attacking gives an expected payoff of at most $(1/2)(-M) + (1/2)(1) = -(M-1)/2$ (recall that M is very large, and in particular greater than 1).

Now the first division general knows that the second division general will never attack if he does not receive any messages (i.e., in states $(0,0)$ and $(1,0)$). Suppose that the first division general knows that the enemy is unprepared (and so sends a message) but never receives a confirmation from the second division general. Thus the first division general believes the true state is either $(1,0)$ or $(1,1)$. He believes that with probability

$$\frac{(1-\delta)\varepsilon}{(1-\delta)\varepsilon + (1-\delta)\varepsilon(1-\varepsilon)} = \frac{1}{2-\varepsilon} > \frac{1}{2}\,,$$

the second division general did not receive any message (i.e., the true state is $(1,0)$) and so will not attack. By the same argument as before, this ensures that the first division general will not attack even if he knows the enemy is unprepared, but has received any confirmation. An unraveling argument ensures that attack never occurs. This argument holds no matter how small the *ex ante* probability that the enemy is prepared (δ) is, as long as the communication system is sufficiently reliable (i.e., ε is sufficiently small).

Despite the resemblance to the argument for the non-probabilistic co-ordinated attack problem, notice that the conclusion is *much* stronger. As in the non-probabilistic version, we have no attack occurring despite the fact that when many messages have been sent, both generals know that both generals know . . . , up to an arbitrary number of levels, that attack is desirable. But in the strategic case studied in this section (unlike in the non-probabilistic version) each general *would* attack if he assigned high probability to the enemy being unprepared and the other division attacking.

It is important to realize that the strategic scenario contains *two* changes from the probabilistic scenario of the previous section. First, instead of evaluating action protocols *ex ante*, we required that each general must have an incentive to attack once he was actually called upon to do so. In game theoretic language, we did not allow the generals to *commit* to strategies before the communication stage. Second, we allowed the generals to have *different* objectives. That is, the first division general would much rather that the second division attacked alone, than that the first division attacked alone. Both these features are necessary for the strategic version of the co-ordinated attack paradox to hold. If the generals just had different objectives, but they could commit *ex ante* to following a particular action protocol, high probability co-ordinated attack *is* possible. We will see in Section 6 that high probability co-ordinated attack is also possible if the generals have the same objectives.

Table IV. The simple communication protocol state space.

State	Enemy's preparedness	Probability
No message	Prepared	δ
Message sent but not recieved	Unprepared	$(1-\delta)\varepsilon$
Message sent and received	Unprepared	$(1-\delta)(1-\varepsilon)$

Table V. An equilibrium action protocol under the simple communication protocol.

State	Enemy's preparedness	Probability	First division general's action	Second division general's action
No message	Prepared	δ	Don't attack	Don't attack
Message sent but not received	Unprepared	$(1-\delta)\varepsilon$	Attack	Don't attack
Message sent and received	Unprepared	$(1-\delta)(1-\varepsilon)$	Attack	Attack

5. Approximate Common Knowledge

The strong conclusion of the previous section, that co-ordinated attack never occurs, is *not* robust to the communication protocol. The generals would indeed be exceptionally foolish to attempt to co-ordinate their attack using the naive communication protocol. Consider the following "simple communication protocol". Suppose that if the enemy is unprepared, the first division general sends one message to the second division general informing him of this state of affairs. The second division general sends *no* confirmation. This communication protocol gives rise to the state space as in Table IV.

Suppose the payoffs of different actions depend as before on the enemy's preparedness, i.e., Tables II and III. The payoffs and state space together define another incomplete information game. An action protocol for the two generals is said to be an *equilibrium* if each general's action is optimal, given his payoffs and information, and given that the other general follows the protocol.

- For sufficiently small ε, there exists an equilibrium of the strategic co-ordinated attack problem with the simple communication protocol where co-ordinated attack almost always occurs whenever the enemy is unprepared.

