NOÛS

NOÛS 0:0 (2017) 1–25
doi: 10.1111/nous.12186

*Two Paradoxes of Common Knowledge: Coordinated Attack and Electronic Mail**

HARVEY LEDERMAN

Abstract

The coordinated attack scenario and the electronic mail game are two paradoxes of common knowledge. In simple mathematical models of these scenarios, the agents represented by the models can coordinate only if they have common knowledge that they will. As a result, the models predict that the agents will not coordinate in situations where it would be rational to coordinate. I argue that we should resolve this conflict between the models and facts about what it would be rational to do by rejecting common knowledge assumptions implicit in the models. I focus on the assumption that the agents have common knowledge that they are rational, and provide models to show that denying this assumption suffices for a resolution of the paradoxes. I describe how my resolution of the paradoxes fits into a general story about the relationship between rationality in situations involving a single agent and rationality in situations involving many agents.

1. Introduction

Two divisions of an army are camped on separate hilltops overlooking a valley. In the valley awaits the enemy. If both divisions attack the enemy simultaneously they will win the battle, while if only one division attacks it will suffer a catastrophic defeat. Each of the generals commanding these hilltop divisions wants to avoid a catastrophic defeat: neither of them will attack unless he believes that the general commanding the other division will attack with him. During the night a thick fog descends over the hilltops; the only way the generals can communicate is by sending a messenger through the enemy camp.[1]

Some agents *commonly know* a proposition just in case they all know it, they all know that they all know it, they all know that they all know that they all know it, and so on. They *commonly believe* a proposition just in case they all believe it, they all believe that they all believe it, and so on.[2] It can be shown that in simple models of this *coordinated attack scenario* the generals will attack only if they commonly believe that they will. It can also be shown that no matter how many messages are sent, the generals will not have common belief that they will attack. Thus the generals represented in these simple models will not attack, no matter how many messages are sent and received.

*Thanks to Mike Caie, Kevin Dorst and Jeremy Goodman for helpful comments on a draft of this paper. Thanks also to Kyle Thomas for discussion of some of these issues, and to Teddy Seidenfeld for correspondence after a talk on related material.

This result conflicts with a powerful intuition about what it could be rational to do. It seems eminently rational to attack after finitely many messages have been sent. Suppose one of the generals repeatedly risks the life of his messenger, transmitting seventeen messages in the course of the night, each saying that he will attack the next day. But still, he does not attack. Even supposing that his counterpart general also does not attack, this behavior seems bizarre. What will they say when they return to court and the Queen wonders why they did not defeat the enemy on that fateful day? "Your highness, with only seventeen messages, we just weren't ready to attack".

This is the paradox of coordinated attack. A powerful intuition about rational action conflicts with the behavior predicted by simple mathematical models of rational agents. The second paradox of my title—the electronic mail game—has a similar structure. In that game, an action which is intuitively rational is also in conflict with the predictions of a simple mathematical models of how rational agents will play the game.

In this paper, I will propose a resolution of these paradoxes. I will argue that in the coordinated attack scenario the problem lies in assumptions which lead to the result that the generals cannot rationally attack unless they commonly believe that they will attack. I will show that this result depends on strong background assumptions of common belief, most notably the assumption that the generals commonly believe that they are rational, in a sense to be made precise. If we give up this assumption of common belief, we eliminate the result that the generals can coordinate rationally only if they commonly believe that they coordinate: we can give models where the generals are rational and attack after sending finitely many messages. A similar point holds for the electronic mail game: relaxing the assumption that the players are commonly certain that they are rational makes it possible for the players to perform the action which is intuitively rational in that game. I conclude that these two paradoxes of common knowledge are best understood as exhibiting false consequences of the assumption that when rational agents or people interact, they commonly believe one another to be rational.

My solution to the paradoxes might seem to be unduly radical, since it might seem to force us to abandon well-established theories of rational behavior in situations involving many agents. Common knowledge of rationality is, to be sure, almost universally assumed in standard models of rational agents in game theory. In the concluding section of the paper, I argue, however, that common knowledge of rationality is not a basic postulate governing the behavior of rational agents; it is merely a technical assumption which makes models of rational behavior more mathematically tractable. An agent may be rational even if he or she fails to know that others are rational; rational agents may thus interact without commonly knowing one another to be rational. In the conclusion I show how this view of common knowledge of rationality fits into a general, independently attractive understanding of the relationship between rationality in situations involving a single agent (often called "decisions") and rationality in situations involving many agents ("games"). This conclusion can be read independently, and may be of interest to readers generally concerned with the relationship between individual

(decision-theoretic) rationality and social (game-theoretic) rationality, even if they are not interested in the formal details of the rest of the paper.[3]

Section 2 describes the coordinated attack scenario in more detail. Section 3 presents a model of the coordinated attack scenario in which agents coordinate rationally without commonly believing that they coordinate. Section 4 describes the electronic mail game. Section 5 provides a model of the electronic mail game in which the players coordinate on the better outcome of the game without having common common certainty that they will coordinate. Section 6 concludes with a discussion of the role of common knowledge in the study of rationality in social situations. An appendix contains formal details omitted from the main text.

2. Coordinated Attack

The paradox of coordinated attack can be seen as deriving from the conflict of three claims. First, if the generals are rational, they will attack only if they commonly believe they will attack. Second, if the generals form beliefs correctly on the basis of the messages they receive, then the generals in this scenario will not achieve common belief that they will attack. These first two claims are supported by simple mathematical models of action on the one hand and message-passing on the other. Taken together the claims imply that if they are rational the generals will not attack, no matter how many messages are passed. This conclusion conflicts with a third claim: that it can be rational to attack after a small finite number of messages. One of these three claims must be rejected.

At first one might be tempted to reject the intuition that attacking could be rational after finitely many messages are sent, holding on to the idea that the formal models accurately represent rational action and belief-formation. But on inspection this reply is unattractive. The word "rational" can be understood here in an undemanding sense, where to say that an action is rational is akin to saying that it "makes sense", or "is explicable". This notion of rationality is central to an important form of theory about human behavior. Assuming that most of the time people's actions make sense or are explicable, the hypothesis that people are rational helps to predict what they will do. While the simple formal models in question here are highly idealized, they still may be understood as intended to be part of a theory of this undemanding form of rationality.

If "rational" is used in this undemanding sense, however, it is difficult to deny that attacking could be rational in this situation. In numerous studies of close variants of the coordinated attack scenario, many people do choose the analogue of attacking after receiving finitely many messages (Camerer (2003, p. 226–232), Heinemann *et al.* (2004), Kneeland (2016), Thomas *et al.* (2014)). It is of course open for a theory of rationality to declare the actions observed in the laboratory to be irrational. But it is hard to see how to confine this verdict of irrationality to people's behavior in the experiments. More normal situations in which people typically coordinate are structurally analogous to the coordinated attack scenario.[4] If one is willing to accept the simple models of rational action and belief formation as reasonable, albeit idealized, descriptions of rational agents in the experimental

set-up, one should also be willing to accept them as reasonable descriptions of what happens "in the wild". The result that people make mistakes in a small class of complex laboratory experiments is a tolerable cost to a theory of this undemanding kind of rationality. The result that people generally make mistakes in much more familiar cases is not.

A second reply to the paradox is to reject the claim that the generals cannot achieve common belief by sending messages to one another. According to this reply, after five or six messages—and certainly after seventeen—the generals do achieve common belief that they have each received one message, and hence, common belief that they will attack.

This reply, while perhaps partially correct, is not general enough to provide a full resolution of the paradox. Perhaps it is true that some people who choose the analogue of attacking in experiments (for example) do so because they are not considering the situation in sufficient detail; they haven't thought precisely about what the other person believes on the basis of the messages that other person has received, and they choose to attack on the basis of their mistaken assessment. But it seems unlikely that this is the whole story. For even supposing that a person does attend closely to how many messages have been sent and received, and is also aware of the kind of reasoning used to argue that the generals do not commonly know that the first message was passed, it still seems as if it could be rational for him or her to attack. The general on the North hilltop might receive his fourth message of the exchange, and reason as follows. "The South General has received three messages. So he'll surely attack. It's true that he doesn't know that I've received this message. He also doesn't know that I know that he received my previous message [two occurrences of "know"], doesn't know that I know that he knows that I received the message before that [three occurrences of "know"], and consequently doesn't know that I know that he knows that I know that he received one message [four occurrences of "know"]. But so what? After receiving three messages, any sensible person would attack. He'll surely attack, and thus, I should attack, too".[5] This train of thought seems perfectly reasonable in the circumstances, but the reply we are considering predicts that it is not. For the reply leaves untouched the result that rational agents should coordinate in this scenario only if they commonly believe (or commonly know) that they will. But the reasoning just described explicitly countenances the failure of common knowledge that the first message was passed (and hence presumably the failure of common knowledge that the generals will attack), while nevertheless concluding that attacking is the best of the available options. So while this second reply may help to explain some behavior and some of our judgments of rationality, it does not get to the heart of the paradox.[6]

I propose that we should escape the paradox by rejecting the remaining claim: that rational agents can coordinate in this scenario only if they commonly believe that they coordinate. This claim can be proven using apparently modest assumptions about the generals. To show how to reject the claim, then, we must examine the assumptions which imply it.

