CeDEx

Centre for Decision Research and Experimental Economics

Discussion Paper Series

ISSN 1749-3293

CeDEx Discussion Paper No. 2008–01

Common Reasoning in Games

Robin Cubitt and Robert Sugden

March 2008

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of Public Economics, Individual Choice under Risk and Uncertainty, Strategic Interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/economics/cedex/ for more information about the Centre or contact

Karina Whitehead
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0) 115 95 15620
Fax: +44 (0) 115 95 14159
karina.whitehead@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/economics/cedex/papers/index.html

Common reasoning in games

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

25 February 2008

[+]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 dedicate this paper to the late Michael Bacharach, as an acknowledgement of his role in pioneering the analysis of players' reasoning in game theory and in inspiring our interest in exploring the foundations of the theory. We are grateful for comments on earlier versions to a referee and an associate editor; to Giacomo Bonnano, Adam Brandenburger, John Collins, and Michael Mandler; and to participants in various seminars, conferences and workshops at which we have presented the paper. Sugden's work was supported by the Economic and Social Research Council (award no. RES 051 27 0146).

Abstract

This paper makes three related contributions to noncooperative game theory: (i) a solution concept (the "ICEU solution"), which is generated by an iterative procedure that constructs trinary partitions of strategy sets and deals with problems arising from weak dominance; (ii) a class of models of players' reasoning, inspired by David Lewis's work on common knowledge, which can together represent common knowledge of rationality for any consistent conception of individual practical rationality; and, using these ingredients, (iii) a diagnosis of paradoxes associated with the concept of common knowledge of rationality, as represented in Bayesian models of games.

Short title

Common reasoning in games

JEL classification

C72

1. Introduction

This paper makes three main contributions to noncooperative game theory: a solution concept which deals in a new way with the problems arising from weak dominance; a class of models of players' reasoning; and a diagnosis of paradoxes associated with the concept of common knowledge of rationality, as represented in Bayesian models of games. Although seemingly of different types, these contributions are closely related.

Our solution concept (the "ICEU solution") is distinctive in being the output of an iterative procedure which constructs *trinary* partitions of the set of strategies for the relevant game. At each stage of the procedure, strategies are categorised into those that have been found to be rationally playable, those that have been found to be not rationally playable, and those that have not yet been found to be either. At the following stage, non-zero probabilities are required for all strategies in the first category, in the spirit of the principle of weak dominance, while zero probability is required for all strategies in the second category. As long as a strategy remains in the "grey area" of the third category, no restriction is imposed on its probability. In comparison with existing methods of iteratively deleting weakly dominated strategies, this procedure has many intuitively attractive characteristics, attributable to the fact that categorisations made at later stages of the procedure cannot undermine the justification for categorisations made at earlier stages. However, there is no general guarantee that the grey area will be empty when the procedure halts. Thus, in general, the ICEU solution is a trinary partition of the set of strategies. This unusual property calls for a systematic justification.

Our analysis of reasoning provides this. Our concept of a "model of reasoning" has its roots in the work of Lewis (1969), who used a similar approach in his path-breaking analysis of common knowledge.[1] A model of reasoning gives an explicit representation of all the steps of reasoning by which, for a given standard of individual decision-theoretic (or *practical*) rationality, players can arrive at propositions about the rational permissibility or impermissibility of strategies. It is intrinsic to this approach that, for any given strategy and in relation to any given standard of practical rationality, there are three possibilities: *either* the permissibility of the strategy can be established by reasoning; *or* its impermissibility can be so established; *or* neither its permissibility nor its impermissibility can be established. Each member of the class of models of reasoning embeds some standard of practical rationality; given that standard, it provides a formal foundation for a solution concept (the "categorisation solution") in the form of a trinary partition of the strategy set. The ICEU

solution is the categorisation solution generated when the standard of practical rationality is expected utility maximisation with beliefs that are independent and cautious (ICEU), in the sense of Pearce (1984) and Borgers and Samuelson (1992).

