Common reasoning in games:

a Lewisian analysis of common knowledge of rationality*

Robin P. Cubitt[+] and Robert Sugden[++]

7 January 2014

[+]School of Economics, University of Nottingham, Nottingham NG7 2RD, United Kingdom
[++]School of Economics, University of East Anglia, Norwich NR4 7TJ, United Kingdom

Email:

Robin.Cubitt@nottingham.ac.uk
r.sugden@uea.ac.uk

* We are grateful for comments on earlier versions of the paper to Giacomo Bonnano, Adam Brandenburger, John Collins, Franz Dietrich, an editor and two anonymous referees, and to participants in various seminars, conferences and workshops at which we have presented the paper. The completion of the paper was supported by the Economic and Social Research Council (award ES/K002201/1).

Abstract

We present a new class of models of players' reasoning in non-cooperative games, inspired by David Lewis's account of common knowledge. We argue that the models in this class formalise common knowledge of rationality in a way that is distinctive, in virtue of modelling steps of reasoning; and attractive, in virtue of being able to represent coherently common knowledge of any consistent standard of individual decision-theoretic rationality. We contrast our approach with that of Robert Aumann (1987), arguing that the former avoids and diagnoses certain paradoxes to which the latter may give rise when extended in particular ways.

Short title

Common reasoning in games

Keywords

Common reasoning; common knowledge; common knowledge of rationality; David Lewis; Robert Aumann.

1. Introduction

It is a fundamental assumption of standard game theory that each player of a game acts rationally and that this is common knowledge amongst them – in short, that there is *common knowledge of rationality* (CKR). In most day-to-day applications, this assumption is not explicit; analysis is conducted using recognised 'solution concepts', such as Nash equilibrium or iterated deletion of dominated strategies. But one of the core foundational enterprises of standard game theory has been to investigate the implications of CKR for players' strategy choices and beliefs, and there has been a long-standing presumption that acceptable solution concepts ought at least to be consistent with CKR.

Intuitively, CKR seems a meaningful idealisation, in the same sense that perfect competition is a meaningful idealisation in economics or frictionless surfaces are in theoretical mechanics. However, attempts to formalise the assumption are notoriously liable to generate paradoxical implications, for example when the concept of rationality that is assumed to be common knowledge includes some principle of 'admissibility' or 'caution'. The results we have in mind are, in our view, not just surprising or technically challenging, nor best seen as merely raising doubts about particular conceptions of rationality. Instead, we think they pose fundamental questions about how CKR should be modelled.

In this paper, we propose a new and distinctive approach to modelling CKR. Our modelling strategy is inspired by Lewis (1969). Although Lewis is widely credited with the first precise definition of common knowledge, it is less well known among game theorists that this definition is only one component of a detailed analysis in which processes of reasoning that are accessible to rational individuals are represented explicitly. Lewis's concept of common knowledge is not simply (or even at all) a specification of what individuals know about one another's knowledge, but instead forms part of an account of their reasoning about one another's reasoning. Building on an analysis of Lewis's game theory by Cubitt and Sugden (2003), we formalise and extend Lewis's approach to represent how individual players may reason about the standards of *practical* – that is, decision-theoretic – rationality that they and other players endorse, and in this way reach conclusions about whether specific strategies are or are not rationally playable.

Our formalisation comprises a new class of 'common-reasoning models' for noncooperative games. In a model of this kind, players have access to a specific mode of reasoning, 'common reason', which constitutes the common rationality being modelled and which embeds some standard of practical rationality. Our formalisation of a mode of reasoning specifies what is taken as given, and what inferences are permitted, by that mode. For a given game and a given standard of practical rationality, the common-reasoning model specifies that this standard is taken as given by common reason, and specifies the inferences that common reason permits, and how the mode of reasoning of each individual relates to common reason. In doing so, the model represents explicitly the steps of reasoning by which players can arrive at conclusions about the rational permissibility or impermissibility of strategies. For a given strategy, there are three possibilities: *either* the permissibility of the strategy can be established by common reason; *or* its impermissibility can be so established; *or* neither its permissibility nor its impermissibility can be so established. In using this trinary partition of strategies, the solution concepts supported by our approach are quite different from those that are normally discussed in game theory.[1]

Many existing solution concepts (for example, those based on iterated deletion of weakly dominated strategies) can be described by algorithms – that is, step-by-step procedures for finding the relevant solution. In interpreting such algorithms, game theorists sometimes suggest that each player could find the solution for herself by working through the same steps, with the implication that the algorithm might track the players' own reasoning. But the Lewisian project of representing CKR *as reasoning* requires more than this. If what is to be represented is common knowledge *of rationality*, that reasoning must be capable of being understood as rational or valid. Puzzling or paradoxical features of conventional solution concepts are often mirrored by steps in the corresponding algorithms that seem not to be describable as steps of valid reasoning. (See Cubitt and Sugden, 2011, for discussion of this.) In contrast, our concept of a common-reasoning model is built out of explicitly-defined axioms and inference rules, the rationality or validity of which can be assessed independently of their roles in generating particular solutions.

We show that, for any game and any coherent standard of practical rationality, the relevant common-reasoning model provides a consistent rendition of CKR. By doing so, we

[1] The spirit, but not the formal structure, of our approach has some affinities with that of Binmore's (1987, 1988) analysis of 'eductive reasoning', further developed by Anderlini (1990). In Binmore's model, each player is represented by a Turing machine. In order to make a rational choice among strategies, each machine attempts to simulate the reasoning of the other machines. Binmore interprets the resulting infinite regress as demonstrating that 'perfect rationality is an unattainable ideal' (1987, pp. 204-209). Bacharach (1987) presents a related argument, questioning whether, even in games with unique Nash equilibria, the playing of equilibrium strategies can always be justified by the players' own reasoning. Like Binmore and Bacharach, we ask what conclusions players can reach by their own reasoning, without presupposing any general properties that those conclusions should satisfy. However, our Lewisian method of modelling reasoning is very different from those used by Binmore and Bacharach, and allows us to derive positive results about the conclusions players *can* reach.

achieve a complete separation between what it is for some conception of practical rationality to be common knowledge and the substantive content of that conception. Since a standard of practical rationality can coherently include appropriately formulated principles of caution, our approach guarantees the compatibility of such principles with CKR. Moreover, for any internally consistent conception of rationality (cautious or otherwise), our approach has a built-in defence against any charge that the conclusions players draw about the rational playability or otherwise of particular strategies are paradoxical. For every common-reasoning model, we can not only show that the players' reasoning is consistent; we can also show, for every strategy whose permissibility or impermissibility is established, a specific line of reasoning which players might use to reach that conclusion. Although we will show that this line can conveniently be tracked using an algorithm that we present, it is the common-reasoning model, and not the algorithm, which represents the reasoning.

The distinctiveness of the Lewisian approach can best be understood by comparison with more conventional representations of CKR. The approach to modelling CKR that has been seen by most game theorists as canonical is the formalisation due to Aumann (1987, 1999a, 1999b). Aumann calls his approach 'Bayesian'. For convenience, we will follow him in using that term in this way.[2] An important objective of our paper is to explain the fundamental differences between the two approaches.

Aumann (1987) offers a Bayesian framework which he sees as providing formal foundations for a solution concept, correlated equilibrium, which generalises Nash equilibrium. The central assumption of the model is that 'it is common knowledge that all the players are Bayesian utility maximizers' (p. 2), which Aumann treats as synonymous with there being 'common knowledge of rationality' (p.12). The model describes a situation in which, at every state of the world, each player's choices are decision-theoretically rational, given her beliefs, and represents that situation as common knowledge. Any model of this kind implies a binary partition of the set of strategies: one element of this partition contains those strategies that are played in *some* state(s) of the world, while the other contains those that are played in *none*. At every state, it is common knowledge that strategies that belong to the second element are not played. Crucially, however, this modelling strategy does not attempt to specify the steps of reasoning by which the players might discover the partition for

[2] In doing so, we are not taking any position about Bayesian doctrine. Some theorists may think that other formulations of CKR are more 'Bayesian' than Aumann's. For example, Aumann's approach uses concepts of both belief and knowledge, but some theorists might claim that a Bayesian approach should be entirely

themselves. On a natural interpretation, there is an implicit assumption that each player arrives at this partition by some process of reasoning from premises that represent (even if they are not limited to) the idea that 'Bayesian utility maximisation' is common knowledge. But, this process is not itself modelled. Thus, the Aumann approach is quite different from the Lewisian one that we develop, in which premises and steps of reasoning are specified explicitly.

This fundamental feature of Aumann's original model is retained in many subsequent developments of his approach. For example, Aumann (1999a) provides a 'dictionary' which allows the set-theoretic concepts of the original 'semantic' model to be translated into 'syntactic' equivalents. The syntactic version of the model allows propositions about the world, and about individuals' knowledge about the world, to be represented by linguistic formulae. The set-theoretic axioms of the semantic model are translated into axioms which require consistency among knowledge propositions and between knowledge and truth. The result is a syntactic description of *what* individuals know in the situation that is being modelled, satisfying certain properties of internal consistency. There is still no representation of the reasoning by which individuals *arrive at* this knowledge.

Although Aumann's model is logically consistent, apparently natural extensions of it, intended to introduce different conceptions of practical rationality involving principles of admissibility or caution, turn out to generate puzzles and even contradictions in some games (Börgers and Samuelson, 1992; Samuelson, 1992; Cubitt and Sugden, 1994). Given the generality of our results about the consistency of common-reasoning models, it is natural to ask whether Aumann's way of modelling CKR is vulnerable to paradoxes in some way that the Lewisian approach is not. Our paper explores this issue.

In doing so, we recognise that Aumann's formulation of CKR can be amended in ways that avoid some of the paradoxes associated with admissibility and weak dominance. For example, Monderer and Samet (1989) use a set-up similar to that of Aumann (1987), and explore the properties of a concept of 'almost common knowledge' as complete common knowledge is approached. A different way of representing 'almost certain' beliefs is developed by Brandenburger (2007) and Brandenburger *et al*. (2008), who amend Aumann's model by using lexicographic probability systems in place of Bayesian probabilities. Each of these adaptations amends some features of Aumann's formulation of the epistemic component

subjective. Aumann uses only standard probabilities, while some more recent 'Bayesian' models allow lexicographic probability systems. For reviews of different approaches, see Bonanno (2012) and Perea (2012).

of CKR, and avoids some of the paradoxes associated with that formulation; but neither attempts to represent the steps of reasoning by which players arrive at the propositions that they believe to be almost certainly true. Thus, they represent different approaches to the modelling of CKR from the Lewisian one that we adopt.[3]

Since our focus is on differences between the Lewisian approach and core features of Aumann's approach, clarity is served by taking his canonical model as our comparator. By supplementing Aumann's assumptions about rationality with a principle of caution, and by adding an assumption about the absence of correlation between the strategies of different players, we create a specific type of Bayesian model (an 'ICEU Bayesian model') which generates specific versions of the paradoxes of combining caution with Aumann's conception of CKR. Later, we use these paradoxes as exhibits and as test cases for our Lewisian approach. Because the concept of an ICEU Bayesian model is a simple and natural extension of Aumann's model, an investigation of the relationship between ICEU Bayesian models and their common-reasoning analogues sheds light both on the differences between the two modelling strategies and on the sources of the paradoxes.

The remainder of the paper is organised as follows: Section 2 presents Aumann's Bayesian approach to the representation of CKR. Section 3 extends this approach to capture a conception of practical rationality in which rational individuals maximise expected utility in relation to beliefs that are independent and cautious, in the sense of Pearce (1984) and Börgers and Samuelson (1992). We show that this extension has puzzling implications in some games and generates contradictions in others.

Our Lewisian approach is presented in Sections 4–8. Sections 4–6 introduce successively its major ingredients. Section 7 defines the class of common-reasoning models and then establishes the consistency of every such model. Section 8 introduces a sense in which a given common-reasoning model defines a 'solution' to the game, and defines a 'recommendation algorithm' which identifies that solution and which is interpretable as tracking specific steps of reasoning that lead common reason to it.

As the primitives of our common-reasoning models are very different from those of the Bayesian models introduced in Section 2, it helps to define a framework within which they can be compared. We present such a framework in Section 9, exploiting concepts

[3] Lewis's analysis is sometimes reconstructed in ways that make it more akin to Aumann's approach (e.g. Vanderschraaf, 1998; Sillari, 2005; Gintis, 2009; Paternotte, 2011). The approach we follow here is in line with

introduced in Cubitt and Sugden (2011). Section 10 then specialises the common-reasoning framework to the case where common rationality embodies the conception of practical rationality that gives rise to paradoxes within the Bayesian approach of Section 2. Using the framework of Section 9 as a bridge, Section 10 establishes precise relationships between the corresponding Bayesian and common-reasoning models. These relationships provide the ingredients for a resolution in Section 11 of the paradoxes presented in Section 3. Section 12 concludes. Between them, two appendices provide proofs of all the formal results.

2. Common knowledge of rationality in a Bayesian model

In this section, we present a Bayesian model of CKR, based on that of Aumann (1987).

We consider the class G of finite, normal-form games of complete information, interpreted as one-shot games. For any such game, there is a finite set $N = \{1, ..., n\}$ of *players*, with typical element i and $n \geq 2$; for each player i, there is a finite, non-empty set of (pure) *strategies* S_i, with typical element s_i; and, for each profile of strategies $s = (s_1, ..., s_n)$, there is a profile $u(s) = (u_1[s], \ldots, u_n[s])$ of real-valued and finite *utilities*. The set $S_1 \times ... \times S_n$ is denoted S; the set $S_1 \times ... \times S_{i-1} \times S_{i+1} \times ... \times S_n$ is denoted S_{-i} and its typical element by s_{-i}. We impose that, for all distinct $i, j, \in N$, $S_i \cap S_j = \varnothing$. This condition has no substantive significance, but imposes a labelling convention that the strategies available to different players are distinguished by player indices, if nothing else. This convention allows a conveniently compact notation in later sections.

Aumann's modelling strategy is Bayesian in the sense that agents' beliefs are described by subjective probabilities, defined on some set of states and updated from initial priors in the light of known information. Thus, we define a Bayesian model, for any game in G, so that it specifies all of the following: a set of states of the world; players' behaviour; players' knowledge; players' subjective beliefs; and a standard of decision-theoretic rationality.

Uncertainty is represented by means of a finite, non-empty, universal set Ω of *states*, whose typical element is denoted ω. A set of states is an *event*.

the view elaborated in Cubitt and Sugden (2003) that reconstruction of Lewis's analysis in a framework akin to Aumann's is liable to edit out much of the originality of Lewis's game theory.

Players' behaviour is represented by a *behaviour function* $b(.)$, which assigns a profile of strategies $b(\omega) = (b_1[\omega], ..., b_n[\omega])$ to each state ω, to be interpreted as the profile of strategies that are in fact chosen by the players at ω. Stochastic choice (such as mixed strategies) is represented as choice that is conditioned on random events. For each profile s of strategies and each strategy s_i, we define the events $E(s) = \{\omega \in \Omega \mid b(\omega) = s\}$ and $E(s_i) = \{\omega \in \Omega \mid b_i(\omega) = s_i\}$. Let $S^* = \{s \in S \mid E(s) \neq \varnothing\}$ and $S_i^* = \{s_i \in S_i \mid E(s_i) \neq \varnothing\}$. S^* (respectively S_i^*) is the set of strategy profiles (respectively strategies for i) *included* in the Bayesian model. Thus, a Bayesian model specifies a binary partition of each player's strategy set S_i, the elements of which are the set of included strategies S_i^* and the set of excluded strategies $S_i \backslash S_i^*$. By construction, each S_i^* is non-empty.

Players' knowledge is represented by an *information structure* $\mathcal{I} = (\mathcal{I}_1, ..., \mathcal{I}_n)$. For each player i, $\mathcal{I}_i$ is an information partition of Ω, representing what i knows at each state. $K_i(E)$, where E is an event, is the event $\{\omega \in \Omega \mid \exists E' \in \mathcal{I}_i$: $(\omega \in E')$ and $(E' \subseteq E)\}$.[4] If $\omega \in K_i(E)$, we say 'i knows E at ω'. An event E is *Bayesian common knowledge* at ω if ω is an element of all events of the finitely-nested form $K_i(K_j(... K_k(E)...))$. (This is the formal definition of 'common knowledge' used in the model. We use the qualifier 'Bayesian' to distinguish this theoretical construct from the intuitive concept.) Since Ω is the universal set, it follows that, for every player i and every state ω, i knows Ω at ω. Thus, Ω is Bayesian common knowledge at every state.