The result can be stated compactly in a simple modal propositional language, containing two propositional atoms, $Attack(N)$ ("North attacks") and $Attack(S)$

("South attacks"), the Boolean connectives $\neg$ and $\wedge$, and three monadic sentential operators, B_N ("North believes"), B_S ("South believes") and C ("North and South commonly believe"). The formation rules for this language are the obvious ones; we use the metalinguistic abbreviations $\vee$ and $\rightarrow$ in the standard way, and in addition use $Attack(NS)$ as a metalinguistic abbreviation for $Attack(N) \wedge Attack(S)$.

In both the formal and the informal discussion, it will be helpful to have some further terminology related to common belief. North and South *mutually believe* (or: mutually believe^1) a proposition just in case they both believe it. In the formal language, we use $M^1\varphi$ to abbreviate the sentence $B_N\varphi \wedge B_S\varphi$, which is interpreted as $\ulcorner$both generals believe that $\varphi\urcorner$. Beyond mutual belief^1, there are "higher orders" of mutual belief. North and South mutually believe^2 a proposition just in case they mutually believe that they mutually believe^1 it. More generally, they mutually believe^n that φ just in case they mutually believe that they mutually believe^{n-1} that φ. In the formal language we use $M^n\varphi$ to abbreviate $M^1(M^{n-1}\varphi)$. This terminology will be useful in part because it allows us to state facts related to common belief concisely: for example, the generals commonly believe something just in case for all n, they mutually believe^n it.

The proof of the result can be conducted in a multi-modal logic where each modal operator—B_N, B_S, C—obeys the normal modal logic **K**; we call this logic $\mathbf{K_{NSC}}$. The details of this system are given in the appendix for those who are interested, but the body of the paper is intended to be legible without those details.

With this logic in the background, the proof requires only two further assumptions. The first can be motivated by considering the story of coordinated attack in greater detail. In the story I suggested that the costs of catastrophic defeat are sufficiently great that if one general does not believe that the other general will attack (for example, if he has not received a single message) he will conclude that the best available action is the safe one, not to attack. It is somewhat odd to use the verb "believe" to describe the scenario, but it is easy to imagine someone describing it using the verbs "know" or "think": "If North doesn't know whether South is going to attack, he shouldn't attack" or "If North doesn't think that South is going to attack, he shouldn't attack".[7] In this particular example, then, it seems plausible that a general will be rational in attacking only if he believes that the other general will attack. In symbols, we may characterize this constraint on rationality schematically as follows:

$$\textbf{Rationality}: Rat(i) := Attack(i) \rightarrow B_i(Attack(j)) \quad \text{where } j \neq i \qquad (1)$$

Thus for example $Rat(N)$ will be used as an abbreviation for $Attack(N) \rightarrow B_N(Attack(S))$. I will also use $Rat(NS)$ to abbreviate $Rat(N) \wedge Rat(S)$.

To some, the use of "rationality" here may seem misleading or even incorrect. In decision theory and game theory it is standard to speak about probabilistic degrees of confidence, and not about "all-out" belief. Rationality is defined as the maximization of subjective expected utility; it is not defined in terms of what one should do given what one believes. But the restriction of the demands of rationality to constraints on what one should do given that one has a particular

pattern of degrees of confidence is artificial. A more general, intuitive notion of "rationality" also imposes demands on what an agent should do given what she believes and knows (not merely "believes to thus and such a degree"). The paradox of coordinated attack operates with this notion of rationality, spelled out in terms of "full" or "qualitative" belief.[8]

In addition to this assumption about rationality, we make a second assumption: that if a general attacks, he believes that he attacks. Even if there are circumstances we could imagine in which a general would attack but not believe that he is doing so, we may assume that the situation we are describing is not a circumstance of that kind. I will use the following schematic abbreviation to describe this:

$$\textbf{Transparency} : Trans(i) := Attack(i) \rightarrow B_i(Attack(i)) \tag{2}$$

Thus $Trans(N)$ will be used as an abbreviation for $Attack(N) \rightarrow B_N(Attack(N))$; $Trans(NS)$ will then be used to abbreviate $Trans(N) \wedge Trans(S)$. Finally, further simplifying notation, I will use $Ideal(NS)$ to abbreviate $Rat(NS) \wedge Trans(NS)$. The theorem then states

Theorem 2.1.

$$Ideal(NS), C(Ideal(NS)) \vdash_{\mathbf{K_{NSC}}} Attack(NS) \rightarrow C(Attack(NS)).^{9}$$

In other words, if the agents attack, they commonly believe that they attack. Common belief is a precondition for coordination.

3. Common Knowledge of Rationality Part I

I have already argued that we should respond to the paradoxes by giving up one of the assumptions which lead to this theorem. In this section I will show that denying common belief in rationality suffices to eliminate the theorem, in the precise sense that for any n, if we require only mutual belief^n in rationality for some finite n, the conclusion no longer follows:

Proposition 3.1. *For all n,*

$$Ideal(NS), C(Trans(NS)), M^n(Rat(NS) \nvdash_{\mathbf{K_{NSC}}} Attack(NS) \rightarrow C(Attack(NS)).^{10}$$

This result is not surprising from a mathematical perspective. Relaxing the assumptions used in the proof of a theorem often allows the statement of the theorem to fail. But my focus here is on the conceptual point: that the paradoxes of common knowledge can be resolved by abandoning this background assumption of common knowledge.

How general is this response? One might worry that a "revenge" paradox can be constructed by simply assuming as a feature of the setup that the generals commonly believe that they are rational. Given this additional assumption, it would still be impossible for the generals to coordinate, at least for all I have said. But this "revenge" paradox is not a paradox: it does not conflict with the intuition that

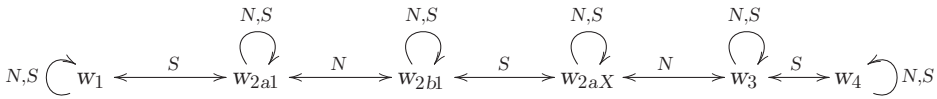
attacking could be rational after finitely many messages have been passed. For the intuition describes what it could make sense to do; it does not describe what it could make sense to do *for agents who have common belief that they are rational*. So all that follows from the "revenge" argument is that, since people do tend to attack and to do so rationally, they do not have common belief that they are rational. But this is a perfectly acceptable conclusion: it does not follow from the fact that two people are rational, that they commonly believe one another to be so.

To show that something does not follow from a given set of premises in a logic we must give a model of the logic in which the premises are true, but the conclusion is not. To do this, we use simple, idealized models of knowledge, broadly in the tradition of Hintikka (1962) (for an introduction, see Fagin *et al.* (1995)). A model is a structure $\mathcal{M} = \langle W, R_N, R_S, v \rangle$. The first element of the model is a set of "worlds", logically possible situations; the second and third elements of the model are binary "accessibility relations" for the two agents, General N and General S. These accessibility relations are used to give the semantics of the belief operators for the two generals. Informally, a world w' is accessible from a world w for an agent i if and only if what is true at w' is compatible with what i believes at w. The truth clauses for the operators are the standard ones; for $i \in \{N, S\}$:

- $\mathcal{M}, w \models \mathit{Attack}(i)$ if and only if $w \in v(\mathit{Attack}(i))$;
- $\mathcal{M}, w \models \neg\varphi$ if and only if $\mathcal{M}, w \not\models \varphi$;
- $\mathcal{M}, w \models \varphi \wedge \psi$ if and only if $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models B_i\varphi$ if and only if for all $w' \in W$ if $w R_i w'$ then $\mathcal{M}, w' \models \varphi$;
- $\mathcal{M}, w \models C\varphi$ if and only if for all n $\mathcal{M}, w \models M^n\varphi$.

A logic Λ is *sound* on a class of models just in case if the elements of any set of formulas Γ are true at some world w in some model $\mathcal{M}$ in the class, and if the members of the premise set Γ can be used to prove a formula φ in the logic Λ ($\Gamma \vdash_\Lambda \varphi$), then φ is also true at w. Standard results straightforwardly imply that the logic $\mathbf{K_{NSC}}$ mentioned earlier as our background logic is sound on the class of models just defined. Thus if we can give a model in this class where the premises of the theorem hold but its conclusion does not, we will have shown that the conclusion does not follow from the premises; we will have proven the proposition.[11]

The figure below depicts a model which relaxes the assumption that the players commonly believe one another to be rational. In the diagram, nodes represent worlds, and labeled directed edges represent the binary accessibility relations R_N and R_S; the row of text below the worlds indicates whether the world in question is an element of $v(\mathit{Attack}(N))$, $v(\mathit{Attack}(S))$, both or neither. As I describe in the appendix, extensions of the model can be used to prove Proposition 3.1 in full generality. Here, I'll just indicate part of that proof by giving the reader a sense for how this model works.



