ZACHARY ERNST

WHAT IS COMMON KNOWLEDGE?

ABSTRACT

Common knowledge is usually defined as a state in which everyone knows that *p*, everyone knows that everyone knows that *p*, and so on, *ad infinitum*. This definition is usually attributed to David Lewis, despite the fact that his own formulation bears no resemblance to common knowledge as it is usually understood. In this paper, I argue that this concept of common knowledge requires revision. Contrary to usual practice, it turns out to be difficult to model formally because existing models fail to distinguish between full-blown common knowledge and merely finite levels of interactive knowledge. Conceptually, the concept is incompatible with Lewis's intended purpose and obscures the explanatory role played by rational choice models. I propose that the concept of common knowledge be brought better into alignment with Lewis's actual formulation. This reconceptualization of common knowledge suggests a greater focus on explanations that make recourse to the cognitive constraints of real-world agents.

1. INTRODUCTION

The concept of common knowledge has come to play a crucial role in nearly every explanation of collective action, equilibrium behavior, cooperation, and convention. This is for a good reason. To take a simple example, if you and I are to meet for lunch, it is not enough for me to know that our meeting is supposed to happen tomorrow at noon. I must also know that you know when our meeting is supposed to take place, and vice-versa. To put the point generally, our collective action will depend upon our having appropriate expectations about each other's behavior. Those expectations can only be formed if we each have knowledge of each other's knowledge. This, of course, is just to say that we have some level of so-called 'interactive knowledge'.

Episteme 8.3 (2011): 209–226
DOI: 10.3366/epi.2011.0018
© Edinburgh University Press
www.eupjournals.com/epi

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

Interactive knowledge comes in degrees. The various degrees of interactive knowledge can be summarized by the following schema:

(1) $$(\text{everyone knows that})^{n} p$$

where p is the proposition that is known, and n tells us how many times to iterate the 'everyone knows that' clause. When $n = 1$, we have only so-called 'mutual knowledge'; that is, the proposition p is known by everyone, but each of us may be unaware that anyone else knows p. If the statement is true up to some particular value m, then we say that the group has interactive knowledge of p 'to level m'. In the extreme case, the group has full-blown 'common knowledge' of p just in case the statement is true for all values of n. This formulation is typically attributed to David Lewis, from his seminal *Convention* (1969), and it accounts for a great deal of the influence that Lewis has had on the philosophical, economic, and computer science literature.

Full-blown common knowledge occupies a prominent position in almost all the literature that deploys some aspect of interactive knowledge. This is mainly for pragmatic reasons. Common knowledge is vastly simpler to characterize and work with in a formal setting, and because common knowledge logically implies every weaker form of interactive knowledge, it is an attractive idealizing assumption to say that all the agents in a model have 'common knowledge' of, for example, the structure of a game, each other's rationality, and so on.

In this paper, I shall argue that the apparent simplicity of the concept of common knowledge belies a set of deep difficulties in interpreting and deploying it. Indeed, the concept is fraught with conceptual difficulties that go beyond any of the typical problems we must be sensitive to with regard to other idealizing assumptions. I shall approach these problems from two perspectives. First, it turns out that Lewis never formulated his concept of common knowledge in any way resembling (1). And this is for good reason – that formulation is inadequate for the purpose for which Lewis defined the concept in the first place. By better appreciating the context in which the concept was originally deployed, we may better appreciate some of its limitations. The second perspective is formal. The best known and widely-deployed formal apparatus for constructing models of common knowledge is Robert Aumann's information partition model (1999a, 1999b). However, this model is unable to distinguish between full-blown common knowledge and mere finite levels of interactive knowledge. It therefore obscures important questions that ought to arise when we consider common knowledge and interactive knowledge generally. I conclude with some positive suggestions for how we ought to revise the concept of common knowledge.

2. THE ORIGINS OF COMMON KNOWLEDGE

If there are analytic truths, the most natural place to find them is in the truths of logic and mathematics; and if there are such analytic logical truths, then their

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

truth is underwritten by the meanings of logical terms. Because the meanings of terms – including the logical terms – are conventional, Quine's attack on analyticity required an argument that the meanings of logical terms are not conventional. This task was taken up in his "Truth by Convention" (1966).

The most influential argument Quine used for this purpose is a regress argument. In short, the argument states that if logical truths were true by convention, then speakers would need to know when that convention applied to any given statement. But in order to determine whether a convention applied, some amount of logic would be necessary. For example, a speaker who was cognizant of the convention regarding *modus ponens* would have know how to deploy the convention 'If a statement is of the form such-and-such, then an application of *modus ponens* is warranted.' However, it is plain that a pre-existing knowledge of *modus ponens* would be required if one were to know when to apply the *modus ponens* convention, which creates a vicious regress.