For any player i, a *prior* is a function $\pi_i: \Omega \rightarrow [0, 1]$ satisfying the conditions (i) $\Sigma_{\omega \in \Omega} \pi_i(\omega) = 1$; and (ii) for every event $E \in \mathcal{I}_i$, there exists some state $\omega \in E$, such that $\pi_i(\omega) > 0$. $\pi_i(\omega)$ is interpreted as a subjective probability. We extend this notation to events by defining, for each event E, $\pi_i(E) = \Sigma_{\omega \in E}\, \pi_i(\omega)$. Posterior probabilities, conditional on events, are defined from priors by means of Bayes's rule. Condition (i) is the obvious condition that prior probabilities sum to unity; condition (ii) guarantees that posterior probabilities, conditional on what player i knows at any state, are well-defined. The latter condition is required if the standard of rationality developed below and drawn from Aumann (1987) is itself to be well-defined.

To represent the normative standard of practical rationality that is to be built into the model, we define a *choiceworthiness function* for each player i as a function $\chi_i: \Omega \rightarrow \wp(S_i)$, where $\wp(S_i)$ denotes the power set of S_i, satisfying two restrictions. First, $\chi_i(\omega)$, the set of strategies that are *choiceworthy* for i at ω, is nonempty for all ω. Second, for all $E \in \mathcal{I}_i$, for

[4] We use $\subset$ (resp. $\subseteq$) to denote 'is a strict (resp. weak) subset of'.

all $\omega, \omega' \in E$: $\chi_i(\omega) = \chi_i(\omega')$. The interpretation is that $\chi_i(\omega)$ is the set of strategies which, according to the standard of rationality, may be chosen by i at ω. The first restriction stipulates that, in every state, there is at least one choiceworthy strategy; the second that what is choiceworthy for a player cannot differ between states that he is unable to distinguish.

In Aumann's case, the standard of rationality is subjective expected utility maximisation. Consider any player i. For any $s \in S$, for any $s_i' \in S_i$, let $\sigma_i(s, s_i')$ denote the strategy profile created by substituting s_i' for s_i in s (i.e. $\sigma_i[s, s_i'] = [s_1, ..., s_{i-1}, s_i', s_{i+1}, ..., s_n]$). For any prior π_i, for any state ω', for any $E \in \mathcal{I}_i$, let $\pi_i(\omega'|E)$ denote the posterior probability of ω', given E. For each player i, for each state ω, a strategy s_i is *SEU-rational* for i at ω with respect to the information partition $\mathcal{I}_i$ and prior π_i if, for each strategy $s_i' \in S_i$, $\sum_{\omega' \in E} \pi_i(\omega'|E)\ (u_i[\sigma_i(b[\omega'], s_i)] - u_i[\sigma_i(b[\omega'], s_i')]) \geq 0$, where E is the event such that $\omega \in E \in \mathcal{I}_i$. Thus, s_i is SEU-rational for i at ω if it maximizes expected utility for i, conditional on his prior beliefs updated by his information at ω.

We define a *Bayesian model* of a particular game as an ordered quintuple $<\Omega, b(.), \mathcal{I}, \pi, \chi>$, where Ω is a finite, nonempty set of states and $b(\omega) = (b_1[\omega], ..., b_n[\omega])$, $\mathcal{I} = (\mathcal{I}_1, ..., \mathcal{I}_n)$, $\pi = (\pi_1, ..., \pi_n)$ and $\chi = (\chi_1, ..., \chi_n)$ are, respectively a behaviour function, an information structure, a profile of priors and a profile of choiceworthinesss functions defined with respect to Ω and the game, such that the following three conditions are satisfied:

> *Choice Rationality*. For all $i \in N$, for all $\omega \in \Omega$: $b_i(\omega) \in \chi_i(\omega)$.
>
> *SEU-Maximisation.* For all $i \in N$, for all $\omega \in \Omega$: $\chi_i(\omega) = \{s_i \in S_i |\ s_i$ is SEU-rational at ω with respect to $\mathcal{I}_i$ and $\pi_i\}$.
>
> *Knowledge of Own Choice*. For all $i \in N$, for all $\omega \in \Omega$: $\omega \in K_i[E(b_i[\omega])]$.

Choice Rationality requires that, at each state, each player's actions are consistent with whatever standard of decision-theoretic rationality is being modelled. SEU-Maximisation stipulates that the standard of rationality is the maximisation of subjective expected utility. Knowledge of Own Choice imposes the obvious restriction that, at each state, every player knows the pure strategy that he chooses.[5] Our definition of a Bayesian model is equivalent to that of Aumann (1987), except that we do not impose (or rule out) Aumann's assumption of *common priors* (that is, that for all $i, j \in N$, $\pi_i = \pi_j$). Aumann (pp. 12–15) notes that, unlike

[5] This is consistent with randomisation by players since, as noted above, play of random strategies is represented in the model by prior uncertainty about which state obtains.

the other features of his model, it is not essential to a Bayesian representation of CKR; he imposes it only to generate sharp results.

The following result is implied by the analysis of Aumann (1987):

Theorem 1: For every game in G, a Bayesian model exists.

This theorem shows that, for every game in G, the concept of a Bayesian model is an internally consistent representation of CKR. In particular, in any such model, Ω is a universal set of states at each of which some profile of strategies is played that contains only choiceworthy strategies; and S^* is the set of profiles played at states in Ω. As Ω is Bayesian common knowledge at all states and $\Omega = \cup_{s \in S^*} E(s)$, there is Bayesian common knowledge at all states of the event that a profile in S^* is played.

3. Three puzzles

Given Theorem 1, it is natural to ask whether further conditions can be imposed on Bayesian models. Our concern is with two restrictions on player's beliefs. To formulate them, we use the following concepts: A prior π_i for player i is *independent* if, for each profile $s = (s_1, ..., s_n)$ of included strategies, $\pi_i(E[s]) = \Pi_{j \in N}\, \pi_i(E[s_j])$. We extend our notation for posterior probabilities to those on events by using $\pi_i(E'|E)$, where $E' \subseteq E$ and E is an element or union of elements of $\mathcal{I}_i$, to denote player i's posterior probability of E' given E. We can now state:

Independence (of Priors). For all $i \in N$: π_i is independent.

Caution (of Posteriors). For all distinct $i, j \in N$, for all $s_i \in S_i$, for all $\omega \in \Omega$: if $s_i \in \chi_i(\omega)$ then $\pi_j(E(s_i)|E) > 0$, where E is the event such that $\omega \in E \in \mathcal{I}_j$.

Independence rules out the possibility that some player i believes that the choices of players from among their included strategies are correlated with one another.[6] Although Aumann's (1987) Bayesian model of CKR allows correlation of strategies between players, game theory needs to be able to model situations in which the players have no mechanisms for achieving such correlation or grounds for believing in it. If the representation of CKR is to apply to such cases, it must be possible to impose Independence on the model.

[6] Note that each player's prior may be independent whether or not players have a common prior.

Caution requires that, if some strategy s_i is choiceworthy for player i at state ω then, at the same state, other players must assign strictly positive posterior probability to s_i being played. The motivation for the condition rests on two background assumptions. The first is the implicit assumption, integral to Aumann's approach, that each player knows, in an informal sense, the information partition and prior of each other player (see Aumann 1987, pp. 9–10). From this, player j should also be taken to know, in the same informal sense, the choiceworthy strategies for player i at every state ω. The second background assumption is that the tie-breaking mechanism that player i uses to discriminate between options which, according to the standard of rationality, are equally choiceworthy is private to him. Since tie-breaking occurs only when rationality fails to determine what should be chosen, the properties of a tie-breaking mechanism must be non-rational. If the representation of CKR is to apply to cases in which non-rational tie-breaking rules are private, it must be possible to impose on the model the background assumption that they are. To do this would be to require that, if s_i is choiceworthy for player i at state ω, then the element of (each other) player j's information partition that contains ω must contain some state(s) at which s_i is played. Given these background assumptions, Caution expresses the following principle of prudential belief for each player $j \neq i$: If s_i is choiceworthy for player i at state ω then, given what player j knows at ω, j's beliefs should allow that s_i might be played.

Note that there is no corresponding prudential argument in relation to j's beliefs about a strategy s_i that is not choiceworthy for i, as (by Aumann's implicit assumption) j is aware of such non-choiceworthiness. If j knows, in the informal sense, that s_i is not choiceworthy for i at *any* ω, then an Aumann-like understanding of CKR (captured by Choice Rationality and SEU-Maximisation) should not allow j to assign positive probability to s_i. By defining the requirement of Caution relative to choiceworthy strategies, we follow Börgers and Samuelson (1992) and Pearce (1984).[7] As with Independence, we are not suggesting that Caution is an implication of Bayesian rationality, but merely that it represents a case of potential interest for the modelling of CKR.

To be more precise, we interpret Bayesian models which satisfy both Independence and Caution as attempting to represent common knowledge of the following standard of practical rationality: each player's beliefs assign independent probabilities to other players' strategies, zero probability to strategies regarded as not rationally playable, and strictly

[7] This approach is distinct from that used by some others in the literature (e.g. Asheim and Dufwenberg, 2003; Perea, 2011), for whom caution requires that *no* strategy is regarded as entirely impossible.

positive probability to all strategies regarded as rationally playable; and each player maximises expected utility relative to these beliefs. We call this standard that of *independent cautious expected utility maximization* (or the *ICEU standard*, for short).[8] Thus, a Bayesian model which satisfies Independence and Caution is an *ICEU Bayesian model.*

It would be puzzling if an otherwise coherent representation of CKR could not accommodate the view of rationality embedded in the ICEU standard without giving rise to paradoxes or impossibility. However, that is how matters turn out. We use three games to illustrate this.

Our first exhibit, illustrating the *Proving Too Much Paradox,* is Game 1.[9]

Game 1:

		Player 2	
		left	*right*
	first	0, 0	0, 0
Player 1	*second*	– 1, 3	2, 2
	third	–1, 3	1, 5

Proposition 1: ICEU Bayesian models of Game 1 exist; and, in every such model, $S_1{*} = \{first\}$ and $S_2{*} = \{left, right\}$.

Proposition 1 is paradoxical because $S_1{*} = \{first\}$ implies that *first* is choiceworthy at every state and the only strategy that Player 1 may play. The strategy *first* can only be choiceworthy at every state if Player 1's posterior probability on *left* is always greater than or equal to 2/3 (since otherwise *second* would be strictly better than *first*). But, at every state, Player 1 knows that Player 2 is indifferent between *left* and *right.* Clearly, player 1 *might* believe that there is a tendency for player 2's tie-breaking mechanism to resolve in favour of *left* even when, as in the situations that ICEU Bayesian models are intended to represent, that mechanism is private. The puzzling feature of Proposition 1 is that it implies that, even when the tie-breaker is private, Player 1 *must* believe it to have that tendency. In this sense, we seem to have proved too much.

[8] The ICEU standard is very closely related to that described in Section 5 of Cubitt and Sugden (2011), where an independence condition is added to the 'reasoning-based expected utility' conception of practical rationality introduced in Section 3 of that paper. We use a different name here, to avoid associating the standard itself with any particular approach to modelling CKR.

[9] Game 1 is the normal-form of a simple extensive-form 'Centipede' game in which the initial move belongs to Player 1, the second move to Player 2 and the third and final move to Player 1. Although Centipede games have most often been discussed in the literature using the extensive form, our analysis here uses the normal form only.

In more general terms, the Proving Too Much Paradox takes the following form. The property that some specific strategy is (or is not) choiceworthy holds for *all* ICEU Bayesian models, with the apparent implication that common knowledge of this property is implied *merely* by the assumption that the ICEU standard of rationality is common knowledge; but there seems to be no way that the players could arrive at that conclusion, using only the reasoning resources attributed to them by that assumption.[10]

Our second exhibit illustrates another way in which a Bayesian modelling approach can seem to generate paradoxical implications. This game is named in memory of a method of marking lanes on single-carriageway roads which was once common in Britain. There were three lanes, one for slow traffic in each direction, and a central lane designated for overtaking in both directions. The players are drivers travelling in opposite directions who have simultaneous overtaking opportunities; each is indifferent between staying in his slow lane and overtaking successfully. Each player *i* chooses whether to pull out to overtake (out_i) or not to do so (in_i):

Game 2: (Three-lane Road)

		Player 2	
		in_2	out_2
Player 1	in_1	1, 1	1, 1
	out_1	1, 1	0, 0

> *Proposition 2:* ICEU Bayesian models of Game 2 exist; and, in every such model, *either* (i) $S_1^* = \{in_1\}$ and $S_2^* = \{in_2, out_2\}$ *or* (ii) $S_1^* = \{in_1, out_1\}$ and $S_2^* = \{in_2\}$.

Proposition 2 implies that, in every ICEU Bayesian model of Game 2, one of the two players plays the ‘risky’ strategy (*out*) in some states, while the other plays it in none. The structure of the game is entirely symmetrical with respect to the two players. But, if one player plays *out* in some states and the other plays it in none, there must be some asymmetry that tells one and only one player that they may play *out*. The apparent implication of Proposition 2 is that the existence of such an asymmetry is implied merely by the assumption

[10] It might be objected that ‘proving too much’ in a model of CKR is a contradiction in terms, as an ideally rational player would be able to reproduce any proof of a property of the model. In the current context, the objection would take the form that, if the players have the reasoning resources attributed to them by the assumption that the ICEU standard of rationality is common knowledge, they can work out that *first* is the unique choiceworthy strategy for Player 1 in Game 1 by reproducing our proof of Proposition 1. This objection rests on a misunderstanding. Even supposing players can reconstruct the proof of Proposition 1, (a supposition which, though coherent, goes beyond what can be represented in a Bayesian model), all they could thereby conclude is Proposition 1 itself, i.e. a claim about ICEU Bayesian models, not about Player 2’s tie-breaker.

that the ICEU standard of rationality is common knowledge; but there seems to be no way in which the players could discover that asymmetry using only the knowledge attributed to them by that assumption, since both the assumption and the game are symmetric. Of course, there is no paradox in the idea that there *could* be common knowledge of an asymmetry in what rationality requires of the players, grounded on information external to the formal description of the game. The *Three-lane Road Paradox* is the apparent demonstration that a conception of CKR implies that there *must* be such knowledge.

Neither Game 1 nor Game 2 yields an outright inconsistency in the conditions that define an ICEU Bayesian model. In fact, these conditions are mutually consistent for every two-player game in *G*.[11] However, an inconsistency can be shown using a game introduced by Cubitt and Sugden (1994), which can be thought of as a three-player extension of Game 2.[12] We call the inconsistency shown by this game the *Tom, Dick and Harry Paradox* to match the story described in that paper. Tom (player 1), Dick (player 2) and Harry (player 3) are guests in an isolated hotel. Tom is trying to avoid Dick, Dick to avoid Harry, and Harry to avoid Tom; yet, there is no alternative to taking their evening meal in the hotel. Guests who eat in the restaurant (*out*) will meet each other, whereas those who eat in their rooms (*in*) will not meet any others.

Game 3 (Tom, Dick and Harry)

Player 3: in_3

		Player 2	
		in_2	out_2
Player 1	in_1	1, 1, 1	1, 1, 1
	out_1	1, 1, 1	0, 1, 1

Player 3: out_3

		Player 2	
		in_2	out_2
Player 1	in_1	1, 1, 1	1, 0, 1
	out_1	1, 1, 0	0, 0, 0

[11] This can be proved by exploiting the existence proof for quasi-strict Nash equilibrium for two-player games due to Norde (1999). Given a quasi-strict Nash equilibrium of a game, a Bayesian model of that game can be constructed, using the technique in our proof of Theorem 1. The properties of quasi-strict Nash equilibrium ensure that Independence and Caution are satisfied.

[12] The reader may wonder why a three-player example is needed, as we motivated our Caution condition with reference to the conception used by Börgers and Samuelson (1992) in defining their "consistent pairs". Börgers and Samuelson present a 2-player game with no consistent pair. Though related, the concept of a consistent pair is not identical to that of an ICEU Bayesian model even for a 2-player game. ICEU Bayesian models do exist for Börgers and Samuelson's example. For detail on the relationship between Cubitt and Sugden (1994) and Börgers and Samuelson (1992), see the expanded version of the former that appeared as Cubitt and Sugden (1997), and Squires (1998).

The paradox consists in the fact that there is no ICEU Bayesian model of Game 3, which constitutes a proof of the following result:

Theorem 2: There are games in *G* for which no ICEU Bayesian model exists.

Theorem 2 establishes that there is at least one standard of rationality, namely ICEU, that cannot be represented without contradiction in a Bayesian model. However, we also know from Theorem 1 that there is at least one standard, namely SEU-maximisation, that *can* be so represented. So should one simply conclude that the Bayesian modelling strategy has been vindicated and that ICEU is unacceptable? We think not.