The equilibrium is described in Table V. Co-ordinated attack occurs with probability $(1-\delta)(1-\varepsilon)$, i.e., almost always when the enemy is unprepared, if communication is very reliable (i.e., ε is small). Note that this does not violate the non-probabilistic co-ordinated attack result, since the first division general is required to attack with no confirmation that his message was received. Thus he attaches pos-

itive probability to attacking alone. But he attaches probability $1-\varepsilon$ to the message being successfully received, and thus is prepared to attack (if $\varepsilon < (1/M+1)$).

Why is attack possible under the simple communication protocol but not under the naive communication protocol? We know that attack is possible if there is common knowledge that the enemy is unprepared. How close to common knowledge do the simple and naive communication protocols get? To answer this question, we need an appropriate notion of approximate common knowledge. We have already observed that a high number of levels of knowledge does not generate outcomes close to common knowledge in strategic environments. Consider the following alternative notion. Let p be some probability. Say that an event is "common p-belief" if everyone believes it with probability at least p; everyone believes with probability at least p that everyone believes it with probability at least p; and so on. In standard probabilistic settings, common knowledge is equivalent to common 1-belief. Common p-belief has a fixed point characterization analogous to that of common knowledge (Monderer and Samet, 1989). In particular, if an event is common p-belief, then everyone believes with probability at least p that it is common p-belief. An action protocol is a *strict equilibrium* if every player's equilibrium action gives a strictly higher payoff than any alternative action, as long as other players follow the protocol.

- Suppose an action protocol is a strict equilibrium if the payoffs of a (finite) game are common knowledge; then there exists $p < 1$ such that that action protocol is a strict equilibrium when the payoffs are common p-belief.

A general statement or proof of this claim is beyond the scope of this paper.* But we can illustrate with our example. Consider the payoffs which result if the enemy is unprepared (i.e., Table III). If it were common knowledge that the enemy was unprepared, this game would have two equilibria: both generals attack and both generals don't attack. Consider the attack equilibrium. The action "attack" is optimal for either of the two generals exactly if he assigns probability at least $(M/M+1)$ to the other general attacking.

If it is common $(M/M+1)$-belief among the two generals that the enemy is unprepared, there is an equilibrium where both attack. A constructive argument shows why: if each general always attacks exactly when it is common $(M/M+1)$-belief that the enemy is unprepared, each general has an incentive to do so. Thus to understand the difference between the naive communication protocol and the simple communication protocol, we must examine when there is common p-belief that the enemy is unprepared, under the two protocols.

* Monderer and Samet (1989) introduced the version of common p-belief discussed here and first studied the connection between common p-belief and game theory, for p close to 1. The above result can be seen as a special case of their results. Morris et al. (1995) have elaborated the mechanism underlying the unravelling argument of the previous section and its relation to the absence of common p-belief.

- In the naive communication protocol, for any $p \geq 1/2$, it is never common p-belief that the enemy is unprepared.

We can show this by explicitly identifying the states where it is p-believed that the enemy is unprepared, up to different levels. First, both generals p-believe that the enemy is unprepared if at least one message has been sent and received, i.e., at those states (n, m) with $n \geq 1$ and $m \geq 1$. When do both generals p-believe that both generals p-believe that the enemy is unprepared? If the first division general has received no confirmation (i.e., $n < 2$), then he assigns probability less than $1/2$ (and thus less than p) to the second division general having received the first message. Thus both generals p-believe that both generals p-believe that the enemy is unprepared if and only if at least one confirmation has been sent and received, i.e., at those states (n, m) with $n \geq 2$ and $m \geq 1$. To get one more level of belief, we must have one more message sent. Thus however many messages are successfully sent, there is never common p-belief that the enemy is unprepared.

- In the simple communication protocol, it is common $(1 - \varepsilon)$-belief that the enemy is unprepared if the one message is sent and received.

Suppose the one message is sent and received. In this state, both generals attach probability 1 to the enemy being unprepared. This implies both generals 1-believe, and thus $(1 - \varepsilon)$-believe, that the enemy is unprepared. In fact, this is the *only* state where they $(1 - \varepsilon)$-believe that the enemy is unprepared. But at this state, the first division general attaches probability $(1 - \varepsilon)$ to being at this state, while the second division general attaches probability 1. Thus both generals $(1 - \varepsilon)$-believe that both $(1 - \varepsilon)$-believe that the enemy is unprepared. This argument iterates to ensure that it is common $(1 - \varepsilon)$-belief that the enemy is unprepared.