Although Quine's regress argument is widely accepted, Lewis noted that it requires a concept of 'convention' that may be inappropriate.[1] Quine thought of a convention as a sort of logical rule to be deployed by speakers who know the convention and know when it applies. In other words, in order to deploy a convention, it would be necessary for speakers to be able to formulate the conditions under which it applies. But, as Lewis notes, we are often party to conventions that we cannot articulate. For example, the meanings of words are clearly conventional, but speakers typically can 'know' how to use those terms appropriately without being able to precisely express the conditions under which that term can be used.

Although Lewis's *Convention* is best known for articulating an equilibrium concept of conventions, his overarching goal was to rehabilitate analyticity. This was to be done by giving an account of conventions that allows for meanings of words to be conventional, and for individuals to know the meanings of words because they are party to the convention, but without requiring that the individuals have explicit knowledge of the conditions under which the words are appropriate. For if they can be party to the convention without being able to precisely articulate the convention, then, *a fortiori*, they would not need to deploy logical inference to correctly use those words. If this project could be completed, then one could claim that the meanings of words – and of logical terms especially – are conventional, and that therefore the truth of some sentences is guaranteed by the meanings of those words, without being subject to Quine's regress argument.

The problem for Lewis is that some sort of knowledge is required in order for individuals to take part in a convention; however, his goal of explaining how conventions can exist without satisfying Quine's assumptions requires him to avoid imposing strong epistemic conditions on the individuals who are party to the convention. In short, the account must allow for the fact that individuals must – in some sense – 'know' how to apply the convention, but this knowledge must not

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

take the form of an inference, or else it would be vulnerable to Quine's regress argument.

In other words, in order to accomplish his goal of rehabilitating truth by convention, the conventions that underwrite logical truths must thread a difficult needle. They must be 'known' to the members of the group or else the account is too weak to underwrite truths by convention. But this form of knowledge cannot be too stringent, or else inferences will be required in order for the members of the group to acquire that knowledge.

Because this difficult middle-ground is required for Lewis's theory, it is odd that he is so often interpreted as requiring common knowledge in the form of (1). This is because common knowledge in the sense of (1) attributes extremely high levels of rationality and cognitive sophistication to the agents. To convince yourself that great cognitive sophistication is indeed implicitly required by this sense of 'common knowledge', try to keep straight in your mind the difference between fourth- and fifth-degree interactive knowledge that *p*; and while you're at it, try to come up with a situation in which that difference is salient. If we are to require full-blown common knowledge, then it is only natural to expect a situation to arise in which one-thousand-degree interactive knowledge is not sufficient, but common knowledge is. And for this to be true, the agents would have to be capable of truly remarkable mental feats. If such unrealistically high levels of rationality are required in order for individuals to achieve common knowledge, then it is difficult to see how Quine's regress argument has been avoided – after all, that argument hinges on the fact that inferences are necessary in order to know when a convention applies. If anything, formulation (1) actually *strengthens* Quine's position.

Perhaps it is for this reason that Lewis's account of convention – and of common knowledge in particular – is criticized for making too stringent demands of the agents. Although it is often useful to make the assumption that common knowledge can be attained by the agents in a particular model, it is surely too strong an assumption for an account that is supposed to apply to real-world agents whose cognitive powers fall short of those of *Homo economicus*. And to return to Lewis's motivations for his account, it is surely wrong-headed to require such cognitive sophistication if the goal is to show how conventions can become established without requiring the agents to make inferences at all.

Luckily, Lewis can plead innocent to the charge of requiring too strong a condition on the cognitive powers of the agents. For he never formulates Condition (1) anywhere in his theory. Indeed, the question of why anyone would attribute this condition to him raises difficult issues, to which we shall return later.

A reconstruction of Lewis's concept of common knowledge that is closer – but still not yet accurate – is that Condition (1) is to be weakened to:

(2) (everyone has reason to believe that)^n p

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

for all values of *n*. Clearly, Condition (2) is not vulnerable to the charge of requiring unrealistically sophisticated cognitive powers of the agents; after all, it does not require that the agents actually 'know' (in the usual sense of the word) anything at all about *p*. Much less does it require that the agents know anything about the others' knowledge. Rather, it only requires that the agents have some reason or other to believe that *p*, to believe that others have reason to believe that *p*, and so on. Although it is far from clear under what conditions the agents can have this infinite hierarchy of reasons (presumably, this would depend upon how we construe the term 'reason'), this reconstructs Lewis's requirement in such a way that ordinary agents can satisfy it.

However, Condition (2) is far too weak; because it does not require that the agents actually know *p*, it is fully compatible with the agents believing that *p* is false. Implausibly, it would turn out that Lewis's 'common knowledge' condition is satisfied by agents who are quite certain that *p* is false, and who believe that others believe that *p* is false, and so on, *ad infinitum*. To put the point rhetorically, what use is an epistemic condition that only requires *reasons* to believe *p*, if the agents never deploy those reasons to form the belief that *p*? Thus, we have a dilemma: either the reasons for believing *p* are responsible for forming the belief that *p*, or they are not. If they are, then we run the risk of imposing unrealistically strong epistemic conditions on the agents. If they are not, then the reasons are irrelevant, and so is the epistemic condition.