Let us begin with w_{2aX}. Here, N attacks even though he does not believe that S is attacking: w_3 is consistent with what he believes (it is accessible), but S does not attack at w_3. In symbols:

$$\mathcal{M}, w_{2aX} \vDash Attack(N) \wedge \neg B_N(Attack(S)).$$

In other words, N is not rational at this world; he attacks without believing that S attacks.

We now move one step to the left, to w_{2b1}, where the generals attack and believe that they both attack, since they also both attack at every world accessible to either of them from w_{2b1}.

$$\mathcal{M}, w_{2b1} \vDash Attack(NS) \wedge M^1(Attack(NS)).$$

Since this conjunction is true, $Rat(NS)$ is also true at w_{2b1}: both agents are rational. But the agents do not commonly believe that they attack; in fact, they do not mutually believe^2 that they attack, since w_{2aX} is accessible for S, and at w_{2aX} (as we saw) N does not believe that S attacks. So S does not believe that N believes that they attack. This is consistent with Theorem 2.1 because the agents do not commonly believe that they are rational. In fact they do not even mutually believe that they are rational, precisely because they do not mutually believe^2 that they attack:

$$\mathcal{M}, w_{2b1} \vDash \neg M^2(Attack(NS)) \wedge \neg M^1(Rat(NS))$$

This failure of mutual belief in rationality is stark. Plausibly the generals at least believe that [each of them will attack only if he believes the other will]. But as we move further to the left in the model, mutual belief^n fails only for greater n, and the failure of common belief becomes more plausible. Here, I simply summarize some key facts (it is easy to check that $Trans(NS)$ is true at every world in the model so that $Trans(NS)$ is also true at every world):

$$\mathcal{M}, w_{2a1} \vDash Attack(NS) \wedge M^2(Attack(NS)) \wedge \neg M^3(Attack(NS))$$
$$\mathcal{M}, w_{2a1} \vDash Rat(NS) \wedge M^1(Rat(NS)) \wedge \neg M^2(Rat(NS))$$
$$\mathcal{M}, w_1 \vDash Attack(NS) \wedge M^3(Attack(NS)) \wedge \neg M^4(Attack(NS))$$
$$\mathcal{M}, w_1 \vDash Rat(NS) \wedge M^2(Rat(NS)) \wedge \neg M^3(Rat(NS))$$

The model thus shows that if the generals do not commonly believe one another to be rational, they can be rational in attacking, without having common belief that they attack.

To bring this model to life, suppose that the generals have never met each other. Prior to their hilltop correspondence, each general takes seriously the possibility that the other general is a fanatic who will attack without any regard for the catastrophe that might befall his soldiers. They also each take seriously the possibility that the other takes seriously the possibility that they themselves are fanatics, and so on.

Instead of or in addition to sending messages which describe their intended actions, the generals send notes about their own motivations. Suppose that each general has sent and received one message saying "I will attack only if I believe that you will". Then the generals have mutual belief^2 that they are rational, but they do not have mutual belief^3 that they are rational. In some ways of spelling out this case, they could attack, and be rational in doing so, just as at w_1. It is striking that they could rationally coordinate in this version of the case, even though they could not rationally attack if they had common knowledge that they were rational.

People in everyday situations are plausibly not unlike the generals in this version of the scenario. They may believe that others will act rationally, and even believe that others believe that they themselves will act rationally. But for some n, they may not mutually believe^n that everyone involved is rational.

Denying common belief of rationality allows us to reject the first claim of the inconsistent triad: that the agents can coordinate only if they commonly believe that they will. There are further ways of rejecting this claim by denying other assumptions about what the agents commonly believe. For example, Theorem 2.1 (implicitly) relies on the assumption of common knowledge that the agents' beliefs are closed under conjunction. Denying this common knowledge assumption would also be sufficient to escape the paradox.[12] In this particular case, it seems to me a natural idealization to assume that the generals' beliefs are closed under conjunction. But throughout the paper I want to remain officially neutral about which assumption of common knowledge is to blame. My aim is to argue that we should take the paradoxes of common knowledge to show that one of the relevant background common knowledge assumptions is false. Although my target is this general claim, I will continue to focus on denying common knowledge of rationality to illustrate it.

4. Electronic Mail Game

The second paradox of common knowledge I will consider is the electronic mail game. One way to think of the game is as providing a detailed specification of probabilities and utilities in an example closely related to the coordinated attack scenario. Once these probabilities and utilities are added, the purely logical argument of the coordinated attack scenario can be reframed in terms of the standard decision-theoretic notion of subjective expected utility maximization.

In the electronic mail game two players, Row and Column, are uncertain which of two coordination games, G_A or G_B, represent the payoffs to their actions (see Figure 4.1). In G_A, the players each have a strictly dominant action (A) which ensures a payoff of 0. In G_B, they receive a payoff of 1 if they coordinate on the best option (both playing B), and a payoff of 0 if they coordinate on the safe option (both playing A). In this game, if only one player tries for the better option, that player pays a penalty of 2. The names of the games are related to the actions which are best in them, as a mnemonic device: G_A is the game in which coordinating on A is best, G_B the game in which coordinating on B is best.[13]

		Column A	Column B
Row	A	$0,0$	$0,-2$
	B	$-2,0$	$-2,-2$

G_A

		Column A	Column B
Row	A	$0,0$	$0,-2$
	B	$-2,0$	$1,1$

G_B

Figure 4.1. The Electronic Mail Game

The players are commonly certain that the games are selected by the toss of a fair coin: each is chosen with probability $p = \frac{1}{2}$. But their knowledge of the game is asymmetric in one important respect. Row alone will be informed of the true game; Column will learn the game only by receiving a message from Row. The communication is structured as follows. Both players have a computer terminal before them: if the game is G_B, Row's computer will send a message to Column's computer. If either player's computer receives a message, it automatically replies with a new message. But each message has a positive, equal and independent rate of failure (ϵ). Since the rate is positive and independent, the process of automatic replies terminates almost surely. When it is over, each player's monitor will display the number of messages he or she has received and sent.

	$(0,0)$	$(1,0)$	$(1,1)$	$(2,1)$	$(2,2)$	$(3,2)$	$(3,3)$
Row	$\textcircled{1}$	$\frac{1}{2-\epsilon}$ $\frac{1-\epsilon}{2-\epsilon}$		$\frac{1}{2-\epsilon}$ $\frac{1-\epsilon}{2-\epsilon}$		$\ldots$	$\ldots$
Column	$\frac{1}{1+\epsilon}$ $\frac{\epsilon}{1+\epsilon}$		$\frac{1}{2-\epsilon}$ $\frac{1-\epsilon}{2-\epsilon}$		$\frac{1}{2-\epsilon}$ $\frac{1-\epsilon}{2-\epsilon}$		$\ldots$
Prior	$\frac{1}{2}$	$\frac{\epsilon}{2}$	$\frac{\epsilon\cdot(1-\epsilon)}{2}$	$\frac{\epsilon\cdot(1-\epsilon)^2}{2}$	$\frac{\epsilon\cdot(1-\epsilon)^3}{2}$	$\frac{\epsilon\cdot(1-\epsilon)^4}{2}$	$\ldots$

Figure 4.2. Certainty and Probability in the Electronic Mail Game

Figure 4.2 represents the information structure of the game. Here, the "worlds" are described by pairs representing the number of messages sent by Row, and the number of messages sent by Column. Thus for example $(0, 0)$ indicates that Row and Column have each sent 0 messages, whereas $(2, 1)$ indicates that Row sent 2 messages, while Column has sent only 1. Row sends 0 messages if and only if the game is G_A, so $(0, 0)$ represents the game being G_A. Every other state occurs only if the game is G_B.

The boxes in the second and third rows of the diagram represent what the agents are certain of, given that the relevant numbers of messages have been passed. If an agent has sent n messages, she is assumed to be certain of the event that she has sent n messages. Thus, for example, if the state is $(0, 0)$, Row is certain that he

has sent 0 messages; in this case he is certain that the state is $(0, 0)$ (and thus that the game is G_A). In that same state, however, Column is certain only that she sent 0 messages; since she is uncertain how many messages Row sent, she is uncertain whether the true message counts are $(0, 0)$ or $(1, 0)$.

The agents are assumed to update by conditionalizing the objective prior probabilities (specified in the last line of the diagram) on what they are certain of. The fractions inside the boxes in the diagram represent the probabilities of each state occurring, given what the agent is certain of at that state. Thus for example, in the state $(0, 0)$, Column assigns probability $\frac{1}{1+\epsilon}$ to the state being one where Row didn't send a message at all. With probability $\frac{\epsilon}{1+\epsilon}$, she thinks Row did send a message but it failed to get through. For higher message counts, the players are uncertain whether communication ended with the message they sent, or the message the other player sent. If the message count is $(2, 1)$, for example, Row is uncertain whether his message failed on the way to Column as in the true state $(2, 1)$, or whether it got to Column but her message failed on the way back, as in the state $(2, 2)$. The probability that it failed on the way there is $\frac{1}{2-\epsilon}$; the probability that it failed on the way back is $\frac{1-\epsilon}{2-\epsilon}$.