Considered in its own right (rather than in the context of Bayesian modelling), the view of rationality embedded in ICEU seems to be internally coherent. The normative issue of adjudicating between alternative standards of rationality seems orthogonal to the modelling issue of how to represent a world in which some standard of rationality is common knowledge. Thus, whether or not one believes that obeying the ICEU standard is a normative requirement of rationality, it is puzzling that common knowledge of that standard cannot always be represented in a Bayesian model. Of course, it is open to a game theorist to treat the Bayesian modelling strategy as a fixed point, and to stipulate that the only relevant standards of rationality are those that can be represented in such models. But that would leave unanswered the theoretical question of *why* the Bayesian approach cannot represent common knowledge of an apparently coherent standard of rationality. And in the absence of a satisfactory answer to that question, privileging that approach seems an arbitrary manoeuvre, and all the more so if common knowledge of the ICEU standard *can* be represented by a different approach.

4. Reasoning schemes

We now develop our Lewisian rendition of CKR in which players' reasoning is represented explicitly. Such representation requires different primitives. As a first step, we introduce our representation of a mode of reasoning.

We model a mode of reasoning as a structure defined in relation to some domain *P* of *sentences*, to be interpreted as well-formed formulae of a formal language. Implicitly, these sentences express propositions which may either be taken as given or inferred within a given mode of reasoning. We use p, q, r as *sentence variables*, that is as placeholders for unspecified sentences in P. In this Section, we impose only very general conditions on P. The logical connectives $\neg$, $\wedge$, and $\Rightarrow$, defined by the rules of classical logic are used, respectively, for negation, conjunction and material implication. We impose throughout that P is closed under the formation of negation, conjunction and material implication, i.e. that if $p, q \in P$ then $\neg p, p \wedge q, p \Rightarrow q \in P$.

An *inference rule* in domain P is a two-place instruction of the form «from ..., infer ... », where the first place is filled by a non-empty, finite subset of P (whose elements are the *premises* of the rule) and the second place by an element of P (the *conclusion* of the rule).

An *inference structure* is a triple $R = \langle P, A(R), I(R) \rangle$, where P is the domain in which reasoning takes place, $A(R) \subseteq P$ is the set of *axioms* of R, and $I(R)$ is a set of inference rules in domain P. The set $T(R)$ of *theorems* of R is defined inductively as follows. We define $T_0(R) = A(R)$. For $k \geq 1$, $T_k(R)$ is defined as $T_{k-1}(R) \cup \{p \in P \mid p$ is the conclusion of an inference rule in $I(R)$, all of whose premises are in $T_{k-1}(R)\}$. Then $T(R) = T_0(R) \cup T_1(R) \cup \ldots$.

Below, we will use a particular class of inference structures called 'reasoning schemes' to represent modes of reasoning that are accessible to the players of a game. The game-theoretic content of a reasoning scheme R will be represented by properties of P, $A(R)$ and $I(R)$. P will be defined in terms of a particular formal language, in which sentences refer to properties of strategies in a game; and principles of game-theoretic reasoning will be represented as axioms and inference rules of R. Thus, we will include inference rules which represent such principles, but are not licensed by classical logic – for example, rules which infer what it would be rational for one player to do from a prediction about what another player will do. Our main interest is in the implications for players' reasoning of specific game-theoretic axioms and inference rules. But since we are modelling rationality, it is incumbent on us also to endow players with the capability to make inferences that *are* licensed by classical logic. As a means of achieving this, we proceed as follows.

Throughout our analysis, we will use the term 'logically' only to refer to operations licensed by classical logic. Thus, we will say, for any non-empty, finite subset $Q = \{q_1, ..., q_m\}$ of P and any $p \in P$, that p is *logically entailed* by Q if $q_1 \wedge ... \wedge q_m \wedge \neg p$ is a contradiction in classical logic (i.e. if every truth-value assignment renders it false); and that an inference rule is *logically valid* if its conclusion is logically entailed (in this sense) by the set of its premises. For any inference structure R, we will say that $I(R)$ *contains the rules of valid inference* if, for every non-empty, finite $Q \subseteq P$ and every $p \in P$, if p is logically entailed by Q then «from Q, infer p» $\in I(R)$. We will say that a set of sentences is *consistent* if it has no finite subset the conjunction of all elements of which is a contradiction in classical logic.

An inference structure R, such that $A(R)$ is non-empty and $I(R)$ contains the rules of valid inference, is a *reasoning scheme*. Thus, for every reasoning scheme R, every tautology in P is an element of $T(R)$. We will say that a reasoning scheme R is *consistent* if $T(R)$ is consistent. Our aim is to model CKR in terms of reasoning schemes that are consistent. But, to demonstrate the feasibility of this goal, we need to use a modelling framework in which consistency can be proved. For this reason, we do not impose consistency as part of the definition of a reasoning scheme.

We will say that a person *endorses* a reasoning scheme R if he takes its axioms to be true and accepts the authority of its inference rules; a person who endorses R has *reason to believe* each of its theorems. Our analysis will be about what players have reason to believe about one another's strategy choices, given that they endorse particular kinds of reasoning schemes. Because we take this approach, we will never need to consider whether the theorems of some reasoning scheme are 'really' true or false.

5. Common reasoning in a population

Our approach is to model CKR among a population of agents as the existence of a core of shared reasoning which is endorsed by each agent in the population and is commonly attributed to other such agents. It allows us to represent the idea that each individual maintains a distinction between (on the one hand) what *everyone* has reason to believe, given the axioms and inference rules that everyone endorses and (on the other hand) what *he* has reason to believe, given the axioms and inference rules that he endorses. Thus, given a finite, non-empty, *population* $N = \{1, \ldots, n\}$ of agents, we postulate the existence, for each agent i, of a reasoning scheme R_i of *private reason* which i endorses, and the existence of a reasoning scheme R^* of *common reason*. Of each sentence p that is a theorem of R^*, we will say that there is *common reason to believe p*.

As a step in defining the formal language of our models, we introduce the following piece of syntax. If p is a sentence and R is a reasoning scheme, $R(p)$ is a sentence to be read as 'p is a theorem of R'. Note that $R(p)$ is a sentence that can feature as an axiom or theorem of a reasoning scheme; it is not how we will express a statement, made from our viewpoint as modellers, about the properties of R. When *we* assert that p is a theorem of R, we use the notation $p \in T(R)$. Thus, for example, if R_1 and R_2 are (possibly distinct) reasoning schemes, $R_2(p) \in T(R_1)$ denotes that the sentence $R_2(p)$, which expresses the proposition that p is a theorem of R_2, is in fact a theorem of R_1.

We take as given a non-empty set P_0 of *primitive sentences*, such that no sentence in P_0 contains any of the terms $R^*(.)$, $R_1(.)$,..., $R_n(.)$. For each $k \geq 1$, we define P_k to contain all of the following sentences (and no others): (i) every sentence which can be constructed from the elements of P_{k-1} using a finite number of connectives from the set $\{\neg, \wedge, \Rightarrow\}$, (ii) every sentence of the form $R^*(p)$ where $p \in P_{k-1}$; and (iii) every sentence of the form $R_i(p)$ where $i \in \{1, ..., n\}$ and $p \in P_{k-1}$. We define $\varphi(P_0) \equiv P_0 \cup P_1 \cup ...$. For any given specification of P_0, $\varphi(P_0)$ is the domain in which the reasoning schemes of our model operate.

We now define the following concept as a representation of the links between private and common reason. An *interactive reasoning system* among the population $N = \{1, \ldots, n\}$ is a triple $< P_0, R^*, (R_1, \ldots, R_n)>$, where P_0 is a set of primitive sentences, R^* is a reasoning scheme, and $(R_1, ..., R_n)$ is a profile of reasoning schemes, such that each of the $(n+1)$ reasoning schemes has the domain $\varphi(P_0)$ and the following conditions hold:

> *Awareness:* For all $i \in N$, for all $p \in \varphi(P_0)$: if $p \in T(R^*)$ then $R^*(p) \in A(R_i)$.
>
> *Authority:* For all $i \in N$, for all $p \in \varphi(P_0)$: «from $\{R^*(p)\}$, infer p» $\in I(R_i)$.
>
> *Attribution (of Common Reason)*: For all $i \in N$, for all $p \in \varphi(P_0)$: «from $\{p\}$, infer $R_i(p)$» $\in I(R^*)$.

We will say that an interactive reasoning system $<P_0, R^*, (R_1, ..., R_n)>$ is *consistent* if each of its component reasoning schemes is consistent.

The Awareness condition stipulates that each agent's private reasoning has access to the theorems of common reason. The Authority condition stipulates that, from the premise that some sentence p is a theorem of common reason, each agent's private reason infers p. The Attribution condition stipulates that common reason attributes its own theorems to the private reason of each agent. Awareness and Authority together imply that each theorem of common reason is also a theorem of the private reason of each agent i. Thus, we may think of common reason, not as involving any unexplained collective consciousness, but simply as a sub-routine that each agent can individually use to generate certain theorems. It is convenient to identify this sub-routine using the device of R^*, and then to impose Attribution upon it, in order to capture the idea that certain conclusions are reached by all agents in the population in the same way and also attributed by each of them to each other.

We will say that there is *iterated reason to believe p* in population N if the proposition $R_j(... R_k(p)...) \in T(R_i)$ is true, for all finitely nested sentences $R_j(... R_k(p)...)$ and all $i, j, ..., k \in N$. The following theorem establishes that, in an interactive reasoning system, there is iterated reason to believe all sentences that are theorems of R^*:

> *Theorem 3*: Consider any population N of agents and any interactive reasoning system $<P_0, R^*, (R_1, \ldots, R_n)>$ among the population N. For every sentence $p \in T(R^*)$, there is iterated reason to believe p in population N.

Iterated reason to believe is the closest Lewisian analogue of Aumann's concept of common knowledge. Theorem 3 allows our analysis of CKR to be carried out almost entirely in terms of R^*. Nevertheless, private reasoning schemes are an essential part of our modelling strategy, since it is only by virtue of the connections between private and common reason that we can claim that an analysis of R^* is informative about iterated reason to believe.

There is a close affinity between Theorem 3 and Lewis's (1969, pp. 52–60) analysis of common knowledge. Lewis defines p to be 'common knowledge' in a population N if some 'state of affairs' A holds, such that (i) everyone in N has reason to believe that A holds, (ii) A 'indicates' to everyone in N that everyone in N has reason to believe that A holds, and (iii) A 'indicates' to everyone in N that p. He defines 'A indicates to person i that p' as 'if i has reason to believe that A holds, i thereby has reason to believe that p'. He sketches a proof of the theorem that if p is common knowledge in this sense, and given (not fully specified) premises to the effect that individuals share, and have reason to believe that they share, certain principles of rationality, inductive standards and background information, there is iterated reason to believe p in N. Cubitt and Sugden (2003) reconstruct the theorem and its proof, using an explicit specification of the properties of 'indication', motivated by interpreting 'i has reason to believe that p' as saying that p is treated as true in a mode of reasoning that i endorses and 'A indicates to i that p' as saying that, in that mode of reasoning, there is an inference from 'A holds' to p.

The same ideas can be represented more directly in an interactive reasoning system, in which 'shared' background information and principles of rationality are represented by axioms and inference rules of R^*. For this purpose, we treat p and 'A holds' as sentences in the formal language of that interactive reasoning system. That some state of affairs *A* is such that, if it occurs, its occurrence is public and self-evident can be represented by the material implication that if (in fact) A holds, then '*A* holds' $\in A(R^*)$. That there are common standards of inductive inference such that, if there is common reason to believe that *A* holds, there is thereby common reason to believe *p* can be represented by «from {'*A* holds'}, infer *p*» $\in I(R^*)$. Given these conditions, it follows from Theorem 3 that if (in fact) *A* holds, then there is iterated reason to believe *p* in *N*.

Our method of modelling CKR in a given game will be to represent practical and game-theoretic rationality in terms of axioms and inference rules, and to attribute these to common reason in an interactive reasoning system among the population comprising the players of the game. By virtue of Theorem 3, any sentences that can be derived using those axioms and inference rules will be the object of iterated reason to believe among the players.

6. Decision rules: practical rationality expressed by sentences

In this section, we develop a general method of representing principles of practical rationality for a game in the form of a particular kind of sentence, which we call a 'decision rule'. This concept uses a purely formal notion of 'permissibility', to be interpreted as permissibility with respect to whatever principles of practical rationality are to be represented.

Here, and throughout Sections 6–9, we fix a given game in *G*. Our analysis applies to any such game but we suppress phrases of the form 'for all games in *G*' except in formal results. Differences between games become important again only in Sections 10 and 11.

We now add two further pieces of syntax to our formal language. For every player *i* and for every $s_i \in S_i$, $p_i(s_i)$ is a sentence to be read as 's_i *is permissible for i*' and interpreted as stating that, normatively, *i* may choose s_i (but not that he must, since two or more strategies might be permissible for him). For every player *i* and for every $s_i \in S_i$, $m_i(s_i)$ is a sentence to be read as 's_i *is possible for i*' and interpreted as stating that s_i might in fact be chosen by *i*. Sentences of the form $p_i(s_i)$ or $\neg p_i(s_i)$ (the latter read as 's_i is impermissible for *i*') are *permissibility sentences*. For each permissibility sentence $p_i(s_i)$ or $\neg p_i(s_i)$, the corresponding *possibility sentence* $m_i(s_i)$ or $\neg m_i(s_i)$ (the latter read as 's_i *is impossible for i*') is its *correlate*, and vice versa.

We will say of any conjunction of sentences that it *states* each of its conjuncts. A *prediction* about a player i is a conjunction of the elements of a consistent set of possibility sentences referring to the strategies available to i, satisfying the conditions that (i) not every strategy in S_i is stated to be impossible; and (ii) that, if every strategy but one in S_i is stated to be impossible, the remaining strategy is stated to be possible. Notice that (ii) is not redundant because, in general, a prediction about i may not refer to every strategy available to i. (Since we want to be able to represent how reasoning can proceed in successive steps from initial premises to progressively richer conclusions, we need to be able to represent predictions that are incomplete in this sense.) Conditions (i) and (ii) express the presumption that, as S_i exhausts the options available to i, it cannot be the case that all its elements are impossible; and, if every element but one is impossible, that establishes that the remaining one is possible (indeed, certain). A conjunction of the elements of a non-empty set of predictions about individual players other than i, where that set contains no more than one non-null prediction about any such player, is a *collective prediction* about $N\backslash\{i\}$.

Analogously with the concept of prediction, a *recommendation* to a player i is a conjunction of the elements of a consistent set of permissibility sentences referring to the strategies available to i, satisfying analogues of conditions (i) and (ii); here the presumption that those conditions embody is that normative requirements must be logically capable of being satisfied.

The definition of a correlate is extended to recommendations and predictions, so that for each recommendation there is a unique correlate prediction and vice versa: the correlate of a recommendation (resp. prediction) is the conjunction of the correlates of its component permissibility (resp. possibility) sentences.[13]

We also need to represent sentences that assert nothing substantive about what player i might or might not play or, correspondingly, about what would or would not be normatively permissible for i. Sentences of the first kind play a crucial initiation role in our model of reasoning as, intuitively put, we need to represent how a player 'starts' with no substantive conclusions about what another player may or may not play and only 'reaches' such conclusions though steps of reasoning. We therefore define a further piece of syntax. We use an arbitrary tautology, denoted by #, to represent a sentence with no content – the *null*

[13] As part of the definition of the correlate of a recommendation (resp: prediction), we require that the order of the component possibility (resp: permissibility) sentences in the correlate matches that of the component permissibility (resp: possibility) sentences in the recommendation (resp: prediction).

sentence. Thus, # can be a null recommendation, a null prediction, or null collective prediction. In view of this, the correlate of # is #.

Recommendations to a player *i*, collective predictions about the set of players $N\backslash\{i\}$, and predictions about *i* are sentences that have special roles to play in what follows. To distinguish them from other sentences, we use the sentence variables y_i for recommendations to *i*, x_{-i} for collective predictions about $N\backslash\{i\}$, and z_i for predictions about *i*. Using this notation, a *maxim* for player *i* is a material implication $x_{-i} \Rightarrow y_i$. The interpretation is that, conditional on the prediction x_{-i} about the behaviour of players other than *i*, the permissibility sentences stated by y_i are mandated by some conception of practical rationality. Note that the maxim $\# \Rightarrow y_i$ is logically equivalent to the recommendation y_i.

A *decision rule* for player *i* is a conjunction of all elements of a set F_i of maxims for *i*, such that F_i satisfies the following conditions: (i) (*Distinct Antecedents*) for all x_{-i}: F_i contains at most one maxim whose antecedent is logically equivalent to x_{-i}; and (ii) (*Deductive Closure*) for all x_{-i}', for all non-null y_i': if the material implication $x_{-i}' \Rightarrow y_i'$ is logically entailed by a conjunction of all elements of F_i, then F_i contains a maxim $x_{-i}'' \Rightarrow y_i''$ such that x_{-i}'' is logically equivalent to x_{-i}' and y_i'' logically entails y_i'. By virtue of Distinct Antecedents, a decision rule for *i* makes a set of recommendations to her that are conditional on logically distinct predictions about the other players. In view of this, the Deductive Closure condition implies that, for any collective prediction, all the permissibility sentences implied by the rule, given that prediction, are summarised by a single maxim of the rule. As the consequent of that maxim is a recommendation, this condition guarantees that the set F_i is consistent, and that F_i does not logically entail the falsity of any collective prediction. In this sense, a decision rule for player *i* is compatible with every possible collective prediction about the other players. However, it need not contain maxims covering all these possibilities. We use D_i as a sentence variable standing in for a decision rule for player *i*.