5.1. Implications for Game Theory and Economics

The strategic co-ordinated attack paradox appears to be an artifact of an artificial and foolish communication system: it is quite simple to achieve the appropriate approximate common knowledge in an unreliable communication system. Does the strategic co-ordinated attack problem have any practical interest, then? It does suggest two natural lines of future research.

First, if you could design the communication system, subject to unavoidable communication errors, that would be used by players in a strategic environment, how would you do it (see, for example, Chwe, 1995)? Not like the naive communication protocol, clearly. But the approximate common knowledge results provide hints about how it should be designed.

However, there are many economic environments where the communication system, or information structure, is exogenously given to decision makers. If those decision makers must then make strategic decisions, the issues raised by the naive

communication protocol are of the utmost importance. The messengers story is of course a little contrived. Rubinstein (1989) used a more contemporary story about the same formal information structure, where players communicate with electronic mail messages which may get lost. But Carlsson and van Damme (1993) have considered an information structure where two players each receive a noisy signal of the true state of payoffs. It turns out that this simple and natural information structure has all the same implications as the naive communication protocol. So the second question for future research is: given that such unfortunate information structures are exogenously given, how should institutions be designed to deal with them (see, for example, Shin, 1996)?

6. Achieving Approximate Common Knowledge

The naive communication protocol revealed a remarkable fact about common p-belief. Suppose that the probability that the enemy is unprepared (δ) is very small, and the communication is very reliable ($\varepsilon < \delta$). Then we have an event E ("the enemy is unprepared") that has probability very close to one ($1 - \delta$), while the probability that it is common $1/2$-belief is zero. This suggests that there need be no connection between the *ex ante* probability that the enemy is unprepared and the possibility of equilibrium co-ordinated attack (as long as the communication protocol is sufficiently bad).

It turns out that the situation is not quite so bad, at least as long as players' beliefs are generated by the "common prior assumption" (Morris, 1995a). That is, there is a commonly agreed prior probability distribution on the set of possible states and players' beliefs are generated by updating by Bayes rule on that state space (this assumption is standard in game theory).

- Suppose that $p < 1/2$ and the probability of event E is $1 - \delta$. Then the probability that event E is common p-belief is at least $1 - \delta(1 - p/1 - 2p)$.*

Remarkably, it is possible to deduce properties of players' higher order beliefs from *ex ante* probabilities. This result can be combined with the earlier common p-belief result to prove equilibrium results. Suppose that we altered the payoffs of the generals in the strategic co-ordinated attack problem so that there was no conflict of interest. In particular, suppose now that each general gets a payoff of $-M$ if his division does not attack, but the other division does. Thus the conflict of interest is removed. The new payoff matrix is given in Table VI.

The key game theoretic change is that now the both attack equilibrium is "risk dominant": that is, there is a probability p *less* than $1/2$ such that if one general assigns probability p to the other's division attacking, his best response is to attack. This p is $(M/2M + 1)$. Thus there is an equilibrium where attack occurs as long

* This result is due to Kajii and Morris (1995).

Table VI. Payoffs if the enemy is unprepared (symmetric case).

	Attack	Don't attack
Attack	$1, 1$	$-M, -M$
Don't attack	$-M, -M$	$0, 0$

as it is common $(M/2M+1)$-belief that the enemy is unprepared. But we know that this is true with probability at least

$$1-\delta\left(\frac{1-p}{1-2p}\right)=1-\delta\left(\frac{1-(M/2M+1)}{1-(2M/2M+1)}\right)=1-\delta(M+1).$$

This is true under *any* communication protocol. Thus we have:

- Under *any* communication protocol, if the enemy is unprepared with sufficiently high probability, then there is an equilibrium of the symmetric strategic co-ordinated attack game where co-ordinated attack occurs with probability close to one.