For each agent, a strategy is a function from the number of messages that agent has sent to probability distributions over actions (the set $\{A, B\}$). A strategy is rational for an agent i if and only if for every n, the strategy maximizes expected utility given the probability distribution obtained by conditionalizing the prior on the event that i sends n messages.[14] A player is rational if and only if she plays a rational strategy. A strategy is rationalizable if and only if it is consistent with the players' having common certainty that they are rational. We then have the following result:

Theorem 4.1 Rubinstein (1989)**.**

For each player, the unique rationalizable strategy is the constant function which takes every number of messages sent to A.

The full proof can be found in many places; I'll just describe what happens in a few cases to give a sense for how the argument goes. For simplicity, I'll discuss only "pure" strategies, where the probability assignments to actions assign an action either 1 or 0, and I'll assume a specific value of ϵ, $\frac{1}{10}$.

If Row sends 0 messages, then the game is G_A, and he's certain the game is G_A, so if he is rational, he must play A. A is the strictly dominant action: it's better no matter what he thinks Column will do. If Column sends 0 messages, on the other hand, then she is certain that the state is either $(0, 0)$ or $(1, 0)$; she assigns $\frac{10}{11}$ to $(0, 0)$ and $\frac{1}{11}$ to $(1, 0)$. Since she is certain that Row is rational, she is certain he plays A in $(0, 0)$. So even if Column were somehow certain that Row would play B in $(1, 0)$, if Column plays B her expectation would be $-\frac{19}{11}$.[15] If she chooses A, by contrast, she can guarantee herself $0 > -\frac{19}{11}$. So Row will play A if he sends no messages, and Column, too, will play A if she sends no messages.

In the full proof, this is the base case of an induction. But we can get a sense for how the induction goes by considering just the next case. Given that Row is certain that Column is rational, and certain that Column is certain that Row is rational,

Row will be certain that Column will play A if Column sends 0 messages. That's what the base case shows. But now supposing that Row has sent only one message, Row will be certain that the state is either (1, 0) or (1, 1); he assigns $\frac{10}{19}$ to (1, 0) and $\frac{9}{19}$ to (1, 1). But he's certain that if the state is (1, 0), Column will play A, and this is already enough to make him play A, too. For even if he were somehow certain that in state (1, 1) Column would play B, his own expectation from playing B would be $-\frac{11}{19}$.[16] Once again, this is less than the 0 he is guaranteed by playing A.

The same reasoning applies to Column if *she* sends only only message. It can also be extended to higher message counts. Given that the players are commonly certain that they are rational, they can always deduce that [if the other person has sent $n-1$ messages, the other will play A]. Moreover, if the players are certain that [if the other person has sent $n-1$ messages, the other will play A], then it always makes sense for them to play A if they themselves have sent n messages. So they play A regardless of how many messages are sent.

This result is extremely surprising. After one message, the players are *certain* that the game is G_B. But it turns out that they are also certain that the other player will *not* play B. If they send two messages, they are not just certain that the game is G_B, they are also certain that the other is certain of this. But they remain certain that neither will play B. And so on for more messages. This is bizarre: if you are certain that the game is G_B, and certain that others are certain of this, and certain that they are certain that you are certain of this, and so on, it seems clear that it should at least be permissible to play B.[17] Rubinstein himself agreed that the result is highly counterintuitive (Rubinstein (1989, p. 389)). His judgment has been confirmed by empirical studies, which show that on their first exposure to close relatives of this game, people tend to play B with comparatively high probability already after only a few messages are sent (Camerer (2003, p. 226–232), Heinemann *et al.* (2004), Kneeland (2016), Thomas *et al.* (2014)).

This is the second paradox of common knowledge. Rational agents who are commonly certain that they are rational will play A invariably. But there is a powerful intuition that it could be rational to play B.

5. Common Knowledge of Rationality Part II

As I will now show, the result that agents cannot rationally coordinate on playing B in the electronic mail game depends on background assumptions of common knowledge (this time, of common certainty). This formal fact is even less surprising than the analogous fact was in the coordinated attack scenario. The whole point of the electronic mail game is to dramatize the role of common certainty of rationality in strategic reasoning. But the conceptual point I wish to make—that the electronic mail game can be understood as an argument against common knowledge assumptions—has not been sufficiently widely appreciated. And to make this conceptual point it will be helpful to have a concrete sense of how exactly relaxing this assumption allows us to escape the paradox.

In the model I'll present, we think of each player as being of a particular "type". A type encodes all of the relevant information about a person (it is what type of

person a person is). Every type is associated with (1) a strategy (a function from messages sent to probabilities over actions); and (2) a distribution over other types. The players may be uncertain what type of person they're playing against; they each have a probability distribution over the Cartesian product of the set of types and the set of pairs of message counts.

The basic idea of the model is for there to be an irrational type, who plays B even when this is not justified by subjective expected utility maximization. This "seed" of irrationality allows even rational types to play B: if a rational type is certain she is playing against an irrational type who plays B regardless of how many messages are sent, then at least if the rational type sends one message, it will make sense for her to play B. This fact in turn allows rational types who are certain that they are playing against rational types also to play B. For suppose a rational type t_1 is certain she is playing against a rational type t_2 who is certain he is playing against an irrational type t_3 who plays B regardless. Then t_1 should play B if he sends two messages. For in this case, he'll be certain that t_2 sent one message, and thus certain that t_2 (rationally) plays B. This basic idea is structurally parallel to the model given in the case of coordinated attack: the irrationality of N at w_{2aX} allowed rational players at other worlds to attack. The discussion which follows merely shows how this basic thought can be spelled out precisely.

I will describe a simple model where there are only two types for each player, t_1^R, t_2^R and t_1^C, t_2^C, and as before ϵ is assumed to be $\frac{1}{10}$. Each type's beliefs about what type the other person is (her marginal distribution over other types) are described in the following pair of tables.[18] The type whose distribution is described is listed on the left of the table; the type it is assigning probability to is listed on the top of the table. Thus for example the Row type t_1^R assigns Column's type t_1^C probability 1.

	t_1^C	t_2^C
t_1^R	1	0
t_2^R	.6	.4

	t_1^R	t_2^R
t_1^C	.6	.4
t_2^C	0	1

The determination of each player's type is assumed not to depend probabilistically on how many messages are sent. We also assume that each type responds to the message count by conditionalizing the prior, so that the players' probabilistic beliefs about message counts (marginal distributions on message counts) are just as in Figure 4.2. The strategies of the types are described as follows, where the columns indicate the number of messages sent by the player in question.

	0	1	$n \geqslant 2$
$s(t_1^R)$	A	B	B
$s(t_2^R)$	A	A	B

	0	1	$n \geqslant 2$
$s(t_1^C)$	A	B	B
$s(t_2^C)$	A	A	B

Row's type t_1^R is thus the seed irrational type. If Row is t_1^R, he is certain that Column will play A if she does not receive any messages, but he nevertheless plays

B when he sends only a single message, expecting $-\frac{11}{19}$,[19] which is less than the 0 he would be guaranteed if he played A.

The rational type t_1^C of Column assigns this irrational type of Row sufficiently high probability that it too can play B. If Column is t_1^C and sends one message, then since she assigns .6 to Row's being t_1^R, she assigns at least .6 to Row's playing B. In fact she also assigns additional probability to Row's playing B: if he is t_2^R and sends two messages (that is, if Column's message has gotten through, but the second of Row's messages did not make it back to her), then he will also play B. Given the total probability she assigns to Row playing B, Column herself expects $\frac{7}{19}$ by playing B, which is greater than the 0 she would get by playing A.[20] So, as I claimed, this type of Column is rational in playing B.

Other types, too, can now rationally play B. Both t_2^R and t_2^C will rationally play B if they send two or more messages, as the reader may easily verify. These types can be rational in playing B precisely because they are not commonly certain that they are rational. If the types are in fact t_2^R and t_2^C, the players are mutually certain that they are rational, but mutual certainty still fails at a low level: they are not mutually certain^2 that they are rational. As before, however, it is easy to see that for any n, the model could be extended so that there would be types which have mutual certainty^n of one another's rationality, but nevertheless coordinate rationally. So long as the agents are not required to be commonly certain that they are rational, they can coordinate after finitely many messages are passed.[21]

My aim has been to show that denying some common knowledge assumptions allows for a resolution of the paradoxes.[22] Officially, I have used common knowledge of rationality only to illustrate this general point. But there is in fact evidence that this common knowledge assumption in particular should be rejected, at least in some applications. The currently most popular approach to explaining observed behavior in the coordinated attack scenario and the electronic mail game uses models of "limited reasoning", known as "level-k" models. In these models, each agent is assigned to a "level": if an agent's level is k, the agent can perform at most k steps of iterated deletion of dominated strategies. Although the details of these models can vary, the general pattern is to assume that certain types are not rational in the sense of maximizing expected utility, and as a result, that none of the types participates in common knowledge that they are playing against a rational opponent. The success of these models in explaining the data may suggest that people do not have common knowledge of one another's rationality; the models do not seem to provide evidence concerning whether other common knowledge assumptions are true or false.[23]