7. Common practical reasoning in a game

We now use the concepts of an interactive reasoning system and of a decision rule, developed in Sections 5 and 6 respectively, to model CKR in a given game. To do so, we first specify P_0, the set of primitive sentences, so that it contains # and, for each $i \in N$ and for each $s_i \in S_i$, the sentences $m_i(s_i)$ and $p_i(s_i)$ (and no other sentences). This specification implies that all decision rules are in $\varphi(P_0)$. Next, we specify a particular profile of decision rules $D = (D_1, ...,$

D_n). We then construct reasoning schemes $R^* = \langle\varphi(P_0), A(R^*), I(R^*)\rangle$, $R_1 = \langle\varphi(P_0), A(R_1), I(R_1)\rangle, \dots, R_n = \langle\varphi(P_0), A(R_n), I(R_n)\rangle$ in the following way. R^* is constructed by using the rules:

(1) $A(R^*) = \{\#, D_1, ..., D_n\}$;

(2) $I(R^*)$ contains the rules of valid inference and those specified below, and no other rules:

(i) for all $p \in \varphi(P_0)$: «from $\{p\}$, infer $R_i(p)$» $\in I(R^*)$;

(ii) for all $i \in N$, for all $y_i, z_i \in \varphi(P_0)$ such that y_i is a recommendation to i and z_i is the prediction about i that is the correlate of y_i: «from $\{R_i(y_i)\}$, infer z_i» $\in I(R^*)$.

For each $i \in N$, R_i is constructed by using the rules:

(3) $A(R_i) = \{p \in \varphi(P_0) \mid p = R^*(q) \text{ for some } q \in T(R^*)\}$;

(4) $I(R_i)$ contains the rules of valid inference and those specified below, and no other rules:

for all $p \in \varphi(P_0)$: «from $\{R^*(p)\}$, infer p» $\in I(R_i)$.

By virtue of rules (2i), (3) and (4), which respectively ensure that the Attribution, Awareness and Authority requirements are satisfied, $\langle P_0, R^*, (R_1, \dots, R_n)\rangle$ is an interactive reasoning system.[14] Rule (1) provides R^* with substantive axioms, in the form of the decision rules in D. Rule (2ii) provides R^* with an inference rule that is specific to our modelling of game-theoretic rationality. This inference rule embeds in common reason the following principle: from i's having reason to believe some recommendation that applies to him, it can be inferred that he will act on that recommendation. In this sense, common reason attributes practical rationality to each player.

An interactive reasoning system $\langle P_0, R^*, (R_1, \dots, R_n)\rangle$ defined in relation to a profile D of decision rules and constructed according to rules (1) to (4), with P_0 as specified at the start of this Section, is a *common-reasoning model* of the game; D is its *common standard of practical rationality*.

It is immediate that, for any profile D of decision rules for any game in G, a corresponding (and unique) common-reasoning model exists: the model is constructed by

[14] Note that rules (3) and (4) imply that, for all players i and j, $A(R_i) = A(R_j)$ and $I(R_i) = I(R_j)$. Though this feature could be relaxed in more complex models, it is helpful to impose it here. Notwithstanding this, it is still important for our Lewisian perspective that R_i and R_j are distinct objects, for reasons made clear by Theorem 3.

defining P_0 as specified and following rules (1) to (4). What is not so obvious (since rules (1) to (4) attribute substantive axioms, as well as some inference rules besides those of logically valid inference, to the component reasoning schemes) is whether the model so constructed is consistent. The following theorem establishes this property:

> *Theorem 4*: For every game in G, for every profile D of decision rules for that game, the common-reasoning model in which D is the common standard of practical rationality is consistent.

Theorem 4 shows that our framework can represent coherently common knowledge of *any* conception of practical rationality that can be formulated as a profile of decision rules. Together with Theorem 3, it establishes the credentials of our Lewisian modelling approach.

8. The recommendation algorithm

We now focus on the content of common reason in the common-reasoning model defined by a given profile D of decision rules, in so far as that content relates to permissibility and impermissibility of strategies. (From rules (3) and (4) and Theorem 4, each R_i will have the same content as common reason in relation to permissibility and impermissibility of strategies.)

For each player i and each strategy s_i, we can ask whether, in the common-reasoning model, it is a theorem of R^* that s_i is permissible for i (i.e. whether $p_i(s_i) \in T(R^*)$ is true). We can also ask whether it is such a theorem that s_i is impermissible for i (i.e. whether $\neg p_i(s_i) \in T(R^*)$ is true). By virtue of Theorem 4, it cannot be the case that $p_i(s_i) \in T(R^*)$ and $[\neg p_i(s_i)] \in T(R^*)$ are both true. But it *can* be the case that neither is true – that is, that common reason is silent about whether s_i is permissible or impermissible. Thus, in general, a common-reasoning model implies a trinary partition of each player's strategy set S_i, the three elements of which are $\{s_i \in S_i | \, p_i(s_i) \in T(R^*)\}$, $\{s_i \in S_i | \, [\neg p_i(s_i)] \in T(R^*)\}$, and $\{s_i \in S_i | \, p_i(s_i) \notin T(R^*)$ and $[\neg p_i(s_i)] \notin T(R^*)\}$. We call this partition the *common-reasoning partition* for player i.

These arguments indicate that the common-reasoning model, for a given profile of decision rules, defines a 'solution' of the game that is interpretable as indicating which strategies are shown by common reason to be permissible (resp. impermissible).[15] In general,

[15] A corresponding argument can be made about possibility and impossibility, leading to the conclusion that each S_i can be partitioned into $\{s_i \in S_i | \, m_i(s_i) \in T(R^*)\}$, $\{s_i \in S_i | \, [\neg m_i(s_i)] \in T(R^*)\}$, and $\{s_i \in S_i | \, m_i(s_i) \notin T(R^*)$ and $[\neg m_i(s_i)] \notin T(R^*)\}$. It is an implication of our proofs in the appendices that, for each player i, this partition coincides with the common-reasoning partition.

a solution may have the property that some s_i is neither shown by common reason to be permissible nor shown by common reason to be impermissible. The reader might want to ask what, in such a case, it is rational for the relevant player i to do. However, the Lewisian approach does not analyse what it is 'really' rational for a player to do, but rather what players have reason to believe, given that they endorse certain modes of reasoning. Within this approach, the closest analogue of the reader's question is 'Which permissibility sentences does i have reason to believe to be true?' The answer (in view of rules (3) and (4) for the construction of R_i in the common-reasoning model) is: 'The set of permissibility sentences that are theorems of R^*'. All that can be said (for the relevant case) is that, using only the reasoning resources that have been attributed to her by the model, i cannot determine whether s_i is permissible or impermissible.

Although Theorem 4 establishes the existence of a solution, it does not in itself show how we, as analysts, can discover that solution; nor does it indicate a specific line of reasoning whereby common reason can reach the conclusions about permissibility (resp. impermissibility) of strategies that are summarised by the profile of common-reasoning partitions. Each of these gaps can be filled by defining a particular algorithm, as we now explain.

For any profile $D = (D_1, ..., D_n)$ of decision rules, we define the *recommendation algorithm* as follows. The algorithm has a succession of *stages* $k = 0, 1, 2, ...$, at each of which, for each player i, it generates as its *output* a recommendation to i, denoted y_i^k. As an initiation rule, we set $y_i^0 = \#$, for each i. Then, for each stage $k > 0$, and for each player i, y_i^k is obtained through three *operations*. *Operation* 1 generates, for each i, a prediction about i, denoted z_i^k, that is defined as the correlate of y_i^{k-1}. *Operation* 2 generates, for each i, a collective prediction about $N\backslash\{i\}$, denoted x_{-i}^k, that is defined as $z_1^k \wedge ... \wedge z_{i-1}^k \wedge z_{i+1}^k \wedge ... \wedge z_n^k$. *Operation* 3 determines y_i^k, for each i, as follows: if there is a component maxim of D_i that has as its antecedent a sentence logically equivalent to x_{-i}^k, then y_i^k is the consequent of that maxim; otherwise, $y_i^k = \#$. The algorithm *halts* if a stage k^* is reached at which $y_i^{k^*} = y_i^{k^*-1}$, for all i. If such a k^* is reached, then, for each player i, $y_i^{k^*}$ is the *final output* of the algorithm. We can now state:

Theorem 5: Consider any game in G and any profile D of decision rules for the game.

(i) The recommendation algorithm for D halts at some finite stage $k^* > 0$.

(ii) For each player i, let y_i^{k*} be the final output of the recommendation algorithm for D, and $R*$ be common reason in the common-reasoning model with D as common standard of practical rationality. For each $i \in N$, and for each $s_i \in S_i$:

(a) s_i is stated by y_i^{k*} to be permissible if, and only if, $p_i(s_i) \in T(R^*)$; and

(b) s_i is stated by y_i^{k*} to be impermissible if, and only if, $\neg p_i(s_i) \in T(R^*)$.

This theorem establishes that, for any profile D of decision rules, the corresponding recommendation algorithm halts and generates, as its final output for each player i, a recommendation for i that conjoins exactly those permissibility sentences for i that are theorems of $R*$ in the common-reasoning model with D as common standard of practical rationality. Thus, the algorithm is a tool by which we, as analysts, can discover the common-reasoning partition for each player i.

The recommendation algorithm can also be interpreted as tracking a line of reasoning by which common reason can establish the conclusions captured by the players' common-reasoning partitions. To see this, consider the algorithm for any given profile D of decision rules. With respect to each player i, the output of stage 0 (i.e. the null recommendation) is an axiom of, and therefore a theorem of, R^* in the corresponding common-reasoning model, by virtue of rule (1) for construction of that model. The output of each subsequent stage can be derived by using, in a specific sequence, axioms and inference rules of R^* together with the output of the previous stage. At each stage, Operation 1 tracks inferences from the output of the previous stage that are licensed by rules in $I(R^*)$ by virtue of rules (2i) and (2ii) for the construction of a common-reasoning model; Operation 2 tracks inferences from the conclusions of those tracked by Operation 1 that are licensed by rules of logically valid inference in $I(R^*)$, by virtue of rule (2); and, finally, Operation 3 tracks inferences licensed by rules of logically valid inference provided by rule (2), using as premises D, which is an axiom of R^* by virtue of rule 1, and the conclusions of inferences tracked by Operation 2.

9. Categorisations

In Section 10 below, we compare our Lewisian common-reasoning approach with the Bayesian approach described in Sections 2 and 3. But first we define some concepts introduced by Cubitt and Sugden (2011) that are useful in making the comparison, because they provide a convenient way to summarise binary and trinary partitions of sets of strategies.

For any player i, an ordered pair $<S_i^+, S_i^->$ of subsets of S_i is a *categorisation* of S_i if it satisfies the following conditions: (i) S_i^+ and S_i^- are disjoint; (ii) $S_i^- \subset S_i$; and (iii) if $S_i \backslash S_i^- = \{s_i\}$ for any $s_i \in S_i$, then $S_i^+ = \{s_i\}$. In general, a categorisation of S_i defines a trinary partition of S_i, whose elements are the *positive component* S_i^+, the *negative component* S_i^-, and the *residual set* $S_i \backslash (S_i^+ \cup S_i^-)$.

Now consider any non-empty set $N' \subseteq N$ of players. For each $i \in N'$, let $<S_i^+, S_i^->$ be any categorisation of S_i. In order to allow us to aggregate across players, we define a 'union' relation $\cup^*$ between such categorisations such that $\cup^*_{i \in N'} <S_i^+, S_i^-> \equiv <\cup_{i \in N'} S_i^+, \cup_{i \in N'} S_i^->$.[16] Each such $\cup^*_{i \in N'} <S_i^+, S_i^->$ is a *categorisation* of $\cup_{i \in N'} S_i$; its *positive component* is $\cup_{i \in N'} S_i^+$; and its *negative component* is $\cup_{i \in N'} S_i^-$. For purposes of the main text, we need only the case where $N' = N$. For this case, we use a shorthand notation in which $\mathbb{S}$ denotes $\cup_{i \in N} S_i$ and $\mathbb{S}^+$ and $\mathbb{S}^-$ denote, respectively, the positive and negative components of a typical categorisation of $\mathbb{S}$. Such a categorisation is *exhaustive* if $\mathbb{S}^+ \cup \mathbb{S}^- = \mathbb{S}$.

Consider any two categorisations $C' = <\mathbb{S}^{+\prime}, \mathbb{S}^{-\prime}>$ and $C'' = <\mathbb{S}^{+\prime\prime}, \mathbb{S}^{-\prime\prime}>$ of $\mathbb{S}$. We define a binary relation $\sqsupseteq^*$ (read as *has weakly more content than*) between such categorisations such that $C'' \sqsupseteq^* C'$ if and only if $\mathbb{S}^{+\prime\prime} \supseteq \mathbb{S}^{+\prime}$ and $\mathbb{S}^{-\prime\prime} \supseteq \mathbb{S}^{-\prime}$. If, in addition, either $\mathbb{S}^{+\prime\prime} \supset \mathbb{S}^{+\prime}$ or $\mathbb{S}^{-\prime\prime} \supset \mathbb{S}^{-\prime}$ holds, we will say that C'' *has strictly more content than* C', denoted $C'' \sqsupset^* C'$.

Consider any Bayesian model M of the game, as defined in Section 2. For each player i, the model specifies a set $S_i^*(M) \subseteq S_i$ of included strategies. Equivalently, for each i, M specifies a categorisation $C_i^M = <S_i^+(M), S_i^-(M)>$ of S_i, where $S_i^+(M) = S_i^*(M)$ and $S_i^-(M) = S_i \backslash S_i^*(M)$.[17] Thus, aggregating across all players, M specifies a single categorisation $C^M = \cup^*_{i \in N} <S_i^+(M), S_i^-(M)>$ of $\mathbb{S}$. We will say that C^M is the *inclusion categorisation* with respect to Bayesian model M. By construction, C^M is exhaustive.

Now consider the common-reasoning model of the game, for a given profile D of decision rules, as defined in Section 7. As Section 8 showed, for any player i, this model defines a trinary common-reasoning partition of S_i, two of whose elements are $\{s_i \in S_i | p_i(s_i) \in T(R^*)\}$ and $\{s_i \in S_i | [\neg p_i(s_i)] \in T(R^*)\}$, where R^* is common reason. Thus, the model defines, for each i, a categorisation $<S_i^+(D), S_i^-(D)>$ of S_i, where $S_i^+(D) = \{s_i \in S_i | p_i(s_i) \in T(R^*)\}$ and $S_i^-(D) = \{s_i \in S_i | [\neg p_i(s_i)] \in T(R^*)\}$.[18] Again aggregating across players, the

[16] Recall that we have imposed that that, for all $i, j, \in N$, $S_i \cap S_j = \varnothing$.

[17] Given our definitions here and in Section 2, non-emptiness of Ω ensures that C_i^M satisfies the definition of a categorisation.

[18] Since the common-reasoning partition for player i is a partition of S_i, condition (i) of the definition of a categorisation is satisfied. That conditions (ii) and (iii) are satisfied too follows from the facts that decision rules

common-reasoning model for the profile D specifies a single categorisation $C^D = \cup*_{i \in N} < S_i^+(D), S_i^-(D) >$ of $\mathbb{S}$. Unlike C^M, C^D may or may not be exhaustive, depending on the common-reasoning partitions resulting from profile D. As the positive (resp. negative) component of C^D is the set of strategies whose permissibility (resp. impermissibility) is established in common reason in the common-reasoning model, we will say that C^D is the *common-reasoning solution* of the game, with respect to the profile D of decision rules.

10. ICEU Bayesian models revisited

In this Section, we compare our approach to modelling CKR, set out in Sections 4–8, to the canonical approach set out in Sections 2 and 3. We focus on the cases in which each approach is adapted to a conception of practical rationality provided by the ICEU standard.

We have already defined the concept of a Bayesian model which incorporates the ICEU standard – the ICEU Bayesian model. To explore the relationship between the two approaches, we need to specify a corresponding class of common-reasoning models in which the conception of practical rationality is ICEU. As explained in Section 7, the common-reasoning model is uniquely defined for any given game and any given profile of decision rules. So what we need to do is to define a profile of ICEU decision rules, for any given game. We do this in the following way.