3. IT'S A SMALL WORLD

Both of the above readings of Lewis are well-represented in the literature (e.g. Burge 1975, Gilbert 1983, Vanderschraaf 1998, Fagin et al. 1999, and Binmore and Samuelson 2001). However, the fact that they do not stand up to any reasonable scrutiny is at least a reason to suspect that they are not intended interpretations of the theory. What the intended interpretation actually is, however, is far from clear.

We can now benefit from discussing Lewis's actual definition of 'common knowledge'. In order to distinguish Lewis's concept from definition (1), I shall refer to Lewis's concept as 'COMMON KNOWLEDGE'.

Definition: A proposition *p* is COMMON KNOWLEDGE in the group *G* if a state of affairs *A* obtains such that:

(1) Everyone in *G* has reason to believe that *A* holds;

  (a) *A* indicates to everyone in *G* that everyone in *G* has reason to believe that *A* holds;
  (b) *A* indicates to everyone in *G* that *p*.

Lewis's concept of COMMON KNOWLEDGE proceeds by way of two other concepts, which Lewis leaves somewhat vague – these are a 'reason to believe'

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

relation and an 'indication' relation. Understanding those two relations will be important for understanding Lewis, diagnosing the difficulties that are raised by other interpretations, and pointing the way to further work that needs to be done in order to derive a satisfactory account of convention.

To borrow a term from the economics literature (Binmore 1985), let us begin by considering a so-called 'small-world' model of knowledge. In a small-world model, there is no room for distinguishing between an agent's believing *p*, knowing *p*, and having justification for *p*. Agents in such models know *p* straightaway if they are in a position to have justification for it; conversely, if they are not in a position to know *p*, then they do not believe it.

Agents in such models are 'logically omniscient', that is, they know all the logical truths and valid rules of inference, and they have the cognitive powers to close the set of their beliefs under implication. This entails that they have perfect powers of introspection with regard to their own beliefs. Among other features of such agents, their beliefs (and hence their knowledge) obeys the following axioms:

$$
\begin{gathered}
K(p \supset q) \supset (Kp \supset Kq) \\
Kp \supset KKp \\
\sim Kp \supset K \sim Kp \\
Kp \supset p
\end{gathered}
\tag{3}
$$

where 'Kp' means 'the agent knows that *p*'.

These (what I shall call) 'small world assumptions' lead naturally from COMMON KNOWLEDGE to Condition (1), so long as we also take a liberal attitude toward the indication relation. If, in keeping with the spirit of the small world assumptions, we say that an agent for whom it is indicated that *p* will thereby be justified in believing that *p*, and therefore know that *p*, then we quickly derive the infinite hierarchy of propositions given by Condition (1). Thus, the conjunction of COMMON KNOWLEDGE and the small world assumptions entails Condition (1). And clearly, because the small world assumptions erase the distinction between 'having a reason to believe that *p*' and 'knowing that *p*', it is equally clear that COMMON KNOWLEDGE plus the small world assumptions entail Condition (2) *a fortiori*.

Because virtually nothing short of the small world assumptions will underwrite a move from COMMON KNOWLEDGE to Condition (1), the question of whether the strong reading of Lewis is justified comes down to determining whether the small world assumptions are anything that Lewis would support, and more interestingly, whether they are true (regardless of what Lewis might have supported). The answer to the first question, I think, is unquestionably 'no', as there is no textual evidence whatsoever for attributing these assumptions to Lewis, and because such assumptions are plainly incompatible with his overarching project. Similarly, the answer to the second question is certainly negative as well. Of course there is a

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

difference between justification, belief, knowledge, and 'indication' (whatever the 'indication' relation might turn out to be).

So we are left with two questions:

(1) What are the effects of the small world assumptions?
(2) If the small world assumptions are undesirable – perhaps because they cannot cohere with the motivations for a program such as Lewis's – how can we best replace or eliminate those assumptions?

I shall address each of these questions in turn.

3.1. Effects of the Small World Assumptions

Like any set of simplifying assumptions, the small world assumptions are useful because they allow us to create simple, mathematically tractable models of complex phenomena. By collapsing justification, belief, and knowledge into a single concept, we eliminate the need to describe the relationships between those concepts; and if the goal of a model of interactive knowledge is to understand how one person's knowledge about another can be derived from other items of knowledge, then the simplification may be harmless.

However, there is a price to be paid for taking these assumptions on board. If we consider the knowledge axioms (3), we find that they are exactly the axioms of the modal logic S5. Thus, an epistemic logic with those axioms inherits the metalinguistic properties of S5. For our purposes here, we should note that one of these properties is the so-called 'finite model property' (Harrop 1958, 1973), which is defined as follows:

Definition: A logic $\mathcal{L}$ has the finite model property just in case for any formula α, if $\mathcal{L} \nvdash \alpha$, then there is a model $\mathcal{M}$ of $\mathcal{L}$ such that $\mathcal{M}$ has only a finite number of elements, and $\mathcal{M} \not\models \alpha$.