Since a model of reasoning can represent players' reasoning about one another's reasoning, it can show how particular propositions about rational play can become common knowledge, in a Lewisian sense. Thus, a model of reasoning can be interpreted as a representation of common knowledge of rationality (henceforth, CKR). In this context, the "rationality" that is common knowledge is partly defined by the standard of practical rationality embedded in the model. Models of reasoning, considered as a class, provide a family of representations of common knowledge of different standards of practical rationality. This allows a new perspective on what have previously been seen as paradoxes of CKR.

Intuitively, CKR seems to be a meaningful idealisation, in the same sense that perfect competition is a meaningful idealisation in price theory, or frictionless surfaces are in theoretical mechanics. (Throughout the paper, we use the terms "common knowledge" and "CKR" to refer to the intuitive concepts that game theory attempts to formalise. Since what is at issue is how these concepts are best formalised, it would be unhelpful to start out by privileging any particular set of formal definitions.[2]) However, formalisation of CKR has often produced surprising results (Brandenburger, 2007; Samuelson, 2004).[3]

The approach to modelling CKR that has become standard in the modern literature is the Bayesian framework due to Aumann (1987). Further developments of this approach have been made by, among others, Tan and Werlang (1988), Dekel and Gul (1997) and Aumann (1999a, b). Although Aumann's model is logically consistent, apparently natural extensions, intended to capture ICEU or similar concepts, generate puzzles and even contradictions in certain games (Borgers and Samuelson, 1992; Samuelson, 1992; Cubitt and Sugden, 1994). One possible response to this is to reject the extensions. However, to those who see the extensions as having compelling motivations, the games in which the puzzles arise are paradoxical exhibits for the Bayesian approach. In our view, there is also a deeper conception of the paradox that does not require ICEU to be seen as a uniquely, or even especially, compelling conception of practical rationality, but only as one that is internally consistent. The deeper paradox is that, within the Bayesian approach, substituting one internally consistent conception of practical rationality for another seems to affect whether CKR is even possible.

We believe that the best response to the paradoxes is not to debate which restrictions on the Bayesian model deliver the "intuitively most reasonable" results, but rather to ask what features of the Bayesian modelling strategy prevent it from encompassing all internally consistent standards of practical rationality. Our approach to this question is to start from models of reasoning that identify the inferences that rational players can make from premises provided by a given standard of practical rationality. This leads to a diagnosis that can be expressed, very roughly, as follows. A Bayesian model requires that strategies are partitioned into those that are "included" in the model (that is, are played at some state of the world) and those that are not, and that this partition is common knowledge. If the model is interpreted as a representation of CKR, there must be common knowledge that included strategies may be chosen by rational players while excluded strategies are not chosen. But our analysis of models of reasoning shows that, in general, players' reasoning about the implications of a given standard of rationality may sometimes be unable to establish whether particular strategies are rationally playable or not. The attempt by the Bayesian approach to impose a binary partition even when such a grey area exists is the source of the paradoxes.

We show how models of reasoning provide non-paradoxical accounts of common knowledge of the ICEU standard of rationality for the games that give rise to paradox when the Bayesian model is extended to represent common knowledge of the same standard. Further, our models of reasoning 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. Thus, crucially for a resolution of the deeper paradox as we conceive it, we can give a consistent representation of CKR, not just for ICEU, but for any standard of practical rationality that is coherent at the individual level, and for any game.