For any player i and for any collective prediction x_{-i} about $N \backslash \{i\}$, we define a probability distribution over S_{-i} as *ICEU-consistent* with x_{-i} if it satisfies the following conditions. First, probabilities are independent in that, for each $s_{-i} \in S_{-i}$, the probability of s_{-i} is the product of the marginal probabilities of the strategies appearing in s_{-i}. Second, every strategy that x_{-i} states to be impossible has zero marginal probability. Third, every strategy that x_{-i} states to be possible has strictly positive marginal probability. An *ICEU maxim* for i is a maxim $x_{-i} \Rightarrow y_i$ such that (i) y_i states $p_i(s_\mathrm{i})$ if, and only if, s_i maximises i's expected utility relative to *all* probability distributions that are ICEU-consistent with x_{-i}, and (ii) y_i states $\neg p_i(s_\mathrm{i})$ if, and only if, s_i does *not* maximise i's expected utility relative to *any* probability distribution that is ICEU-consistent with x_{-i}. Because the elements of a given set of sentences can be conjoined in different orders, there may be collective predictions (resp. recommendations) that are formally distinct from, but logically equivalent to x_{-i} (resp. y_i), and

are defined in terms of predictions and recommendations, and conditions analogous with (ii) and (iii) are embedded in the definitions of 'prediction' and 'recommendation'.

so there may be more than one ICEU maxim with the logical content of $x_{-i} \Rightarrow y_i$. By taking exactly one maxim from every set of logically equivalent ICEU maxims for *i*, we can construct a *non-redundant* set F_i of ICEU maxims for each player *i*.

Consider any two ICEU maxims $x_{-i}' \Rightarrow y_i'$ and $x_{-i}'' \Rightarrow y_i''$. It follows from the definition of an ICEU maxim that if x_{-i}' and x_{-i}'' are logically equivalent then so too are y_i' and y_i''. Hence, given the definition of non-redundancy, F_i satisfies Distinct Antecedents. It also follows from the definition of an ICEU maxim that if x_{-i}' logically entails x_{-i}'', then y_i' logically entails y_i''.[19] Thus, F_i satisfies Deductive Closure. So any conjunction of the elements of a non-redundant set F_i of ICEU maxims for *i* is a decision rule for that player. Since all such conjunctions are logically equivalent, we can fix on any one of them as 'the' ICEU decision rule D_i for player *i*; and, in this way, construct 'the' profile *D* of ICEU decision rules for the game and 'the' *ICEU common-reasoning model*. This model implies a unique common-reasoning solution, which we may unambiguously take as the *ICEU common-reasoning solution* (as every profile of ICEU decision rules for the game yields the same such solution). We denote this solution by the categorisation C^*.

Assume for the moment that the game has at least one ICEU Bayesian model. Consider any such model *M* and let C^M be its inclusion categorisation; necessarily, C^M is exhaustive. We now investigate the relationship between C^M and C^*.

M can be interpreted as a model of a situation in which each player is rational in the ICEU sense, and in which it is common knowledge that each strategy in the positive component of C^M might be played, and that each strategy in the negative component will not be played. The Bayesian modelling strategy does not try to represent how players arrive at this common knowledge; but on the most natural interpretation, there is an implicit assumption that they do so by some process of reasoning whose premises include (but are not necessarily limited to) those of the ICEU standard of rationality (see Section 1). In contrast, the ICEU common-reasoning framework explicitly models what players have reason to believe about the game, given that the ICEU standard of rationality is axiomatic in common reason. Thus, given our interpretation of Bayesian models, it is natural to conjecture that if the permissibility (resp. impermissibility) of some strategy is a theorem of common reason, then every ICEU Bayesian model includes (resp. does not include) that strategy. It is also

[19] The reason is that, as collective predictions become (strictly) stronger, the restrictions on probabilities required by ICEU-consistency with such predictions tighten, so making it 'easier' for a strategy to be expected utility

natural to conjecture that if the ICEU common-reasoning solution is exhaustive, then there exists an ICEU Bayesian model that includes (resp. does not include) every strategy whose permissibility (resp. impermissibility) is a theorem of common reason. The following theorems establish that these conjectures are indeed correct.

> *Theorem 6*: Consider any game in G for which an ICEU Bayesian model exists. Consider any such model M of the game, and let its inclusion categorisation be C^M. Let C* be the ICEU common-reasoning solution. Then $C^M \supseteq^* C^*$.

> *Theorem 7*: For every game in G: If the ICEU common-reasoning solution C* is exhaustive, then (i) there exists an ICEU Bayesian model of the game; and (ii) for every such model M, the inclusion categorisation C^M is identical to C*.

Theorem 7 establishes that, in the special case in which the ICEU common-reasoning solution of the game is exhaustive, and with respect to the resulting categorisations, the Bayesian and common-reasoning approaches are equivalent. In this case, the ICEU common-reasoning model may be seen as *justifying* ICEU Bayesian models, in the sense that it describes explicit steps of reasoning whereby the players could establish the permissibility of those strategies included in each ICEU Bayesian model and the impermissibility of all other strategies.

But, now consider cases where the ICEU common-reasoning solution is *not* exhaustive. (As we will show in Section 11, this is a genuine possibility.) Nothing we have said in this Section excludes the possibility that such a game has an ICEU Bayesian model. However, we know from Theorem 6 that, for every such model M, $C^M \supset^* C^*$. Thus, the unmodelled reasoning that players are implicitly assumed to use in any such M must be (assumed to) enable them to arrive at common knowledge, not captured in C*, about the possibility and impossibility of strategies. That reasoning would have to make use of axioms and inference rules in addition to (or stronger than) those of the ICEU common-reasoning model.[20] But if an ICEU Bayesian model requires support of this kind, there seems no a priori

maximising for all probability distributions satisfying the restrictions and also 'easier' for it to be expected utility maximising for no such probability distribution.

[20] It would be compatible with our general Lewisian approach to analyse interactive reasoning systems in which common reason has additional axioms and inference rules. Indeed, Lewis's analysis of conventions attributes principles of salience and inductive inference to common reason, in the context of a game played recurrently in a population (Lewis, 1969; Cubitt and Sugden, 2003). In the case of one-shot play, game theorists often appeal to additional resources of reasoning when rationalising solution concepts, especially in the presence of multiple equilibria. For example, there is a tradition of postulating that players have access to, and take as authoritative, a book containing suggestions made by the game theorist; and that these may go beyond (but may not contradict) the implications of CKR.

reason to expect such a model to exist for all games. And, even if *some* such model does exist for a given game, the additional axioms and inference rules that it requires may not be appropriate for all contexts in which the game is played.

11. Resolving the paradoxes

We now turn to the paradoxes presented in Section 3. It is convenient to go straight to the most problematic case – the Tom, Dick and Harry Paradox of Game 3. The paradox is that this game has no ICEU Bayesian model.

Nevertheless, this game (like every other) has an ICEU common-reasoning solution, which can be identified by using the recommendation algorithm. The outputs of the algorithm are (for $i = 1, 2, 3$): $y_i^0 = \#$; $y_i^1 = p_i(in_i)$; $y_i^2 = y_i^1$. The ICEU common-reasoning solution is $<\{in_1, in_2, in_3\}, \varnothing>$. This categorisation is non-exhaustive: for each player i, out_i is neither shown to be permissible nor shown to be impermissible.

It should not be surprising if, for *some* reasoning scheme R and *some* sentence p, neither p nor its negation is a theorem of R. The reasoning scheme R has, as resources, only axioms in $A(R)$ and inference rules in $I(R)$; thus, it is hardly surprising if we can find some p such that neither it nor its negation can be derived from those axioms using those inference rules. The significance of the previous paragraph is that it shows that this unsurprising general possibility applies *specifically* to the case where R is common reason in the ICEU common-reasoning model of Game 3 and p any permissibility sentence referring to the strategy out_i for any player i.

As the ICEU common-reasoning solution of Game 3 is not exhaustive, it is natural to ask whether there is any way in which the profile of decision rules in the ICEU common-reasoning model could be strengthened to yield an exhaustive categorisation as the resulting solution. The answer is ‘No’: for each player in Game 3, there is no decision rule stronger than that player’s ICEU decision rule.[21] More generally, one might ask whether it is possible

[21] A proof can be sketched as follows. For each player i of Game 3, let $j(i)$ be the player other than i on whose strategy i’s payoff depends. The set of maxims of an ICEU decision rule for any player i of Game 3 can be partitioned into those whose antecedents refer to both of $j(i)$’s strategies (*type* 1) and the remainder (*type* 2). Let a recommendation to i that refers to both of i’s strategies be *maximally specific*. Every type 1 maxim has, as its consequent, a recommendation to i that is maximally specific. Thus, no consequent of a type 1 maxim can be strengthened, as the definition of a maxim forces consistency on its consequent. One can strengthen the consequent of some type 2 maxims; but, by Deductive Closure, each such strengthening requires an extra permissibility sentence to be conjoined to the consequent of some type 1 maxim, in all cases inducing inconsistency in the consequent of at least one such maxim.

to achieve an exhaustive categorisation by augmenting the interactive reasoning system that constitutes the ICEU common-reasoning model with further axioms or inference rules. Obviously, Theorem 4 cannot vouch for the consistency of such a construction. But we can show that in any such interactive reasoning system that *is* consistent, neither the permissibility nor the impermissibility of any out_i is a theorem of common reason.[22]

If our modelling framework is accepted, the implication is this: given common knowledge of the ICEU standard of rationality, there is no way in which players could reason their way to the conclusion that the permissibility or impermissibility of out_i, for any player i, was common knowledge. However, an ICEU Bayesian model of Game 3, were it to exist, would rest on the implicit assumption that the players *had* reasoned their way to such a conclusion. Given this analysis, the non-existence of an ICEU Bayesian model is not surprising.

This analysis also provides the key to the other paradoxes. Consider the Proving Too Much Paradox of Game 1. For this game, as for all two-player games, a family of ICEU Bayesian models exists. The paradox, as we originally presented it, is that certain propositions about the choiceworthiness of strategies, namely those summarised by $S_1^* = \{first\}$ and $S_2^* = \{left, right\}$, hold for all ICEU Bayesian models of Game 1, but there seems no way that the players could arrive at common knowledge of those propositions using only reasoning resources attributed to them by the assumption that the ICEU standard of rationality is common knowledge. Now that we have developed the concept of a common-reasoning model, we can tighten up the 'there seems no way ...' clause. The ICEU common-reasoning solution tells us what conclusions about choiceworthiness (or permissibility) can be arrived at by players who have just those reasoning resources and how they can arrive at them.

In the case of Game 1, with the ICEU common standard of practical rationality, the outputs of the recommendation algorithm are: $y_1^0 = y_2^0 = \#$; $y_1^1 = \neg p_1(third)$, $y_2^1 = \#$; $y_1^2 = \neg p_1(third)$, $y_2^2 = p_2(left)$; $y_1^3 = y_1^2$, $y_2^3 = y_2^2$. The ICEU common-reasoning solution is $<\{left\}, \{third\}>$. As with Game 1, this solution is not exhaustive: *first*, *second* and *right* are neither shown to be permissible nor shown to be impermissible. So the intuition on which the initial

[22] For any profile D of decision rules, we define a *generalised common-reasoning model* as an interactive reasoning system $< P_0, R^*, (R_1, \ldots, R_n)>$ which satisfies amended versions of the conditions used to define the common-reasoning model. Specifically, '$A(R^*) = \ldots$' and '$A(R_i) = \ldots$' in conditions (1) and (3) respectively are replaced by '$A(R^*) \supseteq \ldots$' and '$A(R_i) \supseteq \ldots$' so as to allow additional axioms, and the clauses 'and no other rules' in conditions (2) and (4) are deleted, so as to allow additional inference rules. It can be shown that, in any generalised common-reasoning model of Game 3, if D is the profile of ICEU decision rules, $p_1(out_1) \in T(R^*) \Rightarrow [\neg p_3(out_3)] \in T(R^*) \Rightarrow p_2(out_2) \in T(R^*) \Rightarrow [\neg p_1(out_1)] \in T(R^*)$, and $[\neg p_1(out_1)] \in T(R^*) \Rightarrow p_3(out_3) \in T(R^*) \Rightarrow$

statement of the paradox is based is correct. Any ICEU Bayesian model of Game 1 must therefore rest on the implicit assumption that players have access to resources of reasoning beyond those arising from the ICEU standard of practical rationality being common knowledge among them, and that those resources enable them to arrive at common knowledge of an exhaustive categorisation. We are not entitled to make this assumption merely by virtue of the existence of *some* Bayesian model of Game 1. Thus, the fact that *first* and *right* are choiceworthy in all Bayesian models of that game does not imply that the choiceworthiness of those strategies is an implication of common knowledge of ICEU rationality *per se*. Rather, it is an implication of the Bayesian representation of that common knowledge. That representation is premised on an implicit assumption to which the ICEU common-reasoning model lends no support.

The Three-Lane Road Paradox of Game 2 can be resolved in a similar way. Proposition 2 establishes that ICEU Bayesian models of Game 2 can be partitioned into two classes – those models M' for which the inclusion categorisation is $C^{M'} = <\{in_1, in_2, out_2\}, \{out_1\}>$, and those models M'' for which the inclusion categorisation is $C^{M''} = <\{in_1, out_1, in_2\}, \{out_2\}>$. The paradox is that each of these models appears to represent a situation in which the players have common knowledge of an asymmetry between what rationality requires of one of them and what it requires of the other, even though the formal structure of the game is symmetric.

Again, we can use an ICEU common-reasoning model to discover what conclusions players can reach by reasoning that uses only those axioms and inference rules that represent common knowledge of the ICEU standard. For Game 2, the outputs of the corresponding recommendation algorithm are (for $i = 1, 2$): $y_i^0 = \#$; $y_i^1 = p_i(in_i)$; $y_i^2 = y_i^1$. The ICEU common-reasoning solution is $<\{in_1, in_2\}, \varnothing>$; again, it is not exhaustive. There is no asymmetry between the players in this solution, but, for each player i, out_i is neither shown to be permissible nor shown to be impermissible. As in the other games, we are not generally entitled to assume that the players have access to additional (or stronger) axioms and inference rules that allow them to arrive at an exhaustive categorisation of strategies by common reasoning, because that assumption can only hold *if* the additional axioms and

$[\neg p_2(out_2)] \in T(R^*) \Rightarrow p_1(out_1) \in T(R^*)$, where we use $\Rightarrow$ now as material implication in our own analysis. Thus, neither $p_1(out_1)$ nor $\neg p_1(out_1)$ can be a theorem of a consistent R^*.

inference rules induce an asymmetry between the players.[23] If there is no source of asymmetry, the Bayesian modelling strategy must fail.

12. Conclusion

Our objective has been to investigate the relationship between two ways of understanding one of the core concepts of game theory – common knowledge of rationality (CKR). On one understanding, canonically expressed by Aumann (1987), CKR is represented by a model of what players know about one another's strategy choices, conditional on every possible event. Each player's choices are required to be decision-theoretically rational with respect to her knowledge and beliefs, and the event represented by the model itself is required to be common knowledge. The strategies available to each player can then be partitioned into those that are and are not played in the model, and since the model is common knowledge, those binary partitions are common knowledge too. On the alternative understanding, inspired by Lewis (1969) and developed here with our concept of a common-reasoning model, CKR is represented by the axioms and inference rules of a mode of 'common reasoning' that each player endorses and attributes to the others. This approach induces a trinary partition of each player's strategy set, since a given strategy can have one of three statuses in common reason – it can be shown to be permissible, shown to be impermissible, or neither.

By modelling reasoning, the Lewisian approach provides formal resources with which to assess the implicit assumption of the Bayesian model, formalised in Section 2 and based on Aumann's canonical contribution, that, if CKR holds, then players can arrive at a binary partition of strategies into those that are rationally playable and those that are not. We have shown that, for a given standard of practical rationality, our common-reasoning model grounds this implicit assumption when, *but only when*, the resulting common-reasoning solution is exhaustive. Since this condition may fail, there can be conceptions of practical rationality common knowledge of which cannot be represented in a Bayesian model without creating the potential for puzzling or contradictory results such as those of Section 3. In contrast, we have shown that the Lewisian approach has no need to restrict the conception of practical rationality beyond a requirement of internal coherence.

[23] Using a similar argument to that deployed for Game 3, it can be shown that there are no decision rules for Game 2 that are stronger than the ICEU decision rules. However, additional axioms or inference rules might be included in a generalised common-reasoning model, and such a model might be consistent, unlike the case considered in footnote 22.

To repeat what we said at the outset, we recognise that the paradoxes of Section 3 can be circumvented by various (non-Lewisian) adaptations of the Bayesian framework. The central contribution of this paper does not depend on our analysis of what we see as paradoxical implications of Aumann's approach, nor on whether we are right to see those implications as paradoxes, rather than as minor curiosities or as evidence against the particular conception of practical rationality that gives rise to them. Our core contribution is the provision of formal and general foundations for Lewisian solution concepts in game theory.