6. Conclusion

A prominent interpretation of the paradoxes of common knowledge draws a quite different conclusion from the one I have argued for in this paper. On this interpretation, the paradoxes are taken to be arguments for the claim that common knowledge is "needed" to explain rational action. The textbook of Fagin *et al.* (1995) is a representative example. Although they recognize that the results are paradoxical—the

authors introduced the term "paradox of common knowledge"—nevertheless they claim to "show that common knowledge is a necessary and sometimes even sufficient condition for reaching agreement and for coordinating actions" (1995, p. 198). Later, they summarize their discussion of the paradoxes by recalling to the reader that common knowledge "can be shown to be a prerequisite for day-to-day activities of coordination and agreement" (1995, p. 454). In context, these claims are best understood as restricted to the classes of models employed in proving the relevant mathematical results. But remarks such as these have contributed to a general attitude that the coordinated attack scenario and the electronic mail game demonstrate that common knowledge is needed to explain rational coordination, and, thus, to explain social behavior more generally.[24]

The mathematical results based on the coordinated attack scenario and the electronic mail game do not, however, establish these unqualified claims about the relationship between coordination and common knowledge. Rather, they establish conditionals—that *if* agents have common knowledge that they are rational (as well as common knowledge of other background facts), *then* they will coordinate only if they commonly know that they will. In these examples, the relevant background assumptions of common knowledge lead to *counterintuitive and implausible* predictions about rational behavior. In these cases we must reject the predictions about behavior given in the consequents of the relevant conditionals. It follows, then, that we must also reject the antecedents of these conditionals, that is, the background assumptions about common knowledge.[25]

If we take this view of the paradoxes seriously, it becomes natural to see common knowledge of rationality as just a simplifying technical assumption, which is useful because it yields tractable models and rich predictions about behavior. In closing, I want to support this view of common knowledge of rationality by showing how it fits into a more general story about rationality in strategic situations and its relation to decision theoretic rationality.

Some of the founders of modern game theory, most notably Nash, are associated with the idea that game-theoretic rationality is a different beast altogether from decision-theoretic rationality.[26] These authors are supposed to have claimed that players act rationally in a game only if their actions result in an equilibrium. Since one player cannot control what other players do, this theory of rational play represents a sharp departure from the usual decision-theoretic notion of rationality, since it puts the rationality of a given player's action out of the control of that player. On the standard theory that agents act rationally if they choose an act which maximizes subjective expected utility, for example, what it is rational for an agent to do depends on what the agent thinks about the world; it does not depend on how the world in fact is. A person may lose by choosing an action which maximizes expected utility relative to her beliefs, but that is just bad luck; it does not call into question the rationality (in this sense of "rationality") of her choice. Decision-theoretic rationality is supposed not to be hostage to the whimsy of the world. It is thus natural to suppose that a subject's game-theoretic rationality should not be hostage to the whimsy of others' play.[27]

This is of course not to say that equilibrium is uninteresting. Equilibrium has many important justifications—for example, particular equilibrium notions can be justified from the perspective of evolutionary biology. But from the perspective of the individual choice of rational agents in social situations, equilibrium is unnatural. Non-equilibrium concepts such as rationalizability offer alternative theories of play which do not make the same unnatural demands on the players. A strategy in a given game is rationalizable just in case it can be played by a player who participates in common certainty that all players are rational. In standard models of belief, this is equivalent to saying that a strategy is rationalizable just in case it can be played by a rational player who is certain that all players have common certainty that they are rational. The criterion that a strategy be rationalizable combines decision-theoretic rationality with restrictions on players' beliefs about the game and each other. It is thus much closer to a theory which uses only decision-theoretic notions of rationality. The rationalizability of an individual's action depends only on what that individual thinks; it does not depend on what others happen to think or do.

But while rationalizability represents a step in the right direction, it does not go far enough. Decision-theoretic notions of rationality generally provide conditional recommendations for action: if one's beliefs are thus-and-so, one should choose thus-and-such an act. These notions of rationality impose minimal restrictions on what agents must be certain of in order to count as rational. For example, it is no part of decision-theoretic rationality that a rational agent be certain of the laws of physics. But once we have this point clearly in view, we can see that there is no *decision-theoretic* justification for the key constraint imposed by rationalizability – that agents are certain that they are commonly certain that they are rational. Standard decision theories do grant a special place to the laws of logic; agents are assumed to be certain of all propositional tautologies, as a consequence of the axioms of the probability calculus. But the fact that some agent other than myself is decision theoretically rational is surely quite different from the laws of logic. Any particular other person could have failed to be rational; one can't deduce that others are rational from any *a priori* axiomatic system. A truly decision-theoretic perspective on game-theoretic rationality would thus not require that agents be certain of one another's rationality, never mind commonly certain of one another's rationality. Even if we are only interested in the pure theory of rational agents interacting with one another, we should take a broader view of what agents can believe about one another while remaining rational. In particular, we must study what happens when they take seriously the possibility that others are irrational.[28]

The idea that common knowledge of rationality is a simplifying technical assumption also sheds new light on the paradoxes themselves. In simple models used in classical physics, objects are treated as point masses. In standard applications, the assumption that the objects are point-like is false. But under a range of well-understood conditions, the models make predictions which are equivalent to the predictions of more complex, accurate models. Similarly, the assumption of common knowledge of rationality is false in general. But there are conditions under which the behavioral predictions of models which assume common knowledge of rationality coincide with those of less tractable, more realistic models. The two

paradoxes studied in this paper show that at least in some cases, however, the assumption of common knowledge of rationality leads to false predictions. The paradoxes of common knowledge thus play a role similar to that of examples of moving objects which cannot be approximated by point masses in classical physics. In each case useful technical assumptions are shown to be responsible for false predictions.

The analogy also helps to make a second, subtler point. It is a mathematical fact that in the models of classical physics just mentioned objects have no spatiotemporal extension. But one doesn't conclude, even in the good cases where these idealized models give the right results about objects' motion, that real objects in fact have no extension. Similarly, even when the predictions of simplified models which assume common knowledge are right about what people will do, we shouldn't conclude that as a matter of fact the people do have common knowledge of one another's rationality. Relatedly, we shouldn't infer from the fact that in the models coordination requires common knowledge, that people, too, need common knowledge to coordinate. The mathematical fact about the models is a direct consequence of simplifying technical assumptions; it does not hold if we work with more complex, realistic models. To claim that everyday coordination requires common knowledge is to mistake a simplifying technical assumption for a truth about human psychology or rational agency. The mistake is not far from the error an engineer would commit if he took mathematical models of point-like objects to be evidence that bridges, too, have no spatial extension.

A. Appendix

A.1 Logic

The set of theorems of the logic $\mathbf{K_{NSC}}$ is the smallest set containing every instance of the following axiom schemata, and closed under the rules. As usual $\vdash_{\mathbf{K_{NSC}}} \varphi$ means that φ is in the set of theorems; for Γ a set of formulas, we write $\Gamma \vdash_{\mathbf{K_{NSC}}} \varphi$ to mean that there are formulas $\psi_1, \ldots \psi_n \in \Gamma$ such that $\vdash_{\mathbf{K_{NSC}}} (\psi_1 \wedge \ldots \wedge \psi_n) \rightarrow \varphi$ (see Chellas (1980, p. 47)). In the following $\Box$ is schematic for the monadic sentential operators B_N, B_S and C:

(PL)	Any substitution instance of a theorem of propositional logic.
(K)	$\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$
$(C-out)$	$C\varphi \rightarrow M^n\varphi$.
(MP)	If $\vdash_{\mathbf{K_{NSC}}} \varphi, \varphi \rightarrow \psi$ then $\vdash_{\mathbf{K_{NSC}}} \psi$.
(N)	If $\vdash_{\mathbf{K_{NSC}}} \varphi$ then $\vdash_{\mathbf{K_{NSC}}} \Box\varphi$.
$(C-in)$	If for all $n \in \mathbb{N}$, $\Gamma \vdash_{\mathbf{K_{NSC}}} \varphi \rightarrow M^n\psi$ then $\Gamma \vdash_{\mathbf{K_{NSC}}} \varphi \rightarrow C\psi$

A.2 Sketch of Proof of Theorem 2.1

The proof of theorem 2.1 is by induction. Here is a sketch of the strategy: first note that

$$\vdash_{\mathbf{K_{NSC}}} Ideal(NS) \rightarrow (Attack(NS) \rightarrow M^1(Attack(NS))).$$

An application of (N) and two instances of (K) allow us also to establish:

$$\vdash_{\mathbf{K_{NSC}}} M^1(Ideal(NS)) \rightarrow (M^1(Attack(NS)) \rightarrow M^2(Attack(NS))),$$

which together with the previous means that

$$Ideal(NS),\ M^1(Ideal(NS)) \vdash_{\mathbf{K_{NSC}}} Attack(NS) \rightarrow M^2(Attack(NS)).$$

More generally it is now routine to show that for all n:

$$\{Ideal(NS)\} \cup \{M^k(Ideal(NS)) | 1 \leqslant k \leqslant n\} \vdash_{\mathbf{K_{NSC}}} Attack(NS) \rightarrow M^{n+1}(Attack(NS))$$

which suffices, given $(C - in)$, for the result.