Informally, the definition states that a logic has the finite model property if every non-theorem can be shown to be a non-theorem by a model that has only a finite domain. For example, if we consider the usual Kripke semantics for S5, and some formula α that is not a theorem of S5, then we are guaranteed that there exists a model with only a finite number of worlds in which α is false at some world.[2]

This fact has unfortunate repercussions for an epistemic logic with the small world assumptions. To see this, we must consider what this epistemic logic would look like in a little more detail. Suppose we were to model the interactive knowledge of two agents, which we shall call Alice and Bob. Each will obviously know different propositions, so we introduce a knowledge operator for each of them, K_A and K_B. With the small world assumptions, each of the knowledge operators will obey (3).[3]

So let us suppose we are trying to model Alice and Bob's knowledge of a particular proposition p. If they have nth degree interactive knowledge of p, then

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

the following formulas will be true in the model:

$$\overbrace{K_A K_B K_A \cdots K_B}^{n \text{ times}} p$$
$$\underbrace{K_B K_A K_B \cdots K_A}_{n \text{ times}} p$$

Let us call these two formulas Γ. Now suppose there is some formula α that we need to show either is or is not a consequence of Alice and Bob's nth degree interactive knowledge. If there is a model that distinguishes between Γ and α, then the finite model property guarantees that there is a finite model that does so. Let us call that model $\mathcal{M}$, and let us say that $\mathcal{M}$ has exactly m possible worlds.

Unfortunately, because $\mathcal{M}$ has only m possible worlds, there is guaranteed to be some critical value c such that if Alice and Bob have interactive knowledge to degree c, then they also have interactive knowledge to any degree $d \geq c$. To see this, consider what it means in a possible worlds model for Alice and Bob to have interactive knowledge of p to some degree n. In the model, they will have such knowledge if and only if p is true at a certain subset of worlds. If there are only m worlds in the model – where m is finite – then there are only 2^m subsets of worlds. Therefore, there will be only a finite number of ways in which p can take truth values at the worlds in the model. By the pigeonhole principle, the model can distinguish (at best) only among 2^m different degrees of interactive knowledge. Thus, if Alice and Bob have interactive knowledge to degree 2^m, then they will automatically have interactive knowledge to every higher degree as well. And this implies, among other things, that the model will treat interactive knowledge to degree 2^m exactly as it treats full-blown common knowledge (in the sense of Condition (1)).

Therefore, the small world assumptions turn out to make the concept of common knowledge redundant in the following sense. If we wanted to show that common knowledge is special in that it implies some fact not implied by any weaker degree of interactive knowledge, no model with the small world assumptions could show this. For the finite model property would guarantee that common knowledge is equivalent, for any given model, with interactive knowledge to some finite degree. So if common knowledge were to imply some fact in the model, so would interactive knowledge that falls short of common knowledge. The conclusion one may draw from this fact is that the reading of Lewis that underwrites the move to Condition (1) requires assumptions that actually defeat the purpose of studying common knowledge in the first place. Recalling that multimodal S5 provides the same models as Aumann's information partition model, we see that the problem generalizes to Aumann's framework as well.

4. POSSIBLE RESPONSES

There are at least three possible responses one could make to the observation that these models do not differentiate between common knowledge and finite levels

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

of interactive knowledge. I shall call these the optimistic, pessimistic, and puzzled responses.

The optimistic response is to make a virtue of necessity, so to speak. Confronted with the observation that cognitively limited agents will automatically acquire a form of full-blown common knowledge in these models, one might note that real-world agents do, in fact, act as if they had common knowledge even when it is unrealistic to attribute such cognitive feats to them. The results of the previous section may do something to harmonize these two observations. Specifically, we might interpret the formal results from the previous section as explaining why real-world human beings do act as if they had common knowledge. That is, we might say that human beings do often reach the critical threshold above which the model generates common knowledge; thus, human beings act as if they had common knowledge.

There are several problems with such a response, however. First, the account it would propose is clearly at odds with what actually occurs when individuals coordinate their behavior. According to the optimistic response, individuals who are in a situation in which they coordinate would do so by achieving a certain level of interactive knowledge. After that point, their cognitive capacities would be reached, and they would default to a behavior that would have been predicted had they achieved full-blown common knowledge. But plainly, this is not the case. It is very rare for individuals to achieve interactive knowledge even to degree two or three, except perhaps in unusually demanding or competitive circumstances (such as when playing poker, or trading in a complex financial market). Typically, individuals coordinate by recognizing that they are in a particular environment, perhaps after assuring themselves that the others also recognize that they are in such an environment. Coordination will usually occur 'straightaway', without further deliberation.

A pessimistic response is simply to dismiss the model. The arguments from the previous section would be interpreted to show that what we had thought was a harmless idealizing assumption is anything but harmless, and in fact represents a fatal flaw in the model.

The pessimistic response is too pessimistic, in my opinion. Even if the model is shown to be hopelessly unrealistic, we still have to recognize that the model gives valuable predictions across a wide range of cases. Thus, we should interpret the model as instrumentally valuable, even if it fails to give us explanatory information (at least in a sufficiently stringent sense of 'explanatory'). On this view, the model is recognized as often yielding the right results, despite the fact that it does not reveal the real mechanisms underlying the behavior.