We have proposed, and derived from a model of reasoning, a very general solution concept – that of a 'common-reasoning solution' – which defines a trinary partition of strategies for *any* given normal-form game and *any* given coherent concept of practical rationality for that game. This proposal defines a path of possible research in which alternative conceptions of practical rationality are specified, properties of the resulting common-reasoning solutions analysed, and their relationships to existing concepts studied. Cubitt and Sugden (2011) can be seen as taking initial steps along this path, in view of the relationship (described in Appendix 1) between their 'categorisation solutions' and common-reasoning solutions. We expect that solution concepts that are generated in this way will often be quite distinct from those of standard game theory. The reason for that expectation is that few existing solution concepts are defined in terms of trinary partitions of strategies, but such partitions reflect the underlying structure of common-reasoning models. More basically, they stem from a fundamental feature of reasoning itself.

Appendix 1: Categorisation procedures and recommendation algorithms

Cubitt and Sugden (2011) (henceforth CS11) defines a class of 'categorisation procedures', which operate on the strategies of a given game. In this appendix, we demonstrate a relationship between CS11's concept and that of a recommendation algorithm introduced in Section 8. This allows us to use a result from CS11, together with new results presented here, as ingredients for our main proofs presented in Appendix 2.[24] As we have demonstrated in Section 8 that the recommendation algorithm, for a given profile of decision rules, tracks steps of common reasoning in the corresponding common-reasoning model, the formal results of this appendix also substantiate CS11's informal claim that categorisation procedures may be interpreted as tracking reasoning that players can undertake.

Our analysis applies to any given game in G. We begin by extending the concepts introduced in Section 9, in a way that follows CS11. Section 9 defined the concepts of a categorisation of S_i, for any player i, and of a categorisation of $\cup_{i \in N'} S_i$, for any non-empty set $N' \subseteq N$. We now require the case where $N' = N \backslash \{i\}$, for any player i, as well as that (already introduced) where $N' = N$. We use $\mathbb{S}_{-i}$ as a shorthand for $\cup_{i \in N \backslash \{i\}} S_i$; the positive and negative components of a categorisation of the latter set will typically be denoted $\mathbb{S}_{-i}^{+}$ and $\mathbb{S}_{-i}^{-}$.

We denote the set of categorisations of S_i, the set of categorisations of $\mathbb{S}_{-i}$ and the set of categorisations of $\mathbb{S}$ by, respectively, $\Phi(S_i)$, $\Phi(\mathbb{S}_{-i})$ and $\Phi(\mathbb{S})$. The *null categorisation* $<\varnothing, \varnothing>$ is an element of each of these sets. Where convenient, we use C_i, C_i', and so on, to denote particular categorisations in $\Phi(S_i)$; C_{-i}, C_{-i}', and so on, to denote particular categorisations in $\Phi(\mathbb{S}_{-i})$; and C, C', and so on, to denote particular categorisations in $\Phi(\mathbb{S})$. We extend to categorisations in $\Phi(S_i)$ and $\Phi(\mathbb{S}_{-i})$, in the obvious way, the definitions of the relations $\sqsupseteq^*$ ('has weakly more content than') and of $\sqsupset^*$ ('has strictly more content than'), introduced in Section 9 for categorisations in $\Phi(\mathbb{S})$. (See CS11, Section 2, for details.)

We define a *categorisation function* for player i as a function f_i: $\Phi(\mathbb{S}_{-i}) \rightarrow \Phi(S_i)$ with the following *Monotonicity* property: for all C_{-i}', $C_{-i}'' \in \Phi(\mathbb{S}_{-i})$, if $C_{-i}'' \sqsupset^* C_{-i}'$ then $f_i(C_{-i}'') \sqsupseteq^* f_i(C_{-i}')$.

The content of a given profile $f = (f_1, \ldots, f_n)$ of categorisation functions can be expressed as a single function ζ: $\Phi(\mathbb{S}) \rightarrow \Phi(\mathbb{S})$, constructed as follows. Let $C = <\mathbb{S}^{+}, \mathbb{S}^{-}>$ be

[24] In proving the results of this appendix, we do not assume the truth of any of the results stated in the main text of the current paper. This ensures that the results of this appendix can be used in the proofs in Appendix 2. In commenting on Proposition A3 (below), we do note its significance when conjoined with Theorem 5; but this interpretative passage is not involved in the proof of either result.

any categorisation of $\mathbb{S}$. For each player i, define $C_{-i} = <\mathbb{S}^{+}\backslash S_i, \mathbb{S}^{-}\backslash S_i>$. Next, define $S_i^{+\prime}$ and $S_i^{-\prime}$ as, respectively, the positive and negative components of $f_i(C_{-i})$. Finally, define $\zeta(C) = \cup^*_{i\in N} <S_i^{+\prime}, S_i^{-\prime}>$. We will say that ζ *summarises* f. A function ζ: $\Phi(\mathbb{S})\rightarrow\Phi(\mathbb{S})$ that summarises some profile f of categorisation functions is an *aggregate categorisation function*.

For any aggregate categorisation function ζ, we define the *categorisation procedure* by the following pair of instructions, which generate a sequence of categorisations $C(k) \equiv <\mathbb{S}^{+}(k), \mathbb{S}^{-}(k)>$ of $\mathbb{S}$, for successive stages $k \in \{0, 1, 2, \ldots.\}$, inductively, as follows:

(i) *Initiation rule.* Set $C(0) = <\varnothing, \varnothing>$;

(ii) *Continuation rule*. For all $k > 0$, set $C(k) = \zeta[C(k-1)]$.

The procedure *halts* at the lowest value of k' for which $C(k') = C(k'-1)$; this value of k' will be denoted by k^*. $C(k^*)$ is the *categorisation solution* of the game, relative to ζ. CS11 proves the following result (their Proposition 1):

> *Proposition A1*: Consider any game in G and let ζ be any aggregate categorisation function for the game. The categorisation procedure for ζ has the following properties:
>
> (i) For all $k \in \{1, 2, \ldots.\}$, $C(k) \supseteq^* C(k-1)$.
>
> (ii) The procedure halts, defining a unique categorisation solution relative to ζ.

We can now relate these concepts from CS11 to those introduced in Sections 6–8, by exploiting a correspondence between decision rules and categorisation functions.

Recall that a decision rule D_i, for player i, is a conjunction of all elements of a set F_i of maxims of the form $x_{-i} \Rightarrow y_i$, where x_{-i} is a collective prediction about $N\backslash\{i\}$, y_i is a recommendation to i, and F_i satisfies Distinct Antecedents and Deductive Closure. The content of any recommendation y_i can be expressed by specifying two subsets of S_i: the set S_i^{+} of strategies which are stated by y_i to be permissible for i, and the set S_i^{-} of strategies which are stated to be impermissible. The definition of a recommendation ensures that $C_i = <S_i^{+}, S_i^{-}>$ is a categorisation of S_i. We will say that C_i *encodes* y_{i}. Similarly, the content of any collective prediction x_{-i} can be encoded as a unique categorisation C_{-i} of $\mathbb{S}_{-i}$, the positive (resp. negative) component of which contains all strategies stated by x_{-i} to be possible (resp. impossible). (The null sentence #, whether viewed as a recommendation or as a collective prediction, is encoded by $<\varnothing, \varnothing>$.) Thus, each maxim in F_i is encoded by an ordered pair of the form $<C_{-i}, C_i>$. Because D_i satisfies Distinct Antecedents, no two such ordered pairs have the same C_{-i}. If there is any C_{-i} which is not the antecedent of any maxim stated by D_i, this fact can be encoded as the ordered pair $<C_{-i}, <\varnothing, \varnothing>>$. Thus, D_i is *encoded* by a set of ordered pairs $<C_{-i}, C_i>$; and, since each $C_{-i} \in \Phi(\mathbb{S}_{-i})$ appears in exactly one of these ordered pairs, D_i itself is encoded by a unique function f_i from $\Phi(\mathbb{S}_{-i})$ to $\Phi(S_i)$.

The following result establishes that, for any decision rule D_i, the function f_i which encodes D_i is a categorisation function.

> *Proposition A2*: For every game in *G*, for every player *i*, and for every decision rule D_i for *i*, the function f_i that encodes D_i satisfies Monotonicity.

Proposition A2 implies that, for any profile $D = (D_1, ..., D_n)$ of decision rules, there exists a unique profile $f = (f_1, \ldots, f_n)$ of categorisation functions and, thus, a unique aggregate categorisation function ζ, such that ζ summarises *f* and, for each player *i*, f_i encodes D_i. We will say that ζ *encodes D*.

Recall from Section 8 that any profile *D* of decision rules also defines a recommendation algorithm. This algorithm generates, for each player *i*, an output y_i^k, for each of its stages $k = 0, 1, 2, ...$, where each such output is a recommendation to *i*. Since any such recommendation is encoded by a categorisation of S_i, and such categorisations can be aggregated across players, the combined output of each stage *k* of the recommendation algorithm is *encoded* by a categorisation of $\mathbb{S}$, defined as $\cup^*_{i\in N} <S_i^+(k), S_i^-(k)>$ where, for each *i*, $<S_i^+(k), S_i^-(k)>$ encodes y_i^k. We can now state:

> *Proposition A3*: Consider any game in *G*, and any profile *D* of decision rules for its players. Let ζ be the aggregate categorisation function that encodes *D*. For each $k \in \{0, 1, 2, \ldots\}$, the categorisation generated for stage *k* of the categorisation procedure for ζ encodes the combined output of stage *k* of the recommendation algorithm for *D*.

Propositions A1 – A3 are the results from this appendix which are used as ingredients for Appendix 2.

Proposition A3, combined with earlier results, is of independent interest in relation to CS11. The analysis of Sections 7-9 above implies that, for any profile *D* of decision rules, the theorems of *R** in the resulting common-reasoning model, insofar as they relate to permissibility and impermissibility of strategies, are identified by the combined final output of the recommendation algorithm for *D* and encoded by the common-reasoning solution for *D*. Thus, Proposition A3 establishes that, if ζ is the aggregate categorisation function that encodes *D*, then the categorisation solution for ζ and the common-reasoning solution for *D* are *identical*. This demonstrates a precise sense in which a Lewisian understanding of CKR underpins CSS11's categorisation solutions.

We end this appendix with proofs of Propositions A2 and A3:

Proof of Proposition A2: By definition, D_i is a conjunction of all elements of a set F_i of maxims for i, which satisfies Distinct Antecedents and Deductive Closure. Let f_i be the function that encodes D_i and suppose that it does not satisfy Monotonicity. Then there are C_{-i}', $C_{-i}'' \in \Phi(\mathbb{S}_{-i})$ such that $C_{-i}'' \supset^* C_{-i}'$ and *not* $f_i(C_{-i}'') \supseteq^* f_i(C_{-i}')$. So F_i contains maxims $x_{-i}' \Rightarrow y_i'$ and $x_{-i}'' \Rightarrow y_i''$, such that (i) x_{-i}'' entails x_{-i}' and (ii) y_i'' does not entail y_i'. Notice that (ii) implies that y_i' is non-null. Because of (i), the conjunction of these maxims entails $x_{-i}'' \Rightarrow y_i'$. So, by Deductive Closure, F_i contains a maxim $x_{-i}^* \Rightarrow y_i^*$ where x_{-i}^* is logically equivalent to x_{-i}'' and y_i^* entails y_i'. But because of Distinct Antecedents, this requires $x_{-i}^* = x_{-i}''$ and hence $y_i^* = y_i''$. Thus y_i'' entails y_i', contradicting (ii). □

Proof of Proposition A3: Consider any profile D of decision rules for any game in G and let ζ be the aggregate categorisation function that encodes D. Let the sequences $C(0)$, $C(1)$, …. and $C'(0)$, $C'(1)$, …. be, respectively, the sequence of categorisations generated by the categorisation procedure for ζ and the sequence of categorisations that encode the combined outputs of successive stages of the recommendation algorithm for D. Consider any $k \in \{1, 2, \ldots\}$. From the continuation rule of the categorisation procedure, $C(k) = \zeta[C(k\text{–}1)]$. Now consider stage k of the recommendation algorithm. As $C'(k\text{–}1)$ encodes the combined output of stage k–1 of the recommendation algorithm, the specification of operations 1, 2 and 3 of that algorithm, together with the fact that ζ encodes D, imply that $C'(k) = \zeta[C'(k\text{–}1)]$. Thus, if $C(k\text{–}1) = C'(k\text{–}1)$, it follows that $C(k) = C'(k)$. The Proposition follows, by induction, if $C(0) = C'(0)$, a condition guaranteed by the respective initiation rules of the categorisation procedure and recommendation algorithm (combined, in the latter case, with # being encoded by $<\varnothing, \varnothing>$). □

Appendix 2: Proofs of results from main text

Proof of Theorem 1: For any game in G, let $\rho: S \rightarrow [0, 1]$ be a probability distribution over the set S of strategy profiles. ρ is a *correlated equilibrium* if, for all $i \in N$, for all functions g_i: $S_i \rightarrow S_i$, $\sum_{s \in S} \rho(s)\ (u_i[s] - u_i[\sigma_i(s, g_i[s_i])]) \geq 0$. From Nash's existence result for finite games (Nash, 1951, Theorem 1) and the fact that any Nash equilibrium corresponds to a correlated equilibrium, existence of a correlated equilibrium is guaranteed for every game in G. Consider any such game and take any correlated equilibrium ρ^* of the game. We can construct a Bayesian model of the game as follows: Define $S^* = \{s \in S \mid \rho^*(s) > 0\}$ and Ω so that there is a one-one mapping from S^* onto Ω. For each $s \in S^*$, let $\omega(s)$ denote the corresponding element of Ω. Define the behaviour function $b(.)$ so that $b(\omega[s]) = s$. Define the information structure $\mathscr{S}$ such that, for each player i, for each strategy $s_i \in S_i^*$: $E(s_i) \in \mathscr{S}_i$. Define a function $\pi^*: \Omega \rightarrow [0, 1]$ such that, for each $s \in S^*$, $\pi^*(\omega(s)) = \rho^*(s)$. Note that this implies that $\sum_{\omega \in \Omega} \pi^*(\omega) = 1$; and that $\pi^*(\omega) > 0$ for all $\omega \in \Omega$ (so that, in addition, for each player i and each $E \in \mathscr{S}_i$, $\sum_{\omega \in E} \pi^*(\omega) > 0$). In view of this, define the profile π of priors such that, for each player i: $\pi_i = \pi^*$. Define the profile χ of choiceworthiness functions such that, for each player i, at each state ω, $\chi_i(\omega)$ is the set of strategies that are SEU-rational at ω with respect to $\mathscr{S}_i$ and π_i. By construction, the Bayesian model $<\Omega, b(.), \mathscr{S}, \pi, \chi>$ satisfies SEU-Maximization and Knowledge of Own Choice. Since ρ^* is a correlated equilibrium, it follows that, for each player i, for each state $\omega \in \Omega$: $b_i(\omega)$ is SEU-rational at ω. Hence, $b_i(\omega) \in \chi_i(\omega)$, which entails that Choice Rationality is satisfied. □

Preliminaries for proofs of Propositions 1 and 2 and Theorems 2, 6 and 7: For results concerning ICEU Bayesian models, it is convenient to begin by establishing some results and terminology used in several subsequent proofs.

> *Lemma A1*: For any game in G, for any Bayesian model of that game, for any player i, for any $s_i \in S_i^*$ such that $\pi_i[E(s_i)] > 0$, and for any $s_{-i} = (s_1, ..., s_{i-1}, s_{i+1}, ..., s_n) \in S_{-i}$ such that, for each player $j \neq i$, $s_j \in S_j^*$, the following is true: If i's prior π_i is independent then:
>
> $\pi_i[E(s_{-i})|\ E(s_i)] = \pi_i[E(s_{-i})]$.

Proof: Consider any $i \in N$; any $s_i \in S_i^*$ such that $\pi_i[E(s_i)] > 0$; and any $s_{-i} = (s_1, ..., s_{i-1}, s_{i+1}, ..., s_n) \in S_{-i}$ such that, for each player $j \neq i$, $s_j \in S_j^*$. By definition, $\pi_i[E(s_{-i})|\ E(s_i)] = \pi_i[E(s_{-i}) \cap E(s_i)]/\ \pi_i[E(s_i)]$. If π_i is independent, $\pi_i[E(s_{-i}) \cap E(s_i)] = \Pi_{j \in N}\ \pi_i[E(s_j)]$ and $\pi_i[E(s_{-i})] = \Pi_{j \neq i}\ \pi_i[E(s_j)]$. Thus, $\pi_i[E(s_{-i})|\ E(s_i)] = \Pi_{j \in N}\ \pi_i[E(s_j)]/\ \pi_i[E(s_i)] = \Pi_{j \neq i}\ \pi_i[E(s_j)] = \pi_i[E(s_{-i})]$. □

For any game in G, consider a Bayesian model of the game in which the profile of priors is $\pi = (\pi_1, ..., \pi_n)$, any player i, any $s_i \in S_i$, and any event E, such that E is the union of one or more elements of i's information partition $\mathcal{I}_i$. Define $U_i(s_i | E)$ as the expected value of $u_i(s)$, given that player i chooses s_i and that the probability distribution over S_{-i} is determined by conditioning i's prior π_i on the event E. We will say that $s_i \in S_i$ is *unconditionally EU-maximising* for player i if, for all $s_i' \in S_i$, $U_i(s_i | \Omega) \geq U_i(s_i' | \Omega)$.