A.3 Knowledge

If we add the axiom schema $B_i\varphi \rightarrow \varphi$ to the above collection of axioms and close under the rules, the resulting stronger logic would also prove the claim. Thus if one replaces "believe" and its cognates above with the word "know" and its cognates, the result can be interpreted as establishing that coordination requires common knowledge (as opposed to common belief). But on this interpretation of the modal operator, the principle 1 will seem to many not to deserve its name. Even if a general must believe the other will attack in order for his attack to be rational, he needn't *know* that the other will attack to be rational in attacking. Since the position that knowledge is what is important in assessing rational action is a minority position (albeit an influential minority, see Williamson (2000); Hawthorne & Stanley (2008)), I've focused on belief in the main text.

A.4 Proof of a Strengthening of Proposition 3.1

If we add the following schemas and close under the rules of $\mathbf{K}_{NSC}$, the logic of each belief operator taken on its own is the system **S5**:

(T) $B_i\varphi \rightarrow \varphi$
(5) $\neg B_i\varphi \rightarrow B_i\neg B_i\varphi$

We write the consequence relation defined in this way as $\vdash_{\mathbf{S5_{NSC}}}$. Since the accessibility relations in the model in Section 3 are equivalence relations, and since $\mathbf{S5_{NSC}}$ is easily shown to be sound on such models, in fact I will establish that for every n,

Proposition A.1.

$$Ideal(NS),\ C(Trans(NS)),\ M^n(Rat(NS) \nvdash_{\mathbf{S5_{NSC}}} Attack(NS) \rightarrow C(Attack(NS)).$$

The model from Section 3 can be extended to prove Proposition 3.1, giving models and worlds in which agents coordinate, and do not commonly believe that they coordinate, but have mutual belief^n that $Rat(NS)$. To produce models of this form, we add worlds $w_{2a2}, w_{2b2}, w_{2a3} \ldots w_{2bN}$ "between" w_{2b1} and w_{2aX}, where each w_{2an} (and each w_{2bn}) has the same relationship to its neighbors as w_{2a1} (respectively

w_{2b1}) does. Adding m such worlds makes it so that, at w_1, the agents mutually believe^{m-1} that (each acts only if he or she believes the other does). They would also mutually believe^m that they are coordinating.

Notes

[1] This is a much modified and redacted version of Fagin *et al.* (1995, p. 190–1 cf. Chapters 6 and 11). The authors first formulated the idea that these results are "paradoxes of common knowledge" (Moses (1986) and Halpern & Moses (1990); cf. Fagin *et al.* (1999)). To my knowledge the first use of the name "coordinated attack" was in Gray (1978, p. 465), but Gray's result does not employ a formalization of knowledge and belief, and so does not represent the presence or absence of common knowledge. Arguments similar to Gray's were given apparently independently by Akkoyunlu *et al.* (1975, p. 73–4), and Cohen & Yemini (1979), Yemini & Cohen (1979).

[2] Some authors use "common knowledge" ambiguously, both for the technical notion defined in the main text, and also to describe an informal notion of "public information", an idea that is typically introduced by way of examples. In this paper, I'll reserve "common knowledge" and "common belief" (as well as "common certainty") for notions which are introduced by definitions such as those just given; when I want a term for the informal notion, I'll use the phrase "public information". I will however often follow the standard practice in game theory of using "common knowledge" to mean "common knowledge or common belief or common certainty". The paradoxes of my title are best taken to be paradoxes of common knowledge in this sense of the term.

[3] In Lederman (forthcoming) I argue that people never have common knowledge or any approximation of it. In the concluding section of that paper, I provide a theory of how people can coordinate rationally without common knowledge, on the basis of past experience of others' actions in relevantly similar situations. The theory there does not appeal to any common knowledge in explaining coordination. In this paper, I argue that a standard package of common knowledge assumptions gives rise to false predictions both about what it is rational to do, and about what people will in fact do. As far as this paper is concerned, it may be that some of these common knowledge assumptions are true, even though all of them together cannot be. The arguments and positions of the two papers thus may be taken on their own, but they also support each other by elucidating complementary problems with various common knowledge assumptions.

[4] In perhaps the most common cases of coordination in everyday life, if one person sends a message to another, the sender knows that the receiver has received the message. In this regard, the cases are disanalogous to the coordinated attack scenario. But in other cases, people don't know whether a message they send will be received, and even then after a few messages people will typically coordinate. The behavior in these latter cases still seems familiar enough to undermine the claim that not attacking is rationally mandatory. Thanks to Ginger Schultheis here.

[5] I've used "know" instead of "believe" in this description to avoid confusion arising from "neg-raising", the fact that $\ulcorner S$ doesn't believe that $p\urcorner$ is naturally interpreted as $\ulcorner S$ believes that $\neg p\urcorner$.

[6] The reply also threatens to undermine some of the most prominent arguments for the importance of common knowledge (Heal (1978) (now elegantly presented by Greco (2015)), Clark & Marshall (1981), Fagin *et al.* (1995), among many others): if one attempted to save common knowledge assumptions from their paradoxical consequences by blaming the message-passing setup, one would thereby lose some important motivations for the importance of common knowledge in the first place. In these prominent arguments, a paradigm example of public information – such as a public announcement of some claim – is contrasted with a series of cases of private information (that is, information which is not public, for example, secrets told privately to each individual). The series of cases of private information is constructed so that it is plausible that the agents do not have common knowledge of the relevant claim, but they do know that they know... that they know it, where for every natural number n, there is some case in the series such that "..." can be replaced with n occurrences of "that they know". It is claimed in these arguments that there is no n such that the subjects have public information in the n^{th} private information case in the same way as they did in the original example of the public announcement; common knowledge is thus claimed to be needed to explain the publicity created by

the public announcement. But if seventeen messages are sufficient for the achievement of common knowledge, then these standard arguments could no longer be used to isolate a distinctive role for common knowledge, since the agents would have common knowledge in the 17^{th} case in the series of private information cases.

[7] Although the latter may not make the point since "think", like "believe", is subject to neg-raising, see n. 5.

[8] Some might see this as a flaw of the argument based on coordinated attack. If only we described the details of the generals' probabilities and utilities, they might contend, it would be clear which assumption to reject, and the paradox would disappear. This response could help to eliminate the paradox only if belief is identical to some state specified purely in terms of degrees of confidence (for example, confidence above some threshold t). For if belief is not identical some such state, then since rationality imposes constraints directly on the relation between belief and action, it is irrelevant how we answer further questions of what degrees of confidence the agents may have. In this paper, I wish to remain neutral on how belief and degrees of confidence are related. But for the purpose of presenting this first paradox I'll continue to assume that belief cannot be reduced to a state specified purely in terms of confidences. My second paradox, the electronic mail game, will be formulated using probabilities and utilities directly, and so won't raise this kind of concern. Thanks to an anonymous reviewer for pressing me to elaborate on this point.

[9] Mathematically, this is essentially Halpern & Moses (1990, Propositon 4) (cf. Theorem 5 and Corollary 6, together with Fagin *et al.* (1995, Theorem 6.1.1 and Ch. 11)): see n. 11 for further discussion. A sketch of the proof is given in the appendix, where I also discuss whether the result can be developed for knowledge, as opposed to belief. I take the argument from coordinated attack to be based on this result. What Chant & Ernst (2008) call "coordinated attack" is what I call "the electronic mail game". It is unclear what Chant and Ernst's basis is for calling the argument they describe the argument from coordinated attack; oddly one of the main citations they give for their argument, Halpern (1986), appears to prove neither theorem, and discusses coordinated attack only in passing. It's clear that the results *are* different: one employs probabilities and utilities, while the other doesn't. But this is just a terminological point, and won't matter to any of the substance here.

[10] In fact I will show something substantially stronger; see appendix A.4.

[11] Earlier, I presented the coordinated attack scenario (and Theorem 2.1) using the proof system of a formal logic. More standard presentations of the result directly use models of the kind just described. These model-theoretic presentations might seem to lead to a more powerful result. They begin by assuming that $Ideal(NS)$ is true at every world in every model under consideration (it is *valid* on the class of models). This one assumption can then be used to show that the conclusion of Theorem 2.1 is also valid on the class of models. But this model-theoretic version of the result conceals the strength of its assumptions. In the models just described, if φ is valid on a class of models (as opposed to just true at some world of interest), then $C\varphi$ is also valid on the class of models. Thus if $Ideal(NS)$ is valid, $C(Ideal(NS))$ is also valid.

Actually something even stronger than the validity of $Ideal(NS)$ is assumed in the classic presentation of Halpern & Moses (1990). Halpern and Moses describe "protocols" which guarantee coordination in the coordinated attack scenario. The assumption that coordination is guaranteed corresponds to the assumption that each agent attacks if and only if the other does. Model-theoretically, they assume essentially that $v(Attack(N)) = v(Attack(S))$, that is, that $Attack(N) \leftrightarrow Attack(S)$ is valid. This assumption may be well motivated in their setting, but in the present context, where the focus is rational action, it is an unnatural starting assumption. For it would imply the validity of $C(Attack(i) \rightarrow Attack(NS))$: the agents would commonly believe that if one of them attacked, they both would. But this intuitively is not what they believe in my presentation of the case. It is precisely because each of them does not believe that if he himself attacks the other will too that it is difficult to decide what to do; even if they do come to believe that they will both attack after deliberation, it is implausible that it becomes common belief.