Instead, I would endorse the puzzled response. Starting from the assumption that the models are valuable but not realistic, we are led to wonder why the models ever work at all. Either it is an inexplicable coincidence that the models generate the right predictions, or else the mechanisms that are posited in the model bear some substantial relationship to the actual mechanisms that underly

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

the behavior of real-world human beings. I will conclude by considering this possibility.

5. LEWIS AGAIN

Conventions, on a Lewis-style account, are behavioral regularities that stem from a stable (equilibrium) pattern of expectations. They are not necessarily behaviors that arise as a result of a long and complex chain of reasoning. Thus, when two drivers each stop at their respective stop signs, they do so straightaway upon seeing the signs, and not after first inferring various facts about the other's beliefs. Of course, if we were to ask each of them for a reason why they stopped, they would presumably be capable of doing so in terms of the expected behavior of the other person. But this *post hoc* justification of their behavior does not explain why they stopped; rather, it *justifies* their stopping. We therefore need to distinguish between the actual events leading up to their conventional behavior on the one hand, and the rational reconstruction of their behavior on the other. The former constitutes an *explanation*; the latter constitutes a *justification*.

In drawing this distinction, one runs the risk of trivializing the project of explaining conventions. For if we agree that the rational reconstruction of the behavior is not explanatory – at least not in the sense of describing the actual reasons why the agents behaved in this way – then we might be led to the conclusion that there is little of philosophical interest in this subject, especially if we are interested, as Lewis was, in the descriptive project of explaining why real individuals in the real world adopt conventions.

The danger of trivializing the subject comes to the foreground again in *Convention*. For as Lewis's discussion becomes more sophisticated, and simplifying assumptions are left behind, he comes to rely more heavily on the concept of 'salience', which remains an unanalyzed concept, explained only through examples. For instance, consider two friends who are visiting an unfamiliar city together, and who become separated. In this situation, they face a coordination problem – each does better if they choose the same location to meet, but no location is intrinsically better than any other for this purpose. Lewis alleges – I think correctly – that a typical pair of individuals will have no difficulty meeting; this is because they will choose a salient option, such as their hotel, or a local landmark they have recently visited.

The fact that salience would play an important role in explaining conventional behavior poses a deep difficulty. For if the rational reconstruction has such severe explanatory limits, then we might be left in the unfortunate position of having to rely quite heavily on salience to explain coordination. And of course, if the explanatory burden is placed so heavily upon an unanalyzed concept such as 'salience', this weighs against the value of the explanation.

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

6. RECOVERING THE PROJECT

So far, we have observed several features of COMMON KNOWLEDGE, convention, and salience:

(1) Common knowledge, in the sense of Condition (1), trivializes the concept by erasing the distinction between full-blown common knowledge and merely high levels of interactive knowledge.
(2) The requirement of COMMON KNOWLEDGE must be interpreted as placing a condition on the agents' environment, and only very minimal (but not negligible) conditions on their actual epistemic states.
(3) As the rational choice account is recognized as less explanatory in the sense of failing to describe the sequence of events leading to an individual's behavior, a greater explanatory burden falls upon unanalyzed concepts such as 'salience'.

These remarks have been largely negative – I have argued for what we must *not* do if we are to understand convention and COMMON KNOWLEDGE. In this section, I will begin to outline what a positive project ought to be. I will argue for a picture of convention, COMMON KNOWLEDGE, and salience that avoids trivializing important problems, and that defines a research project that has been neglected.

6.1. Deductive, Eductive, and Constraint

The project I will describe is an alternative to the two projects that Ken Binmore has termed 'deductive' and 'eductive' (1985). A deductive project is to model the behavior of individuals or groups according to a rational choice paradigm, in which the agents in the model are assumed to have high levels of rationality, calculating ability, background knowledge, and so on. The model aims at showing what a fully rational agent would do when faced with a particular problem or strategic situation. The consequences of the model are derived from certain assumptions and axioms that describe the reasoning process of rational individuals. On the other hand, an 'eductive' model is largely free of rational choice assumptions; it models the behavior of individuals as the outcome of an evolutionary process, which may be understood as either biological or social. The evolutionary process will typically pick out stable or highly fit behaviors as the ones that a population of non-rational agents will eventually settle upon. Clearly, deductive and eductive models can make similar or dramatically different predications; and they also are capable of showing how far the rational choice assumptions of a deductive model can be relaxed, while yielding the same results.

The fact that the behavior of individuals can be modeled within a rational choice framework gives us an important clue as to how conventional behaviors can be

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

explained. Across a wide range of environments, our behavior is guided to an important degree by a set of heuristics. These heuristics are aimed at yielding the optimal payoffs for various common situations. So we might think of an optimal agent as one whose behavior is always perfect for every situation she encounters; we might imagine that she chooses a particular behavior by calculating perfectly the expected value of every possible behavior available to her. Thus, if there were even the smallest advantage for choosing one behavior over another, she would choose the most advantageous. There would remain, of course, some set of environments in which it is beneficial to coordinate her behavior with other (perfectly rational) agents, but in which no outcome is even the slightest bit to be preferred over any other, so long as coordination is achieved. In this type of case, a perfectly rational agent will benefit from pre-established conventions.