7. Common Knowledge and Timing

Let us give one last example illustrating the importance of strategic issues. Common knowledge is fragile in many dimensions. Consider the following variation of the co-ordinated attack problem. Suppose now that messages are perfectly reliable – they arrive with probability one – but the length of time they take is uncertain. For simplicity, suppose that a message arrives either instantaneously or after ε seconds.*

At some point in time, the first division general will receive his intelligence about the preparedness of the enemy, and immediately send a message to the second division general. This is the *only* message sent. If the enemy is unprepared, the generals would like to co-ordinate a *simultaneous* attack. Unfortunately, the generals do not have synchronized clocks and they do not know (we assume) whether the message took 0 seconds or ε seconds to arrive. For future reference, let us assume that each outcome occurs with probability $1/2$ and ε is small.

It never becomes common knowledge that the enemy is unprepared. The first division general knows that the second general has received his message ε seconds after he sent it. Since the second division general thinks that the message may have taken only 0 seconds, he knows that the first division general knows that he has received the message only ε seconds after receiving it. Thus the first division general knows that the second division general knows that the first division general knows that the message has arrived only 2ε seconds after the message is sent.

* This example is due to Halpern and Moses (1990).

This argument iterates to ensure that it never becomes common knowledge. Thus co-ordinated attack is impossible.

But suppose that it is enough to ensure that the length of time when only one general is attacking is short. Such approximate co-ordination is easy to achieve with non-strategic generals if the delay ε is small. Consider the action protocol where the first division general attacks as soon as he knows that the enemy is unprepared (simultaneously sending a message to the second division general). The second general attacks as soon as he receives the message. Thus unco-ordinated attack occurs for at most ε seconds.

But consider the following strategic version of the problem (Morris, 1995b). Suppose that the payoffs of Tables II and III represent the generals' instantaneous payoffs at any moment in time. Thus each general gets a flow payoff of $-M$ each second that his division is attacking alone, but a flow payoff of 1 whenever both divisions are attacking and the enemy is unprepared. An action protocol specifies when each general starts attacking, as a function of when he receives a signal (remember, there are no synchronized clocks). For example, the first division general might plan on attacking x seconds after he hears that the enemy is unprepared while the second division general might plan to attack y seconds after he receives the message from the first division general. Thus an action protocol is described by the two numbers, (x, y).

For what values of x and y is this action protocol an equilibrium? That is, when is it the case that each general would actually want to follow this plan if he expected the other general to follow that plan? Suppose the first division general expects the second division to attack y seconds after receiving message: y seconds after *sending* his message, the first division general attaches probability $1/2$ to the second division general attacking, and $1/2$ to his waiting another ε seconds. Thus his flow payoff from attacking (over the next ε seconds) is $(1/2)(-M) + (1/2)(1) < 0$. So he waits until $y + \varepsilon$ to attack. Thus if (x^*, y^*) is an equilibrium, we must have $x^* = y^* + \varepsilon$. On the other hand, suppose the second division general expects the first division general to attack x seconds after *sending* his message: $x - \varepsilon$ seconds after *receiving* his message, the second division general attaches probability $1/2$ to the first division general attacking, and $1/2$ to him waiting another ε seconds. Thus his flow payoff from attacking (over the next ε seconds) is $(1/2)(-M) + (1/2)(1) < 0$. So he waits until x seconds after receiving his message to attack. Thus if (x^*, y^*) is an equilibrium, we must have $y^* = x^*$. But we earlier showed that $x^* = y^* + \varepsilon$, a contradiction. So the only equilibrium action protocol has each general never attacking.

Thus we see again that strategic problems require a stringent form of approximate common knowledge in order to attain approximate co-ordination, while non-strategic problems do not. We showed this first in Section 4, where common knowledge failed because of communication *errors*. We showed a similar conclusion in this section, where common knowledge failed because of *asynchronous* communication.

Table VII.

Home town general's payoff when dry	$-(x-y)^2 - y$
Out-of-town general's payoff when dry	$-(x-y)^2 - x$
Home town general's payoff when wet	$-(x-y)^2 + y$
Out-of-town general's payoff when wet	$-(x-y)^2 + x$

Table VIII.