> *Lemma A2*: For any game in G, for any ICEU Bayesian model of that game, for any distinct players i and j, for any $s_i \in S_i$, the following three statements are equivalent:
>
> (a) $s_i \in S_i^*$.
>
> (b) $\pi_j[E(s_i)] > 0$.
>
> (c) s_i is unconditionally expected utility maximising for i.

Proof: Consider any game in G, any ICEU Bayesian model of that game, any distinct players i and j, and any $s_i \in S_i$.

First, suppose (a) holds. By Choice Rationality, there is some $\omega \in \Omega$ such that $b_i[\omega] = s_i$ and $s_i \in \chi_i(\omega)$. Let E be the event such that $\omega \in E \in \mathcal{I}_j$. Caution implies $\pi_j(E(s_i)|E) > 0$ which in turn (as posteriors are obtained from priors by Bayesian updating) implies (b). Conversely, suppose (b) holds. It is then immediate (from the definition of a prior) that there is some $\omega \in \Omega$ such that $b_i[\omega] = s_i$. Thus, (a) and (b) are equivalent. This is *Result 1*. From Result 1, it will suffice to show that (a) implies (c) and (c) implies (b).

Suppose (a) holds. Then, from Choice Rationality and SEU-Maximisation, $U_i(s_i | E) \geq U_i(s_i' | E)$ for all $s_i' \in S_i$ and for all E such that $E \subseteq E(s_i)$ and $E \in \mathcal{I}_i$. Since this inequality holds for each such E, it must also hold for their union. By Knowledge of Own Choice, the union of all such events E is $E(s_i)$. Thus, for all $s_i' \in S_i$, $U_i(s_i | E[s_i]) \geq U_i(s_i' | E[s_i])$. By Independence, using Lemma A1, the probability distribution over S_{-i} that is determined by conditioning π_i on $E(s_i)$ is identical to that determined by conditioning π_i on Ω. Thus, for all $s_i' \in S_i$, $U_i(s_i | \Omega) \geq U_i(s_i' | \Omega)$, i.e. s_i is unconditionally EU-maximising. Thus, (a) implies (c).

Finally, suppose (c) holds but (b) does not. From Result 1, (a) does not hold. But, since S_i^* is non-empty, there must be some $s_i' \neq s_i$ such that $s_i' \in S_i^*$. Consider any such s_i'. By Knowledge of Own Choice, $E(s_i')$ is the union of some elements of $\mathcal{I}_i$. By Independence, using Lemma A1, and the fact that s_i is unconditionally EU-maximising, $U_i(s_i | E[s_i']) \geq U_i(s_i' |$

$E[s_i'])$. So there must be some event $E' \subseteq E(s_i')$ such that $E' \in \mathcal{P}_i$ and $U_i(s_i | E') \geq U_i(s_i' | E')$. Since, by Choice Rationality and SEU-Maximisation, s_i' is SEU-rational for i at each state $\omega \in E'$, the same must be true of s_i. So, by Choice Rationality, $s_i \in \chi_i(\omega)$, for all $\omega \in E'$; and hence, by Caution, the following is true: for all $\omega \in E'$, $\pi_j(E(s_i)|E) > 0$, where E is the event such that $\omega \in E \in \mathcal{P}_j$. Since posteriors are obtained from priors by Bayesian updating, (b) must hold – a contradiction. Thus, (c) implies (b). □

Proof of Proposition 1. Suppose initially that an ICEU Bayesian model of Game 1 exists; and consider any such model. For player 1, *third* is not unconditionally EU-maximising with respect to any probability distribution over player 2's strategies. Thus, by Lemma A2, *third* $\notin S_1$*. Suppose (this is *Supposition 1*) that *second* $\in S_1$* and *right* $\in S_2$*. By Lemma A2, this implies that $\pi_2(E(\textit{second})) > 0$. Then *right* is not unconditionally EU-maximising, and so by Lemma A2, *right* $\notin S_2$*, contradicting Supposition 1. Therefore Supposition 1 is false. Now suppose (this is *Supposition 2*) that *second* $\in S_1$*. By the falsity of Supposition 1, *right* $\notin S_2$*. Since S_2* is non-empty, S_2* = {*left*}. Then, *second* is not unconditionally EU-maximising, and so by Lemma A2, *second* $\notin S_1$*, contradicting Supposition 2. Therefore Supposition 2 is false. Since S_1* is non-empty, S_1* = {*first*}. This implies that each of *left* and *right* is unconditionally EU-maximising and hence, by Lemma A2, S_2* = {*left*, *right*}. Now, to prove existence of ICEU Bayesian models of Game 1, we proceed by construction in the light of these conditions. First, set $\Omega = \{\omega_1, \omega_2\}$ and specify a behaviour function and choiceworthiness functions such that $b_1(\omega_1) = b_1(\omega_2) = \chi_1(\omega_1) = \chi_1(\omega_2) = \{\textit{first}\}$, $b_2(\omega_1) = \{\textit{left}\}$, $b_2(\omega_2) = \{\textit{right}\}$ and $\chi_2(\omega_1) = \chi_2(\omega_2) = \{\textit{left, right}\}$. Then, set $\mathcal{P}_1 = \{\{\omega_1, \omega_2\}\}$ and $\mathcal{P}_2 = \{\{\omega_1\}, \{\omega_2\}\}$ and specify a profile of priors such that $\pi_1(\omega_1) = \pi_2(\omega_1) = \nu$, where $1 > \nu > 0$. Choice Rationality, Knowledge of Own Choice, Independence and Caution hold by construction; and it is trivial to show that SEU-Maximisation is satisfied whenever $\nu \geq 2/3$, so demonstrating existence and multiplicity. □

Proof of Proposition 2. Suppose initially that an ICEU Bayesian model of Game 2 exists; and consider any such model. Using Lemma A2, it is straightforward to show that $S_1^* = \{in_1\} \Rightarrow S_2^* = \{in_2, out_2\}$; that $S_1^* = \{out_1\} \Rightarrow S_2^* = \{in_2\}$; and that $S_1^* = \{in_1, out_1\} \Rightarrow S_2^* = \{in_2\}$. Symmetrically, by Lemma A2, $S_2^* = \{in_2\} \Rightarrow S_1^* = \{in_1, out_1\}$, $S_2^* = \{out_2\} \Rightarrow S_1^* = \{in_1\}$, and $S_2^* = \{in_2, out_2\} \Rightarrow S_1^* = \{in_1\}$. Given that S_1* and S_2* are non-empty, these material implications can be satisfied simultaneously only if *either* (i) $S_1^* = \{in_1\}$ and $S_2^* = \{in_2, out_2\}$

or (ii) $S_1^* = \{in_1, out_1\}$ and $S_2^* = \{in_2\}$. Now, we prove by construction existence and multiplicity of ICEU Bayesian models of Game 2 in which (i) holds. (An analogous procedure establishes the corresponding result for models in which (ii) holds.) First, set $\Omega = \{\omega_1, \omega_2\}$ and specify a behaviour function and choiceworthiness functions such that $b_1(\omega_1) = b_1(\omega_2) = \chi_1(\omega_1) = \chi_1(\omega_2) = \{in_1\}$, $b_2(\omega_1) = \{in_2\}$, $b_2(\omega_2) = \{out_2\}$ and $\chi_2(\omega_1) = \chi_2(\omega_2) = \{in_2, out_2\}$. Then, set $\mathcal{P}_1 = \{\{\omega_1, \omega_2\}\}$ and $\mathcal{P}_2 = \{\{\omega_1\}, \{\omega_2\}\}$. Finally, specify a profile of priors such that $\pi_1(\omega_1) = \pi_2(\omega_1) = \nu$, where $1 > \nu > 0$. For any such ν, Choice Rationality, SEU-Maximisation, Knowledge of Own Choice, Independence, and Caution all hold. □

Proof of Theorem 2. Suppose that an ICEU Bayesian model of Game 3 exists. First, suppose (*Supposition 1*) that there are two distinct players i, j such that $out_i \in S_i^*$ and $out_j \in S_j^*$. Because of the symmetries of the game, there is no loss of generality in setting $i = 1$ and $j = 2$. This implies, by Lemma A2, that $\pi_1(E(out_2)) > 0$. Hence, out_1 is not unconditionally EU-maximising and so, by a further application of Lemma A2, $out_1 \notin \mathrm{S}_1^*$, a contradiction. So Supposition 1 is false. Since there are three players, this entails that there are two distinct players i, j such that $out_i \notin \mathrm{S}_i^*$ and $out_j \notin \mathrm{S}_j^*$. Without loss of generality, set $i = 1$ and $j = 2$. Since S_2^* is non-empty, $S_2^* = \{in_2\}$. Then out_1 is unconditionally EU-maximising and so, by Lemma A2, $out_1 \in S_i^*$, a contradiction. Thus, Game 3 has no ICEU Bayesian model. □

Proof of Theorem 3. Consider any interactive reasoning system $<P_0, R^*, (R_1, \ldots, R_n)>$ among the population N. Suppose that, for some $p \in \varphi(P_0)$, $p \in T(R^*)$. The proof works by repeated application of the same sequence of steps, using the three conditions of the definition of an interactive reasoning system, beginning as follows:

(1) $p \in T(R^*)$ (by supposition)

(2) for all $i \in N$: $R^*(p) \in T(R_i)$ (from (1), using Awareness)

(3) for all $i \in N$: $p \in T(R_i)$ (from (2), using Authority)

(4) for all $j \in N$: $R_j(p) \in T(R^*)$ (from (1), using Attribution)

(5) for all $i, j \in N$: $R^*(R_j[p]) \in T(R_i)$ (from (4), using Awareness)

(6) for all $i, j \in N$: $R_j(p) \in T(R_i)$ (from (5), using Authority)

(7) for all $i, j \in N$: $R_i(R_j[p]) \in T(R^*)$ (from (4), using Attribution)

… and so on, indefinitely.

The role played by p in (1), (2), (3) is played by $R_j(p)$ in (4), (5), (6), by $R_i[R_j(p)]$ in (7), (8), (9), … and so on. Lines (3), (6), (9), … establish that there is iterated reason to believe p in N. □

Proof of Theorem 4: Consider any game in G, and any profile $D = (D_1, ..., D_n)$ of decision rules for its players. Let $< P_0, R^*, (R_1, …, R_n)>$ be the common-reasoning model of the game, defined in relation to D.

We begin by defining, as a counterpart to R^*, an inference structure $R_{–}^*$ which has the same domain and axioms as R^* but different inference rules. The strategy of the proof is to define $R_{–}^*$ so that it can replicate all the steps of the recommendation algorithm, but can do little else. To do this, we define the following sets of inference rules. I_1 consists of the rules of logically valid inference. I_2 is the set of inference rules of the form «from $\{p\}$, infer $R_i(p)$», where $p \in \varphi(P_0)$ and $i \in N$. I_3 is the set of inference rules of the form «from $\{R_i(y_i)\}$, infer z_i», where $i \in N$, y_i is a recommendation to i, and z_i is the prediction about i that is the correlate of y_i. I_4 is the set of inference rules of the form «from $\{y_i\}$, infer z_i», where $i \in N$, y_i is a recommendation to i, and z_i is the prediction about i that is the correlate of y_i. I_5 is the set of inference rules of the form «from $\{z_1, ..., z_{i-1}, z_{i+1}, ..., z_n\}$, infer $z_1 \wedge ... \wedge z_{i-1} \wedge z_{i+1} \wedge ... \wedge z_n$», where $i \in N$ and each z_j is a prediction about the relevant player j. I_6 is the set of inference rules of the form «from $\{D_i, x_{-i}\}$, infer y_i», where $i \in N$, x_{-i} is a collective prediction about $N\backslash\{i\}$, x_{-i} is logically equivalent to the antecedent of some maxim stated by D_i, and y_i is the consequent of that maxim.

R^* is fully specified by its domain $\varphi(P_0)$ and axiom set $A(R^*)$, and by the condition that the set of its inference rules is $I_1 \cup I_2 \cup I_3$. We define $R_{–}^*$ as the inference structure that has the domain $\varphi(P_0)$, the axiom set $A(R_{–}^*) = A(R^*)$ and the set of inference rules $I_4 \cup I_5 \cup I_6$. Note that this implies that $R_{–}^*$ does not have all rules of logically valid inference.

As established in Appendix 1, there is a unique aggregate categorisation function ζ which encodes D. Let $<\mathbb{S}^{+}*, \mathbb{S}^{-}*>$ be the categorisation solution of the game relative to ζ, existence of which is established by Proposition A1.

The proof of Theorem 4 uses the following lemmas:

> *Lemma A3*: For each $i \in N$ and for each $s_i \in S_i$: (i) $s_i \in \mathbb{S}^{+}*$ if, and only if, $p_i(s_i)$ is stated by some theorem in $T(R_{–}^*)$; (ii) $s_i \in \mathbb{S}^{-}*$ if, and only if, $\neg p_i(s_i)$ is stated by some theorem in $T(R_{–}^*)$; (iii) $s_i \in \mathbb{S}^{+}*$ if, and only if, $m_i(s_i)$ is stated by some theorem in $T(R_{–}^*)$; (iv) $s_i \in \mathbb{S}^{-}*$ if, and only if, $\neg m_i(s_i)$ is stated by some theorem in $T(R_{–}^*)$.

Proof: For the purposes of this proof, we extend the definitions of 'encoding', given in Section 9 and Appendix 1, to allow consistent sets of permissibility or possibility sentences for the set of players N to be encoded by categorisations of $\mathbb{S}$. In the case of permissibility, a strategy s_i is assigned to the positive (resp. negative) component of the encoding categorisation if, and only if, $p_i(s_i)$ (resp. $\neg p_i[s_i]$) is in the relevant encoded set. Similarly, in the case of possibility, s_i is assigned to the positive (resp. negative) component of the encoding categorisation if, and only if, $m_i(s_i)$ (resp. $\neg m_i[s_i]$) is in the relevant encoded set.

We now define a *proof algorithm* which progressively 'discovers' the content of the set $T(R_{-}^*)$ by following a particular sequence of steps of reasoning that are licensed by the axioms and inference rules of R_{-}^*. The set of theorems discovered up to the end of any step l is denoted $T_l(R_{-}^*)$. The algorithm is initiated by defining $T_0(R_{-}^*) = A(R_{-}^*)$. At each step $l = 1, 2, \ldots$, one of the three sets of inference rules is used. I_4 is used at $l = 1, 4, 7, \ldots$, I_5 is used at $l = 2, 5, 8, \ldots$, and I_6 is used at $l = 3, 6, 9\ldots$. For each step $l > 0$, T_l is defined as the union of T_{l-1} and the set of sentences that can be inferred from subsets of T_{l-1} using inference rules in the set specified for step l. For $k = 0, 1, 2, \ldots$ we define $C'(k)$ as the categorisation that encodes the intersection of $T_{3k}(R_{-}^*)$ and the set of permissibility sentences. In other words, after every 'cycle' of three steps, the set of 'permissibility theorems so far discovered' is encoded. Since the intersection of $T_0(R_{-}^*)$ and the set of permissibility sentences is $\{\#\}$, $C'(0) = <\varnothing, \varnothing>$. If a step $3k^* > 0$ is reached at which $C'(k^*) = C'(k^*-1)$, the proof algorithm halts. The specification of the algorithm guarantees that, if it halts, no theorems of R_{-}^* remain to be discovered, i.e. that $T_{3k^*}(R_{-}^*) = T(R_{-}^*)$.

In each cycle of three steps, the reasoning carried out by the proof algorithm corresponds with the three operations of one stage of the recommendation algorithm, defined in Section 8. Thus, since ζ encodes D, Proposition A3 implies that, $C'(k) = C(k) = \zeta[C'(k)] = \zeta[C(k-1)]$, for all $k > 0$, where $C(k)$ and $C(k-1)$ are categorisations generated by the categorisation procedure for ζ, defined in Appendix 1. As $C'(0) = C(0) = <\varnothing, \varnothing>$, by definition, the sequence of categorisations $C'(0)$, $C'(1)$, …. defined by the proof algorithm is identical to that generated by the categorisation procedure for ζ. Thus, by Proposition A1, the proof algorithm halts at some finite k^*. Since that algorithm halts only when all theorems of R_{-}^* have been discovered, the categorisation solution $C(k^*)$ encodes all (and only) those permissibility sentences that are stated by theorems of R_{-}^*. This proves parts (i) and (ii) of Lemma A3. The 'only if' implications of parts (iii) and (iv) follow from parts (i) and (ii), together with R_{-}^* having the inference rules in I_4. The 'if' implications of parts (iii) and (iv) also follow from parts (i) and (ii) because $A(R_{-}*)$ contains no possibility sentences other than #, and $I(R_{-}*)$ contains no inference rules which have possibility sentences as conclusions, other than those in I_4. □

Lemma A4: $T(R_{-}*)$ is consistent.