This range of situations in which an optimal agent would benefit from conventions would probably turn out to be quite limited. For example, it could turn out that there is some tiny advantage to driving on the right side of the road (perhaps the advantage arises from asymmetries in the brain, or some other contingent fact). A population of optimal agents would know this, and would therefore choose to drive on the right side of the road, knowing that all the others in the population would be aware of this tiny benefit. If, however, there were no advantage whatsoever to having one type of coordination over another when individuals approach a four-way stop, then these rational agents would benefit from a coordinating convention.

Thus, the frequency and range of the conventions that operate within a population is inversely proportional to the cognitive powers of the agents in the population; agents who are capable of very little reasoning benefit more from conventions than agents who have greater cognitive capacities. This observation suggests that settling on a set of conventions is the result of a constrained optimization problem. That is, there is a (possibly not unique) optimal set of behaviors for a population of agents; an optimal set of conventions would endorse the optimal behavior in every case. But of course, we are limited in our capacities to keep track of and deploy conventions. So the solution to the optimization problem is constrained by our available cognitive resources.

In this way, we explain conventional behavior by reconceptualizing the problem as an optimality problem under constraints. If this picture is correct, then an explanation of our conventional behavior is to be explained by a two-step process that is neither deductive nor eductive. First, we must identify an optimal set of conventions. Second, we must understand the constraints on the available sets of heuristics available for forming and deploying those conventions.

6.1.1. Optimality

The problem of determining which set of conventions is optimal may appear to have a relatively straightforward answer; specifically, we might take the optimal

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

set of conventions to coincide with the behavior of a perfectly rational agent. After all, we commonly understand the term 'rational agent' as just *meaning* an agent whose behaviors have the greatest expected payoff. We might then identify a set of reasoning patterns – such as dominance arguments, expected utility calculations, and so on – that a rational agent would employ. Then, the optimal set of conventions would simply be the conventions that could be endorsed by those reasoning patterns.

However, this concept of rationality is of limited use here. Putting aside cases in which it is unclear what a 'rational' agent would do (e.g. Newcomb's Problem), there are deeper problems with the concept of rationality as the ability to calculate optimal behaviors. Some of these issues are naturally illustrated by problems concerning interactive knowledge, and common knowledge in particular. In short, the problem is that for any situation in which an ideally rational agent would require interactive knowledge before acting, there are similar situations in which higher levels of interactive knowledge would be required, but in which real-world agents benefit by acting without such knowledge.

For example, suppose that Mark and Bob want to arrange a clandestine meeting, and that they have arranged a system in which Mark will put a flag outside his window if he wants to meet. When Bob sees the flag, they are to meet in a particular parking garage at a prearranged time. If Mark and Bob both know that the flag has been placed outside Mark's window, then they have 'mutual knowledge' of that fact. However, mutual knowledge is clearly not sufficient for agents to coordinate if each incurs a great risk in doing 'his part' of the coordinated action. For example, even if Mark and Bob both know that the flag has been placed outside the window, this does not entail that Mark knows that Bob saw it. Now suppose that there is a significant downside risk for Mark if he shows up at the garage alone – perhaps he will draw unwanted attention to himself if he does. Under such conditions, it may be fully rational for Mark to stay home. This is because as far as he knows, Bob did not see the flag and so will not know to go to the garage, which leaves Mark in a bad situation.

So it is quite natural to learn from such considerations that the agents will need more than mutual knowledge. Perhaps, one might suppose, Bob needs to be able to signal to Mark that he saw the flag. So let us suppose that they have arranged that when Bob sees the flag, he is to move a particular flower pot that is visible from Mark's window.

Of course, a vicious regress quickly ensues if we get in the habit of worrying about the agents reasoning in such a way. For if Bob does move the flower pot, and even if Mark sees that the flower pot has been moved, Bob does not know that Mark has seen it. Thus, Bob may reason that if Mark has not seen the flower pot, then the situation is no better than if the flower pot signal had never been arranged. Thus, it will be rational for Bob to stay home. The regress can be iterated as many times as we like. No matter how many flower pots and flags they have arranged as signals, so long as the downside risk of showing up alone at the garage is high

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

enough, they will fail to meet. And this is in spite of the fact that they both want to meet, and they both know this about each other, and they both know that any finite set of signals has been successfully sent and received.

So what has happened to Mark and Bob that has prevented their coordination? Being rational, and possessing an appropriate level of risk aversion, Mark and Bob have correctly reasoned – in accordance with any typical standard of rationality – that they require some level of interactive knowledge before taking the risk of acting. However, the very same reasoning that they have used in order to demand one level of interactive knowledge generalizes to their detriment. That is, as soon as they have seen that it is rational to demand one level of reassurance, they must also recognize that it is equally rational to demand another level of reassurance, and so on, *ad infinitum*.