	Dry forecast	Wet forecast
Dry	$(1-\varepsilon)/2$	$\varepsilon/2$
Wet	$\varepsilon/2$	$(1-\varepsilon)/2$

8. Approximate Common Knowledge Does Not Allow Intricate Conventions

Our discussion of the sufficiency of common p-belief in securing coordination needs a caveat. So far, the actions taken by the participants have been binary (to attack or not to attack), and in general finite. If, however, the participants are choosing from a continuum, qualifications are necessary. As many interesting coordination problems belong to this class, this is a potentially important point to bear in mind, and we can also draw some conclusions on what sorts of conventions might emerge in such situations.

Imagine that the two generals are now back home in peace time. They live in neighbouring towns, and meet regularly every weekend in the first general's town. This town has a long Main Street, which is represented by the unit interval $[0, 1]$. At one end of Main Street (at point 0) is an open air bar, while at the other end (at point 1), there is an indoor restaurant. The two generals prefer to meet at the open air bar if the weather is dry, and prefer the indoor restaurant if the weather is wet. In any case, they aim to meet at the same point on Main Street, but have to choose a spot somewhere on Main Street to turn up. Denote by x the point on Main Street chosen by the home town general, and denote by y the point chosen by the out-of-town general. The coordination problem of the generals is implicit in their payoffs, which are given in Table VII.

Clearly, the best outcome would be if they could coordinate a convention in which they both turn up at the open air bar (point 0) if the weather is dry, and they both turn up at the indoor restaurant (point 1) if the weather is wet.

The home town general can observe the weather in his own town perfectly. If the other general could also observe the weather situation perfectly, then such coordination could be achieved. There would then be common 1-belief of the weather in the first general's town. However, suppose that the out-of-town general only has a noisy signal of the true weather situation in his friend's town. The out-of-town general receives a forecast which is very accurate, but not perfectly so. The joint distribution over the true weather situation and the forecast is shown

in Table VIII, where ε is a small positive number. Notice that there is common $(1-\varepsilon)$-belief of the true weather situation, so that when the noise is small, there is common p-belief with a high p. If we were to extrapolate the results discussed so far, we might venture to speculate that when ε is small enough the generals can achieve full coordination. However, this is not the case. However small ε is, the best that they can do is to ignore their information completely and turn up at some point on Main Street whether the weather is dry or whether the weather is wet.

To see this, denote by x_d the spot chosen by the home town general when the weather is dry, and by x_w the spot chosen when the weather is wet; y_d and y_w are defined analogously for the out-of-town general. When the weather is dry, the home town general puts probability $(1-\varepsilon)$ to his friend having the accurate forecast, and probability ε to his having an inaccurate forecast. In the former, the out-of-town general takes action y_d, and in the latter, he takes action y_w. So, the home town general's expected payoff is given by

$$-(1-\varepsilon)((x_d-y_d)^2-y_d)-\varepsilon((x_d-y_w)^2-y_w).$$

This is maximized when the home town general chooses

$$x_d=(1-\varepsilon)y_d+\varepsilon y_w.$$

Proceeding in this way, we can derive the following pair of matrix equations as necessary conditions for *any* convention.

$$\begin{bmatrix} x_d \\ x_w \end{bmatrix} = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix} \begin{bmatrix} y_d \\ y_w \end{bmatrix}$$
$$\begin{bmatrix} y_d \\ y_w \end{bmatrix} = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix} \begin{bmatrix} x_d \\ x_w \end{bmatrix}$$

from which we get

$$\begin{bmatrix} x_d \\ x_w \end{bmatrix} = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix}^2 \begin{bmatrix} x_d \\ x_w \end{bmatrix}.$$

As long as ε is positive, any solution of this equation satisfies $x_d = x_w$, so that any convention has $x_d = x_w = y_d = y_w$. The best the two generals can do is to ignore their information altogether.

The lesson to be drawn from this example is that when coordination is to be achieved over a large set of possible outcomes, common p-belief of the underlying circumstances will not be enough. What is necessary is common 1-belief, or what amounts to the same thing, common knowledge. Shin and Williamson (1996) show that this argument holds generally, and that any convention in which participants choose from a continuum can only rely on events which are common knowledge whenever they occur.