[12] In fact even if each agent's beliefs are assumed to satisfy the laws of the logic **EMD45** for belief (see, e.g. Chellas (1980)), we can also show an analogue of proposition 3.1. In this setting, various different semantic clauses for the operator C, which coincide when individual belief satisfies the logic

K, are no longer equivalent (see e.g. Lismont & Mongin (2003)). But for any of the standard semantics we can give models demonstrating:

$$Ideal(NS),\ C(Ideal(NS)) \nvdash_{\mathbf{EMD45_{NS}}} Attack(NS) \rightarrow C(Attack(NS)).$$

Models can even be given witnessing this claim where agents satisfy this strong logic for belief, as well as the logic **S5** for certainty, and where their beliefs are derived by conditionalizing on a common prior. For reasons of space I leave the description of such models to the reader.

[13] I present a variant of the original electronic mail game, which appeared first (to my knowledge) in Morris (2002), but which is now also used by Strzalecki (2014). In Rubinstein's original game, there is a unique Nash Equilibrium *where player 1 plays A in* G_A, but there is also a Nash equilibrium where the players play B regardless of what has happened. In the game above, there is only one Nash Equilibrium. In fact, whether we use a definition of rationality as *ex ante* – before the messages have been sent – or *ex post* – after the agents have updated on the number of messages sent – there is a single strategy profile which is consistent with common knowledge of rationality: that the agents play A regardless of the number of messages sent. This means that the result in this variant is formally harder to escape than Rubinstein's original. In that game, the payoffs were as follows, with $L > M > 0$, and where it is assumed that G_A occurs with probability $p > 1/2$:

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

G_A

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

G_B

[14] A different definition of rationality (*ex ante*) would simply say that the function from information states to actions maximizes expected utility in the prior. Both definitions lead to the result below: since every information state has positive probability in the prior, the only rationalizable strategies would still be those which play A forever.

[15] $-2 \cdot \frac{10}{11} + 1 \cdot \frac{1}{11}$

[16] $-2 \cdot \frac{10}{19} + 1 \cdot \frac{9}{19}$

[17] The result also depends heavily on the fact that communication is automatic; see n. 22 for further discussion.

[18] I'll assume that each player is certain of her own type, and hence perfectly introspective: if she assigns an event probability p, she is certain that she assigns it probability p. This implies – but is not implied by – "positive" and "negative" introspection for each individual's certainties.

[19] $1 \cdot \frac{9}{19} - 2 \cdot \frac{10}{19}$

[20] $.6 + .4(-2\frac{1}{2-\epsilon} + \frac{1-\epsilon}{2-\epsilon}) = .6 + .4(\frac{-2+9/10}{19/10}) = \frac{6}{10} - \frac{44}{190} = \frac{7}{19}$.

[21] In the model in this section, each player can in principle reason about every order of beliefs; this makes the models quite different from the popular "level-k" models, in which players' reasoning is restricted to some finite depth. Strzalecki (2014) provides models of this kind for the electronic mail game; for related models see Kneeland (2016). More abstract models are given by Kets (2014). For a review of work on level-k models, see Costa-Gomes *et al.* (2013).

[22] For a different style of response, not focused on common knowledge, see Binmore & Samuelson (2001). Binmore and Samuelson build on Rubinstein's observation that if the number of messages sent is destined to be truncated (the most that could be sent is some finite n), then new equilibria, which allow A for low message counts and B for higher ones, appear. As Binmore and Samuelson emphasize, this displays a disanalogy between Rubinstein's setup and ordinary communicative situations, suggesting that the application of the argument may be quite specialized.

[23] Level-k models do better in the lab than models which assume common knowledge of rationality; classic studies are Nagel (1995), Stahl & Wilson (1994, 1995). An important recent study of a very different kind of game is Kneeland (2015); see also Arad & Rubinstein (2012). For more references see

above n. 21. Of course there isn't a deductive argument from the data in these studies to the claim that common knowledge of rationality is to blame, and further studies may point in a different direction. But they are highly suggestive in the present context. It's not just that these studies tell against the assumption of common knowledge of rationality; I'm not aware of any studies which suggest that the assumption of common knowledge of rationality is true. Some studies, e.g., Heinemann *et al.* (2004), Chaudhuri *et al.* (2009), Thomas *et al.* (2014) claim that common knowledge is relevant to their predictions. But on inspection, the hypothesis that people are sensitive to an informal notion of "public information" (which need not be analyzed as common knowledge in the technical sense) explains the data; nothing in the studies isolates the role of common knowledge as opposed to the more informal notion of public information.

[24] Cf. Halpern & Moses (1990, pp. 2, 19). Michael Chwe similarly writes "even narrowly rational Homo economicus when solving coordination problems must form common knowledge" (Chwe (2001, p. 96)). In a section of the appendix entitled "Why common knowledge is good for solving coordination problems", he illustrates his claim by describing an argument based on what is essentially the coordinated attack scenario. The point is made repeatedly in Heifetz (2004) "It is well known among Economists and Game Theorists that common knowledge... is essential for coordination..." See also Greco (2014a,b). In a slightly different vein, Stephen Morris suggests that the electronic mail game "provides a useful, if extreme illustration of the logic by which higher-order beliefs and knowledge might influence outcomes in strategic settings". Morris asks us to take seriously the possibility that the game gives us "a sensible starting point for modeling the role of higher-order beliefs in applied settings" (Morris (2002, p. 434)). From this perspective, too, the electronic mail game is not a paradox, but rather an illustration of important and prevalent psychological phenomena.

[25] There are other cases, in addition to coordinated attack and electronic mail, which have been thought to "require" common knowledge, for example puzzles such as "muddy children". In this example, some number of children have mud on their foreheads, and each wants to determine whether she herself has mud on her forehead. The children can't do this unless someone makes an announcement saying that at least one of them has mud on her forehead. Once the announcement is made – making it "common knowledge" among the children that one of them has mud on her forehead – if each child announces, in a series of rounds of public announcements, whether they have learned that they have mud on their foreheads, they can each figure out whether they are muddy or clean in a finite number of rounds.

I am unsure what the data here are supposed to be. Of course I recognize the mathematical fact that in order to explain how this could be solved for all finite n, we must hold that the announcement delivers mutual knowledge^n for all finite n. In other words, it gives the children common knowledge. But does this mathematical fact tell us anything about our psychology? Surely not. People in the "wild" will hardly ever solve such problems, especially if the number of people involved is large, for example, larger than 1,000. The fact that we "need" common knowledge to solve some puzzles in epistemic logic shows that common knowledge is useful in epistemic logic. But it does not show us anything about how people or even rational agents who do not have common knowledge that they reason appropriately will behave.

[26] See for example the survey article Brandenburger (2010).

[27] Versions of these ideas can be found in many, many places, and I have not made any attempt to assemble comprehensive references here. Notable early sources are Luce & Raiffa (1957, e.g. p. 306) and Robert Nozick's 1963 dissertation (published as Nozick (1990)). Other major works include: Harsanyi (1968, Section 15) (on "inconsistent game specifications") and Aumann (1974, 1987). Significant related conceptual discussion can be found in in Kadane & Larkey (1982a,b), Harsanyi (1982a,b) and Roth & Schoumaker (1983). Bacharach (1987) and Skyrms (1990) have fallen out of the mainstream of the technical literature, but are important works related to the general question. Recent expressions of related thoughts are Aumann & Drèze (2008) and the introduction of Brandenburger (2014). On Nash equilibrium in particular see Aumann & Brandenburger (1995) and Stalnaker (1994) (the former antedates the latter, and the two were not independent see Stalnaker (1994, n. 6)). Although my remarks in the main text may sound most similar to the received interpretation of Kadane & Larkey (1982b), see n. 28 for an important qualification.

[28] Let me distinguish the view in the main text from what has been called the "naïve Bayesian" view (often attributed to Kadane & Larkey (1982b), e.g. by Aumann & Drèze (2008, p. 81)). According to what I take to be this naïve view, (*i*) rational agents have a probability distribution over all events, and (*ii*) if an agent has a probability distribution he or she must (simply) maximize expected utility with respect to it. Let me here grant (*ii*): if an agent does have a probability distribution over others' actions, no distinctive game-theoretic considerations need be used in deciding what to do. It is still open that (*i*) is false, and that this could leave space for a way in which expected utility theory does not determine what a rational agent should do in every circumstance. For if a player does not have a probability distribution over other players' actions (the agent faces "Knightian uncertainty" or "ambiguity" concerning others' actions) strategic reasoning may help resolve this uncertainty. One's ability to resolve this Knightian uncertainty about others' actions may go beyond the norms (if any) which govern the resolution of other forms of Knightian uncertainty, for example, in simple decision problems that don't involve strategic reasoning. I take the standard criticism of the "naïve Bayesians" that they ignore how "game theoretic reasoning constrains priors" to be this very point: naïve Bayesians fail to take account of the ways in which game theoretic reasoning can constrain how Knightian uncertainty is resolved (Harsanyi (1982a, p. 120) and Binmore (1987, pp. 209–212)). In my view the interesting debate to be had between "naïve Bayesians" and theorists such as Harsanyi, Binmore and Aumann and Drèze, concerns (a) whether it makes sense for (idealized) agents to face Knightian uncertainty of the kind just described; and (b) whether game-theoretic reasoning adds anything (over and above the dictates of decision theory) to how one should choose actions when facing such uncertainty. The view in the main text is consistent with a huge range of answers to these questions. All it advocates on this count is that if we investigate the ways in which Knightian uncertainty can be resolved in the presence of common knowledge of rationality (and common knowledge of other background facts), then we should also be interested in the ways it can be resolved without common knowledge of rationality.