The observation that Mark and Bob cannot coordinate, given the assumption of perfect rationality, strikes us as extremely counterintuitive, to say the least. Surely, one would naturally protest, any rational individual will coordinate in such a situation while having only one or at most two levels of interactive knowledge. But the reason why this result is so counterintuitive is that we are often faced with similar situations, and we successfully act without making such strong demands for reassurances. However, the reason why we do not make 'unreasonable' demands for an increasing number of reassurances is because we are not typically engaged in a process of reasoning in accordance with these 'rational' requirements at all. Rather, we approach such situations in the real world through a set of simple heuristics that we take as sufficient for acting, even if those heuristics would ultimately lead to different behavior than a rational choice model would endorse. And of course, as real world agents, we benefit from using those heuristics; we would be worse off if we consistently deployed the rational principles that Mark and Bob are forced to use.

And again, it seems that there has been some trick here. For we began by assuming that 'rational' agents would simply be the ones whose reasoning led to the highest possible payoffs. But we have concluded that there are situations in which a 'rational' agent will fail to achieve the payoffs that a more realistic agent would receive. The problem is not with the example. Rather, the problem is with the concept of 'rationality' as payoff maximization. In fact, the concept of 'rationality' should be thought of as having deep similarities to the concept of 'convention'. Specifically, conventions are the result of our being faced with an optimization problem under significant constraints, where those constraints are due to our cognitive limitations. So, too, 'rationality' is best thought of as a sub-optimal solution to an optimization problem under the same set of constraints.

When we are concerned with understanding the nature of rationality, we are typically concerned with extracting a set of rational principles that an agent ought to use in order to solve decision problems. Those rational principles – such as dominance, expected payoff calculations, and so on – are useful heuristics that generally lead to optimal behavior across a wide range of realistic circumstances.

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

But like any set of heuristics, there are circumstances, some of which may strike us as highly artificial, in which they fail. One such case is Newcomb's Problem, in which dominance arguments fail to yield the highest available payoff. But the right response to Newcomb's Problem is not that there is anything wrong with dominance arguments, nor is it that rational individuals should not use considerations of dominance. Rather, the appropriate response is that Newcomb's Problem and cases like it show that there are limits to any reasonable set of heuristics. Because rationality consists in using the right heuristics for a certain range of problems under constraints, it is entirely possible – and probably inevitable – that there will be cases outside of that range in which those heuristics fail.

The conclusion that one may draw from these considerations is that the optimality of a set of conventions should not be defined in terms of any absolute claim that they will yield the best outcomes. Rather, optimality of conventions – and indeed, of rationality itself – is relative to a certain range of 'normal' cases, and also to the realistic cognitive limitations of the agent. A rational choice model, in the economist's sense, will approximate the optimal behavior of rational agents only to the extent that the principles used by an ideally rational agent coincide with the optimal principles for reasoning about cases that fall within a certain specifiable range.

6.1.2. Constraints

An ideal agent who did not have any practical limits to her cognitive capacities would be able to take every environment 'on its own terms', and calculate exactly the correct response on a case-by-case basis. This agent would have no reason to employ any of the general principles and procedures that we would find in a typical decision theory textbook. Thus, we might not even recognize this ideal agent as 'rational' in the usual sense, because that agent would have no use for the usual list of 'rational principles'.

A realistic agent, one with cognitive limits, would not be able to approach each choice situation individually. Rather, she would have to partition the set of possible situations into a manageable number, and then treat the members of each partition according to the same decision procedures. So for instance, a realistic agent might have a partition of cases that are structurally similar to the Prisoner's Dilemma. For such an agent, it would be rational (assuming a typical range of likely circumstances) to play the dominant strategy every time she encounters a problem from this partition. Such an agent might receive a lower payoff than an 'irrational' agent would if she were to confront Newcomb's Problem. Conversely, this agent might also receive a lower payoff than would an agent who was able to partition the possible cases more finely, so as to be able to recognize at least some circumstances in which dominance considerations are inappropriate.

Thus, the task of understanding the constraints under which an agent forms a set of heuristics comes down, at least to the first approximation, to estimating

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

how finely the agent can partition the range of available circumstances. If it were necessary to have only a very small number of partitions – perhaps because the agent's cognitive capacities were severely limited, or because she lacked the experience necessary to conceptualize a broader range of cases – then there would be greater opportunity for her rational principles to yield suboptimal payoffs. This, however, does not impugn her rationality; it only implies that her rationality is a response to relatively severe constraints.

7. LESSONS FOR INTERACTIVE EPISTEMOLOGY AND CONVENTION

We cannot achieve a full understanding of convention, nor of COMMON KNOWLEDGE, by taking either the deductive or eductive approach to the study of rationality. For the deductive approach necessarily employs a set of rational choice assumptions that are supposed to characterize an 'ideally rational' agent, who lacks any cognitive limitations whatsoever. But conventions – and indeed, the concept of rationality itself – are intimately tied up with the cognitive limitations of the agent; this is especially clear when we observe that a so-called 'ideally rational' agent would have little need of conventions or of rational principles in the first place. Thus, it is especially confused to try to explain our own deployment of conventions and rational principles by making recourse to a model in which the *explanans* obviates the need for the *explanandum*.