8.1. An Implication for the Philosophy of Language

If the players, taken individually, are capable of making fine distinctions across the states of the world, and the players were able to utilize this information in their convention, we could regard such a convention as being 'intricate' relative to the information of individual players. In contrast, if an equilibrium involves strategies which do not utilize to any great extent the private information of individual players, then we may informally dub such an equilibrium "simple", relative to the information of individual players. We have seen that for coordination problems with large action sets, the latter is the case with a vengeance. Any information is useless, unless it is commonly shared by all.

This conclusion is intriguing, since many of the conventions around us tend to be simple relative to the set of signals on which we could condition our actions. For example, it could be argued that our conventions concerning appropriate dress are too rigid relative to rules which are superior. Why isn't there a convention that jackets and ties should be worn by men to dinner unless the temperature is above a threshold level, in which case all the men turn up in T shirts? On a less frivolous note, the archetypal convention is that of language, and the meaning we associate with sentences. Lewis's (1969) account of the use of language as an equilibrium in a coordination game has been very influential among philosophers.* In this context, it is an intuitively appealing principle that the meaning of sentences should not vary with, say, the season or day of the week. For example, it is difficult for us to picture a language in which the sentence "the cat is on the mat", means that the cat is on the mat, on weekdays but means that the cherry is on the tree, on weekends. However, if we take at face value the claim that language is merely a coordination device, we are hard put to give a quick rebuttal to such oddities. Why are such perverse languages not observed in practice? We have provided one possible answer. We have shown that for coordination problems over large sets (and indeed, the set of possible languages is rather large), the optimizing behaviour of individuals eliminate all rules of action except those which rely only on the signals shared commonly by all. Our beliefs of the empirical features of the world will often fail this test of commonality. Conventions defined over large action sets cannot rely on information which is only available to a strict subset of the participants, however large his subset might be, and however small the possibility of error might be.

The argument we outlined above concerned a static or "one-shot" strategic situation. However, similar issues will arise in studies of the evolution of conventions through time. In the simple story outlined above, the best-reply structure generates a stochastic process on actions which ensures convergence to the simple convention, and Shin and Williamson's argument relies on the construction of a martingale on the set of actions, and on its convergence. Explicit modeling of learning and other

* Although the strategic analysis of language and conventions has a long history, it is not a feature emphasized in modern logical semantics. But see the chapter by Hintikka and Sandu on game-theoretic semantics in van Benthem and ter Meulen (1996).

dynamic aspects of conventions can be expected to strengthen the presumption in favour of simple conventions.

9. Conclusion

The significance of the results outlined in our paper depend, to a large extent, on what notions of knowledge and belief are applicable to a particular situation. If knowledge is interpreted as a logical notion – as something which is implied logically by more basic propositions – then common knowledge could be regarded as being easily attained, since it is merely the consequence of the logical competence of agents. However, such a view of knowledge is of limited use in many fields in which beliefs are derived from empirical sources, or at least, sources which are not infallible. We may often have good evidence that a proposition is true, but almost invariably, this is not a watertight guarantee. If coordination requires common knowledge concerning such empirical facts about the world, then common knowledge seems to be an excessively strong requirement that will rarely be achieved in practice.

Thus, although we have the benchmark result that perfect coordination requires common knowledge, this need not be very informative to researchers in the field if common knowledge is unlikely to be attainable in practice, or if perfect coordination can be supplanted by some notion of "sufficient" coordination which would suffice for most purposes. Our goal in this paper has been to outline some of the ways in which the two notions (common knowledge and perfect coordination) can be relaxed, and how the relaxed notions can be related to one another.

Recent developments in game theory have offered several insights which may be of relevance to neighbouring disciplines which employ the notion of common knowledge. Firstly, if agents behave strategically, coordination may be difficult to achieve, even if coordination in the corresponding non-strategic situation is easy to attain. Secondly, the appropriate notion of "almost common knowledge" in strategic situations involving a finite number of actions is that of common p-belief, rather than large numbers of iterations of knowledge among the participants. We have also alluded to how the notion of common p-belief interacts with the idea of a common prior probability distribution over the space of uncertainty to impose bounds on how much coordination might be achievable on average. Finally, we have ended on a cautionary note that when coordination is to take place over a continuum, full common knowledge (rather than its approximate cousin) is required. In this extreme case, only the most stringent notion of common knowledge will achieve coordination, and it serves as a benchmark on how far common knowledge can be relaxed.