References

Akkoyunlu, Eralp A., Ekanadham, Kattamuri, & Huber, R. V. 1975. Some Constraints and Tradeoffs in the Design of Network Communications. *SIGOPS Operating Systems Review*, 9(5), 67–74.

Arad, Ayala, & Rubinstein, Ariel. 2012. The 11–20 Money Request Game: A Level-k Reasoning Study. *American Economic Review*, 102(7), 3561–73.

Aumann, Robert J. 1974. Subjectivity and Correlation in Randomized Strategies. *Journal of Mathematical Economics*, 1(1), 67–96.

———. 1987. Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, 55(1), 1–18.

Aumann, Robert J., & Brandenburger, Adam. 1995. Epistemic Conditions for Nash Equilibrium. *Econometrica*, 63(5), 1161–80.

Aumann, Robert J., & Drèze, Jacques H. 2008. Rational Expectations in Games. *The American Economic Review*, 98(1), 72–86.

Bacharach, Michael. 1987. A Theory of Rational Decision in Games. *Erkenntnis*, 27(1), 17–55.

Binmore, Ken. 1987. Modeling Rational Players: Part I. *Economics and Philosophy*, 3(2), 179–214.

Binmore, Ken, & Samuelson, Larry. 2001. Coordinated Action in the Electronic Mail Game. *Games and Economic Behavior*, 35(1), 6–30.

Brandenburger, Adam. 2010. Origins of Epistemic Game Theory. Hendricks, Vincent F., & Roy, Olivier (eds), *Epistemic Logic: Five Questions*. (Pages 59–69). Automatic Press.

———. 2014. *The Language of Game Theory: Putting Epistemics into the Mathematics of Games*. Vol. 5. World Scientific.

Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Chant, Sara Rachel, & Ernst, Zachary. 2008. Epistemic Conditions for Collective Action. *Mind*, 117(467), 549–73.

Chaudhuri, Ananish, Schotter, Andrew, & Sopher, Barry. 2009. Talking Ourselves to Efficiency: Coordination in Inter-Generational Minimum Effort Games with Private, Almost Common and Common Knowledge of Advice. *The Economic Journal*, 119(534), 91–122.

Chellas, Brian F. 1980. *Modal logic: An Introduction*. Vol. 316. Cambridge University Press.
Chwe, Michael Suk-Young. 2001. *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press.
Clark, Herbert H., & Marshall, Catherine R. 1981. Definite Reference and Mutual Knowledge. Joshi, A. K., Webber, B., & Sag, I. (eds), *Elements of Discourse Understanding* (Pages 10–63). Cambridge University Press.
Cohen, Danny, & Yemini, Yechiam. 1979. Protocols for Dating Coordination. *Proceedings of the Fourth Berkeley Conference on Distributed Data Management and Computer Networks* (Pages 179–188).
Costa-Gomes, Miguel A., Crawford, Vincent P., & Iriberri, Nagore. 2013. Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications. *Journal of Economic Literature*, 51(1), 5–62.
Fagin, Ronald, Halpern, Joseph Y., Moses, Yoram, & Vardi, Moshe Y. 1995. *Reasoning about Knowledge*. MIT Press.
———. 1999. Common Knowledge Revisited. *Annals of Pure and Applied Logic*, 96(1–3), 89–105.
Gray, Jim. 1978. Notes on Data Base Operating Systems. *Operating Systems, An Advanced Course*. (Pages 393–481). Springer-Verlag.
Greco, Daniel. 2014a. Could KK be OK? *Journal of Philosophy*, 111(4), 169–97.
———. 2014b. Iteration and Fragmentation. *Philosophy and Phenomenological Research*, 88(1), 656–73.
———. 2015. Iteration Principles in Epistemology I: Arguments For. *Philosophy Compass*, 10(11), 754–64.
Halpern, Joseph Y. 1986. Reasoning about Knowledge: An Overview. *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge* (Pages 1–17). Morgan Kaufmann Publishers Inc.
Halpern, Joseph Y. & Moses, Yoram. 1990. Knowledge and Common Knowledge in a Distributed Environment. *Journal of the ACM (JACM)*, 37(3), 549–587.
Harsanyi, John C. 1968. Games with Incomplete Information Played by "Bayesian" Players Part III. The Basic Probability Distribution of the Game. *Management Science*, 14(7), 486–502.
———. 1982a. Comment—Subjective Probability and the Theory of Games: Comments on Kadane and Larkey's Paper. *Management Science*, 28(2), 120–4.
———. 1982b. Rejoinder to Professors Kadane and Larkey. *Management Science*, 28(2), 124–5.
Hawthorne, John, & Stanley, Jason. 2008. Knowledge and Action. *The Journal of Philosophy*, 105(10), 571–90.
Heal, Jane. 1978. Common knowledge. *The Philosophical Quarterly*, 28(111), 116–131.
Heifetz, Aviad. 2004. Review of Michael Chwe: Rational Ritual: Culture, Coordination and Common Knowledge. *The Economic Journal*, February, F146–7.
Heinemann, Frank, Nagel, Rosemarie, & Ockenfels, Peter. 2004. The Theory of Global Games on Test: Experimental Analysis of Coordination Games with Public and Private Information. *Econometrica*, 72(5), 1583–99.
Hintikka, Jaako. 1962. *Knowledge and Belief*. Cornell University Press.
Kadane, Joseph B., & Larkey, Patrick D. 1982a. Reply to Professor Harsanyi. *Management Science*, 28(2), 124.
———. 1982b. Subjective Probability and the Theory of Games. *Management Science*, 28(2), 113–20.
Kets, Willemien. 2014. *Finite Depth of Reasoning and Equilibrium Play in Games with Incomplete Information*. Submitted MS.
Kneeland, Terri. 2015. Identifying Higher-Order Rationality. *Econometrica*, 83(5), 2065–79.
———. 2016. Coordination under Limited Depth of Reasoning. *Games and Economic Behavior*, 96, 49–64.
Lederman, Harvey. Forthcoming. Uncommon Knowledge. *Mind*.
Lismont, Luc, & Mongin, Philippe. 2003. Strong Completeness Theorems for Weak Logics of Common Belief. *Journal of Philosophical Logic*, 32(2), 115–37.
Luce, Robert Duncan, & Raiffa, Howard. 1957. *Games and Decisions: Introduction and Critical Survey*. John Wiley and Sons.
Morris, Stephen. 2002. Coordination, Communication, and Common Knowledge: A Retrospective on the Electronic-mail Game. *Oxford Review of Economic Policy*, 18(4), 433–45.

Moses, Yoram O. 1986. *Knowledge in a Distributed Environment*. Ph.D. thesis, Stanford University.
Nagel, Rosemarie. 1995. Unraveling in Guessing Games: An Experimental Study. *American Economic Review*, 85(5), 1313–26.
Nozick, Robert. 1990. *The Normative Theory of Individual Choice*. Garland Books.
Roth, Alvin E., & Schoumaker, Francoise. 1983. Note—Subjective Probability and the Theory of Games: Some Further Comments. *Management Science*, 29(11), 1337–40.
Rubinstein, Ariel. 1989. The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge". *The American Economic Review*, 79(3), 385–91.
Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. Harvard University Press.
Stahl, Dale, & Wilson, Paul W. 1994. Experimental Evidence on Players' Models of Other Players. *Journal of Economic Behavior & Organization*, 25(3), 309–27.
———. 1995. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, 10(1), 218–54.
Stalnaker, Robert C. 1994. On The Evaluation of Solution Concepts. *Theory and Decision*, 37(1), 49–73.
Strzalecki, Tomasz. 2014. Depth of Reasoning and Higher Order Beliefs. *Journal of Economic Behavior & Organization*, 108, 108–22.
Thomas, Kyle A., DeScioli, Peter, Haque, Omar Sultan, & Pinker, Steven. 2014. The Psychology of Coordination and Common Knowledge. *Journal of Personality and Social Psychology*, 107(4), 657–76.
Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.
Yemini, Yechiam, & Cohen, D. 1979. Some Issues in Distributed Processes Communication. *Proceedings of the 1st International Conference on Distributed Computing Systems* (pages 199–203).