Rather, the approach I am advocating here is to understand these concepts of convention and rationality as necessarily requiring us to take into account the cognitive limitations of the agents. In this way, we recognize that conventions and rationality are solutions to optimization problems under constraints. This is not an eductive procedure either, if we take eductive explanations as discarding the need for modeling an agent's rationality. For the correct explanatory strategy will model an agent's rationality and seek to derive rational principles for the agent's behavior, but will re-conceptualize rationality outside of the rational choice framework.

The project I am advocating makes it possible to explain, finally, what COMMON KNOWLEDGE really amounts to. Agents face a set of conditions under which they must decide whether it is rational for them to coordinate with others. An optimal procedure for solving this problem may require the agent to have some reassurance – in the form of interactive knowledge – that the other agents are aware of the relevant facts of the situation. It is obviously impossible for a realistic agent to treat every case individually, so she will instead settle on some set of heuristics for determining whether it is likely that the epistemic conditions have been met. Whether Lewis thought of his definition in this way or not, we may think of these heuristics as being triggered by the 'situation *A*' and the 'indication' relation. The conditions for COMMON KNOWLEDGE require that the agent is, indeed, aware that the situation is appropriate, but this does not entail that an infinite hierarchy of facts has been derived, or even that it could be derived. The problem is exactly

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

like that of determining what the weather is likely to be tomorrow on the basis of a few indications such as the direction of the wind and the color of the clouds. An agent is fully aware of these indications and comes to a decision based upon them, but does not – and perhaps cannot – make a detailed chain of inferences about the weather based upon those simple observations. So, too, an agent has COMMON KNOWLEDGE by becoming aware of certain facts about the situation, but does not make any complex inferences that use those facts as premises.[4]

REFERENCES

Aumann, Robert J. 1999a. "Interactive Epistemology I: Knowledge." *International Journal of Game Theory* 28: 263–300.

Aumann, Robert J. 1999b. "Interactive Epistemology II: Probability." *International Journal of Game Theory* 28: 301–314.

Binmore, Ken. 1985. "Equilibria in Extensive Games." *The Economic Journal* 95: 51–9.

Binmore, Ken and Larry Samuelson. 2001. "Coordinated Action in the Electronic Mail Game." *Games and Economic Behavior* 35: 6–30.

Boghossian, Paul. 1996. "Analyticity Reconsidered." *Noûs* 30(3): 360–91.

Burge, Tyler. 1975. "On Knowledge and Convention." *The Philosophical Review* 84(2): 249–55.

Fagin, Ronald, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. 1999. "Common Knowledge Revisited." *Annals of Pure and Applied Logic* 96(1–3): 89–105.

Gilbert, Margaret. 1983. "Notes on the Concept of a Social Convention." *New Literary History* 14(2): 225–51.

Harrop, Ronald. 1958. "On the Existence of Finite Models and Decision Procedures for Propositional Calculi." *Proceedings of the Cambridge Philosophical Society* 54: 1–13.

Harrop, Ronald. 1973. "On Simple, Weak and Strong Models of Propositional Calculi." *Proceedings of the Cambridge Philosophical Society* 74: 1–9.

Kripke, Saul A. 1959. "A Completeness Theorem in Modal Logic." *The Journal of Symbolic Logic* 24(1): 1–14.

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

Quine, Willard Van Orman. 1966. "Truth by Convention." In *Ways of Paradox*. New York: Random House.

Vanderschraaf, Peter. 1998. "Knowledge, Equilibrium and Convention." *Erkenntnis* 49(3): 337–69.

NOTES

1 In this respect, Lewis's position agrees with that of Paul Boghossian, who writes:

> [The regress argument] of Quine's has been very influential; and I think that there is no doubt that it works against its target as specified. However, it is arguable that its target as specified isn't the view that needs defeating. (1996, 381)

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press

2 This is a result of Kripke's original completeness proof for S5. In the course of the proof, Kripke shows how to construct a countermodel for any non-theorem; this countermodel is guaranteed to have a number of worlds that is directly proportional to the number of subterms in the non-theorem, which is always finite (Kripke 1959).
3 This feature makes the epistemic logic a so-called 'multimodal' system because it contains more than one modal operator, unlike S5. However, this system will still have the finite model property.
4 An earlier version of this paper was presented in 2009 at the University of California-Irvine workshop on game theory and the social contract. I am grateful to the participants of that conference, and especially to Brian Skyrms and Ken Binmore.

Zachary Ernst is an Associate Professor of Philosophy at the University of Missouri-Columbia. His primary interests are logic, game theory, and the philosophy of science.

https://doi.org/10.3366/epi.2011.0018 Published online by Cambridge University Press