References

Aumann, R., 1992, "Interactive Epistemology," unpublished paper, Hebrew University of Jerusalem.

Barwise, J., 1988, "Three views of common knowledge," pp. 365–379 in *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, M. Vardi, ed., San Francisco: Morgan Kaufmann.

van Benthem, J. and ter Meulen, A., 1996, *Handbook of Logic and Language*, Amsterdam: Elsevier Science (forthcoming).

Bonanno, G., 1996, "On the logic of common belief," *Mathematical Logic Quarterly* **42**, 305–311.

Carlsson, H. and van Damme, E., 1993, "Global games and equilibrium selection," *Econometrica* **61**, 989–1018.

Clark. H. and Marshall, C., 1981, "Definite reference and mutual knowledge," in *Elements of Discourse Understanding*, A. Joshi, B. Webber, and I. Sag, eds., Cambridge: Cambridge University Press.

Chwe, M., 1995, "Strategic reliability of communication networks," University of Chicago Graduate School of Business, Studies in Theoretical Economics Working Paper #21.

Dekel, E. and Gul, F., 1996, "Rationality and knowledge in game theory," in *Advances in Economic Theory: Seventh World Congress of the Econometric Society*, D. Kreps and K. Wallace, eds., Cambridge: Cambridge University Press (forthcoming).

Fagin, R., Halpern, J., Moses, Y., and Vardi, M., 1995, *Reasoning about Knowledge*, Cambridge, MA: MIT Press.

Geanakoplos, J., 1994, "Common knowledge," *Handbook of Game Theory*, Chapter 40 of Volume 2, R. Aumann and S. Hart, eds., Amsterdam: Elsevier Science.

Halpern, J. and Moses, Y., 1990, "Knowledge and common knowledge in a distributed environment," *Journal of the ACM* **37**, 549–587.

Halpern, J. and Moses, Y., 1992, "A guide to completeness and complexity for modal logics of knowledge and belief," *Artificial Intelligence* **54**, 319–379.

Halpern, J. and Tuttle, B., 1993, "Knowledge, probability and adversaries," *Journal of the ACM* **40**, 917–962.

Hintikka, J., 1962, *Knowledge and Belief*, Ithaca, NY: Cornell University Press.

Kajii, A. and Morris, S., 1995, "The robustness of equilibria to incomplete information," University of Pennsylvania CARESS Working Paper #95-18, forthcoming in *Econometrica*.

Kripke, S.A., 1959, "A completeness theorem for modal logic," *Journal of Symbolic Logic* **117**, 1–14.

Lewis, D., 1969, *Conventions: A Philosophical Study*, Cambridge, MA: Harvard University Press.

Lismont, L. and Mongin, P., 1995, "Belief closure: A semantics of common knowledge for Modal Propositional Logic," *Mathematical Social Sciences* **30**, 127–153.

Monderer, D. and Samet, D., 1989, "Approximating common knowledge with common beliefs," *Games and Economic Behavior* **1**, 170–190.

Morris, S., 1995a, "The common prior assumption in economic theory," *Economics and Philosophy* **11**, 227–253.

Morris, S., 1995b, "Co-operation and timing," University of Pennsylvania CARESS Working Paper #95-05.

Morris, S., Rob, R., and Shin, H., 1995, "p-Dominance and belief potential," *Econometrica* **63**, 145–167.

Osborne, M. and Rubinstein, A., 1994, *A Course in Game Theory*, Cambridge, MA: MIT Press.

Rubinstein, A., 1989, "The electronic mail game: Strategic behavior and almost common knowledge," *American Economic Review* **79**, 385–391.

Shin, H.S., 1993, "Logical structure of common knowledge," *Journal of Economic Theory* **60**, 1–13.

Shin, H.S., 1996, "Comparing the robustness of trading systems to higher-order uncertainty," *Review of Economic Studies* **63**, 39–60.

Shin, H.S. and Williamson, T., 1996, "How much common belief is necessary for a convention?," *Games and Economic Behavior* **13**, 252–268.