Proof: Let $\#^{n-1}$ denote the conjunction of $n - 1$ instances of #. By inspection of the axioms and inference rules of $R_{-}*$, $T(R_{-}*)$ can be partitioned into three subsets T^1, T^2, and T^3, defined as follows: $T^1 = A(R_{-}*) \cup \{\#^{n-1}\}$; $T^2 = \{p \in T(R_{-}*) \mid p$ is a conjunction of one or more predictions about players, at least one of which is non-null$\}$; $T^3 = \{p \in T(R_{-}*) \mid p$ is a non-null recommendation to some $i\}$. From the definitions of these subsets, Lemma A3 implies that, for each player i, the set of strategies for i whose permissibility (resp. impermissibility) is stated by some recommendation in T^3 is identical to the set of strategies for i in the positive (resp. negative) component of the categorisation solution. As that solution is a categorisation of $\mathbb{S}$, it follows from the definition of a categorisation that T^3 is consistent. Since each element of T^2 is a conjunction of a set of correlates of elements of T^3, and since T^3 is consistent, T^2 is consistent. The non-null elements of T^1 are decision rules for different players, so that, from the definition of a decision rule, T^1 is consistent. Since the elements of T^2 are conjunctions of predictions, since the non-null elements of T^1 are conjunctions of material implications whose consequents are recommendations, and since T^1 and T^2 are each consistent, $T^1 \cup T^2$ is consistent. Finally, by the specification of I_6 and the fact that every sentence in T^3 is the conclusion of an application of an inference rule in I_6, every sentence in T^3 is logically entailed by $T^1 \cup T^2$. Thus, $T^1 \cup T^2 \cup T^3$, i.e. $T(R_{-}*)$, is consistent. □

> *Lemma A5*: (i) $T(R^*)$ is consistent. (ii) For each $i \in N$, and for each $s_i \in S_i$: (a) $p_i(s_i) \in T(R^*)$ if, and only if, $p_i(s_i) \in T(R_{-}*)$; (b) $[\neg p_i(s_i)] \in T(R^*)$ if, and only if, $[\neg p_i(s_i)] \in T(R_{-}*)$.

Proof: By Lemma A4, $T(R_{-}*)$ is consistent. Recall that $A(R^*) = A(R_{-}*)$. R^* differs from $R_{-}*$ only in the following respect: R^* has the set of inference rules $I_1 \cup I_2 \cup I_3$ while $R_{-}*$ has the set $I_4 \cup I_5 \cup I_6$. The only effect of substituting $I_2 \cup I_3$ for I_4 is to allow additional theorems of the form $R_i(p)$ to be derived. This cannot be a source of inconsistency in $T(R^*)$ because R^* has no inference rule by which theorems of the form $\neg R_i(p)$ can be derived. The only effect of substituting I_1 for $I_5 \cup I_6$ is to give R^* all (rather than only some) rules of valid inference. Since (by definition) all decision rules satisfy Deductive Closure, I_6 allows $R_{-}*$ to infer, for any player i, from any given collective prediction x_{-i} about the other players, a recommendation y_i which conjoins all the permissibility sentences for i that are logically entailed by $\{D_i, x_{-i}\}$. Thus, given that $T(R_{-}*)$ is consistent, the substitution of I_1 for $I_5 \cup I_6$ cannot induce inconsistency in $T(R*)$. This proves part (i) of the lemma. Given that $T(R*)$ and $T(R_{-}*)$ are consistent, that $A(R*) = A(R_{-}*)$, and that all decision rules satisfy Deductive Closure, any permissibility sentence that can be derived from $A(R*)$ using inference rules in $I_1 \cup I_2 \cup I_3$ can also be derived from $A(R_{-}*)$ using inference rules in $I_4 \cup I_5 \cup I_{6,}$ and vice versa. This proves part (ii). □

Lemma A6: For each $i \in N$, $T(R_i)$ is consistent.

Proof: By part (i) of Lemma A5, $T(R*)$ is consistent. Consider any $i \in N$. It follows from rules (3) and (4) of the definition of the common-reasoning model that $T(R_i)$ can be partitioned into the subsets T^1, T^2 and T^3, defined as follows: $T^1 = \{p \in \varphi(P_0) \mid p = R*(q)$ for some $q \in T(R*)\}$; $T^2 = T(R*)$; $T^3 = \{p \in \varphi(P_0) \mid p$ is logically entailed by, but not contained in, $T^1 \cup T^2\}$. Since $T(R*)$ is consistent, so is T^2. Since T^1 contains only sentences of the form $R*(.)$, while T^2 is a consistent set which contains no sentence of the form $\neg R*(.)$, $T^1 \cup T^2$ is consistent. Since T^3 contains only sentences that are logically entailed by $T^1 \cup T^2$, $T^1 \cup T^2 \cup T^3$ is consistent. □

Finally, Theorem 4 follows immediately from part (i) of Lemma A5 and Lemma A6. □

Proof of Theorem 5: Consider any profile D of decision rules for any game in G. Part (i) of Theorem 5 follows from Proposition A3, together with part (i) of Proposition A1. Now, define the common-reasoning model with D as its common standard of practical rationality and let $R*$ be common reason in this model. To establish part (ii) of Theorem 5, we have to show that the sentences in the set $\{p \in T(R*) \mid p$ is a permissibility sentence$\}$ are precisely those stated by the final output of the recommendation algorithm for D.

To do this, we define the corresponding inference structure $R_{-}*$, as in the proof of Theorem 4. By part (ii) of Lemma A5, the set $\{p \in T(R*) \mid p$ is a permissibility sentence$\}$ is identical to the set $\{p \in T(R_{-}*) \mid p$ is a permissibility sentence$\}$. By Lemma A3, the sentences in the latter set are encoded by the categorisation solution for the game relative to ζ, where ζ is the aggregate categorisation function which encodes D. Finally, by Proposition A3, the categorisation solution is identical to the categorisation that encodes the combined final output of the recommendation algorithm. □

Proof of Theorem 6: Consider any game in G for which an ICEU Bayesian model exists. Consider any such model M and let its inclusion categorisation be C^M. Let ζ be the aggregate categorisation function which encodes the profile of ICEU decision rules. Let $C(0)$, $C(1)$, ... be the sequence of categorisations of $\mathbb{S}$ induced by the categorisation procedure for ζ.

Lemma A7: For every categorisation C of $\mathbb{S}$: $[C^M \supseteq^* C] \Rightarrow [C^M \supseteq^* \zeta(C)]$.

Proof: For any player i, we define a probability distribution over S_{-i} as *ICEU-consistent* with a categorisation C of $\mathbb{S}$ if (i) for each strategy profile $s_{-i} \in S_{-i}$, the probability of s_{-i} is the product of the marginal probabilities of the individual strategies appearing in s_{-i}; (ii) for each player $j \neq i$, for each $s_j \in S_j$, if s_j is in the positive (resp. negative) component of C, then s_j has strictly positive (resp. zero) marginal probability.

By Lemma A2, if some strategy $s_i \in S_i$ is in the positive component of C^M, it is unconditionally EU-maximising, for some probability distribution over S_{-i} that is ICEU-consistent with C^M; if it is in the negative component of C^M, there is some such distribution for which it is *not* unconditionally EU-maximising (this is *Result 1*). Now consider any categorisation C of $\mathbb{S}$ such that $C^M \supseteq^* C$. Since $C^M \supseteq^* C$, every probability distribution over S_{-i} that is ICEU-consistent with C^M is also ICEU-consistent with C (this is *Result 2*). Because ζ encodes the profile D of ICEU decision rules, if some strategy $s_i \in S_i$ is in the positive component of $\zeta(C)$, it is unconditionally EU-maximising for every probability distribution over S_{-i} that is ICEU-consistent with C; if it is in the negative component of $\zeta(C)$, it is unconditionally EU-maximising for no such distribution (this is *Result 3*). Now suppose Lemma A7 is false. Then, using the fact that, by definition, C^M is exhaustive: *either* (i) for some player i, some strategy $s_i \in S_i$ is in the positive component of C^M and the negative component of $\zeta(C)$, *or* (ii) for some player i, some strategy $s_i \in S_i$ is in the negative component of C^M and the positive component of $\zeta(C)$. Using Results 1, 2 and 3, it can be shown that each of these possibilities implies a contradiction. □

We now complete the proof of the theorem. Trivially, $C^M \supseteq^* <\varnothing, \varnothing>$. By repeated application of Lemma A7, $C^M \supseteq^* \zeta(<\varnothing, \varnothing>)$, $C^M \supseteq^* \zeta\,[\zeta(<\varnothing, \varnothing>)]$, and so on. But, by the initiation and continuation rules for categorisation procedures, $<\varnothing, \varnothing>$, $\zeta(<\varnothing, \varnothing>)$, $\zeta\,[\zeta(<\varnothing, \varnothing>)]$, ... are respectively the categorisations $C(0)$, $C(1)$, $C(2)$, ... induced by the categorisation procedure for ζ. By Proposition A1, this procedure halts at some finite stage k^*. By Proposition A3, Theorem 5, and the definition of the ICEU common-reasoning solution C^*, $C(k^*) = C^*$. Thus, $C^M \supseteq^* C^*$. □

Proof of Theorem 7: Consider any game in G and suppose that its ICEU common-reasoning solution $C^* = <\mathbb{S}^{*+}, \mathbb{S}^{*-}>$ is exhaustive. This implies that $\mathbb{S}^{*+} \cap S_i$ is non-empty and finite, for each player i.

We prove part (i) of the theorem by constructing an ordered quintuple M from C^* and then showing that this M is an ICEU Bayesian model of the game. We construct $M = <\Omega, b(.), \mathcal{I}, \pi, \chi>$ as follows, where Ω is a set of states, and $b(\omega) = (b_1[\omega], ..., b_n[\omega]), \mathcal{I} = (\mathcal{I}_1, ..., \mathcal{I}_n)$, $\pi = (\pi_1, ..., \pi_n)$ and $\chi = (\chi_1, ..., \chi_n)$ are, respectively a behaviour function, an information structure, a profile of priors and a profile of choiceworthiness functions defined with respect to Ω. Set $S_i^* = \mathbb{S}^{*+} \cap S_i$, for each player i, and define $S^* = S_1^* \times ... \times S_n^*$. Define Ω so that there is a one-one mapping from S^* onto Ω; for each $s \in S^*$, let $\omega(s)$ denote the corresponding element of Ω. Thus, by construction, Ω is non-empty and finite, as required. Now define the behaviour function $b(.)$ on Ω so that $b(\omega[s]) = s$, for each $s \in S^*$. Define the information structure $\mathcal{I}$ such that, for each player i, for each strategy $s_i \in S_i^*$: $E(s_i) \in \mathcal{I}_i$. For each player i, fix any independent prior π_i, such that $\pi_i(\omega) > 0$ for all $\omega \in \Omega$. Define χ so that, for each player i, for each state ω, $\chi_i(\omega) = S_i^*$.

It is immediate that, by construction, M satisfies Independence and Knowledge of Own Choice. To verify that it also satisfies Caution, consider any distinct $i, j \in N$, any $s_i \in S_i$, and any $\omega \in \Omega$ such that $s_i \in \chi_i(\omega)$. By construction, $s_i \in \chi_i(\omega)$ implies $s_i \in S_i^*$. The event E such that $\omega \in E \in \mathcal{I}_j$ is $E(b_j(\omega))$. Because π_j is independent, $\pi_j(E(s_i)|E) = \pi_j(E(s_i))$, which is strictly positive by construction.

Now consider any player i and any strategy $s_i \in S_i^*$. As, by Theorem 5 and Proposition A3, C^* is identical to the categorisation solution of the game relative to the aggregate categorisation function ζ which encodes any profile of ICEU decision rules, s_i is unconditionally EU-maximising with respect to all probability distributions over S_{-i} which assign strictly positive probability to strategies in $\mathbb{S}^{*+} \cap \mathbb{S}_{-i}$ and zero probability to strategies in $\mathbb{S}^{*-} \cap \mathbb{S}_{-i}$. Hence, s_i is unconditionally EU-maximising with respect to π_i. Because π_i is independent, and because of the specification of $\mathcal{I}_i$, s_i is expected utility maximising at every state $\omega \in \Omega$. Now consider any strategy $s_i' \notin S_i^*$. A parallel argument shows that s_i' is not expected utility maximising at any state $\omega \in \Omega$. Putting these arguments together: at each state $\omega \in \Omega$, the set of strategies that are SEU-rational for i is S_i^*. Thus, the specification that $\chi_i(\omega) = S_i^*$ for each ω ensures that M satisfies Choice Rationality and SEU-Maximisation. Consequently, M is an ICEU Bayesian model of the game, so proving part (i) of the theorem.

To prove part (ii) of the theorem, consider any ICEU Bayesian model of the game. Since its inclusion categorisation C^M is exhaustive by definition, it follows immediately from Theorem 6 that, if C^* is exhaustive, $C^M = C^*$. □

References

Anderlini, Luca (1990). Some notes on Church's thesis and the theory of games. *Theory and Decision* 29, 19-52.

Asheim, Geir B. and Martin Dufwenberg (2003). Admissibility and common belief. *Games and Economic Behavior* 42, 208-34.

Aumann, Robert (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55, 1–18.

Aumann, Robert (1999a). Interactive epistemology I: knowledge. *International Journal of Game Theory* 28, 263–300.

Aumann, Robert (1999b). Interactive epistemology II: probability. *International Journal of Game Theory* 28, 301–314.

Bacharach, Michael O.L. (1987) A theory of rational decision in games. *Erkenntnis* 27, 17-55.

Binmore, Ken (1987). Modeling rational players: Part I. *Economics and Philosophy* 3, 179-214.

Binmore, Ken (1988). Modeling rational players: Part II. *Economics and Philosophy* 4, 9-55.

Bonanno, Giacomo (2012). Epistemic foundations of game theory. University of California, Davis, Department of Economics working paper 12-11.

Börgers, Tilman and Larry Samuelson (1992). "Cautious" utility maximisation and iterated weak dominance. *International Journal of Game Theory* 21, 13–25.

Brandenburger, Adam (2007). The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory* 35: 465-92.

Brandenburger, Adam, Amanda Friedenberg and H. Jerome Keisler (2008). Admissibility in Games. *Econometrica*, 76, 307-52.

Cubitt, Robin P. and Robert Sugden (1994). Rationally justifiable play and the theory of noncooperative games. *Eonomic Journal* 104, 798–803.

Cubitt, Robin P. and Robert Sugden (1997). Rationally justifiable play and the theory of noncooperative games. In M. Bacharach, L.-A. Gérard-Varet, P. Mongin and H.S. Shin (eds.) *Epistemic Logic and the Theory of Games and Decisions*, Dordrecht: Kluwer Academic Publishers.

Cubitt, Robin P. and Robert Sugden (2003). Common knowledge, salience and convention: a reconstruction of David Lewis's game theory. *Economics and Philosophy* 19, 175–210.

Cubitt, Robin P. and Robert Sugden (2011). The reasoning-based expected utility procedure. *Games and Economic Behavior*, 71, 328-338.

Gintis, Herbert (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton: Princeton University Press.

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

Monderer, Dov and Dov Samet (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior* 1, 170–190.

Nash, John F. (1951). Non-cooperative games. *Annals of Mathematics* 54, 286-95.

Norde, Henk (1999). Bimatrix games have quasi-strict equilibria. *Mathematical Programming* 85, 35–49.

Paternotte, Cedric (2011). Being realistic about common knowledge: a Lewisian approach. *Synthese* 183, 249-76.

Pearce, David G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52, 1029-1050.

Perea, Andrés (2011). An algorithm for proper rationalizability. *Games and Economic Behavior* 72, 510-25.

Perea, Andrés (2012). *Epistemic Game Theory: Reasoning and Choice*. Cambridge, UK: Cambridge University Press.

Samuelson, Larry (1992). Dominated strategies and common knowledge. *Games and Economic Behavior* 4, 284–313.

Sillari, Giacomo (2005). A logical framework for convention. *Synthese* 147, 379-400.

Squires, David (1998). Impossibility theorems for normal form games. *Theory and Decision* 44, 67-81.

Vanderschraaf, Peter (1998). Knowledge, equilibrium and convention. *Erkenntnis* 42, 65–87.