As we have said, our approach develops aspects of Lewis (1969) that are often overlooked. But many other game theorists have offered models of players' reasoning towards conclusions about the playability or non-playability of strategies of different kinds. One well-known approach introduces a dynamic element into Bayesian reasoning, as in Harsanyi's "tracing procedure" (Harsanyi, 1975; Harsanyi and Selten, 1988) and Skyrms's (1989, 1990) "dynamic deliberation". In these models, players have common knowledge of their Bayesian rationality and update their subjective probabilities in the light of information generated by their knowledge of how other players have updated theirs. The main results derived from these models depend on the assumption that players' prior probabilities are common knowledge. These priors are taken as data by the model, which then represents

how players' beliefs are modified. In our approach, in contrast, standards of practical rationality are taken as data; the reasoning model represents how, using those standards, players can arrive at beliefs about the possibility or otherwise of different strategies.

An approach somewhat more similar to ours is taken by Binmore (1987, 1988) in his 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). This analysis might be interpreted as demonstrating the general impossibility of justifying Bayesian binary partitions as the product of players' reasoning. 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.[4] We see our work as in a similar spirit as Binmore's and Bacharach's. However, we focus less on negative results and more on what conclusions *can* be reached by coherent modes of reasoning that individuals might endorse.[5]

We do not claim that our Lewisian approach is the *only* way to resolve the paradoxes of CKR. For example, some paradoxes can be eliminated if, following the approach proposed by Brandenburger (2007) and Brandenburger *et al* (2008), one replaces the standard Bayesian concept of probabilistic beliefs (which our approach can accommodate) with that of lexicographic probability systems. However, we believe that our approach throws new light on the difficulties that game theorists have repeatedly found in trying to formalise apparently intuitive ideas about CKR.

The remainder of the paper is organised as follows. Section 2 presents a Bayesian framework for modelling CKR and, using two paradoxical exhibits, illustrates puzzles that arise when it is extended to capture the ICEU conception of rationality, using the concept of an "ICEU Bayesian model". Section 3 presents our general concept of a categorisation solution and the special case of the ICEU solution. For the benefit of readers more interested in comparative properties of solution concepts than foundational issues, Section 3 (and Appendix 1, which compares the ICEU solution to existing concepts) can be read in isolation. However, a full justification for the ICEU solution requires the concept of a common reasoning model, which we develop in Sections 4-7. Section 7 presents our main results about such models. We establish that, for every coherent standard of practical rationality, a consistent common reasoning model exists, and that this model justifies a

corresponding categorisation solution. Section 8 investigates the implications of the preceding analysis for ICEU Bayesian models. Finally, Section 9 presents our diagnosis of the paradoxes. Appendix 2 contains proofs of all formal results not proved in the main text.

2. CKR in a Bayesian model: two paradoxes

In this section, we present two paradoxes stemming from the Bayesian approach to modelling games.

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[6] of strategies $s = (s_1, ..., s_n)$, there is a profile $u(s) = (u_1[s], \ldots, u_n[s])$ of 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}. We impose that, for all $i, j, \in N$, $S_i \cap S_j = \varnothing$. 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.[7] This convention allows a conveniently compact notation in later sections.

We define a Bayesian model, for any game in G, so that it specifies a set of states of the world; players' behaviour; players' knowledge; players' subjective beliefs; and a standard of decision-theoretic rationality.

Uncertainty is represented in the Bayesian model 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*.

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 chosen by the players at ω. Stochastic choice 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^* (resp. S_i^*) is the set of strategy profiles (resp. strategies for i) *included* in the Bayesian model.

Players' knowledge is represented by an *information structure* $\mathtt{I} = (\mathtt{I}_1, ..., \mathtt{I}_n)$. For each player i, $\mathtt{I}_i$ is an information partition of Ω, representing what i knows at each state. $K_i(E)$, where E is an event, is the event $\{\omega \in \Omega \mid \exists E' \in \mathtt{I}_i: (\omega \in E') \wedge (E' \subseteq E)\}$.[8] 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 Bayesian modelling framework. We use the qualifier "Bayesian" to distinguish this theoretical construct from the intuitive concept.) Since Ω is the universal set, then for all i, $K_i(\Omega)$ at all ω; thus Ω is Bayesian common knowledge at all states.

For any player i, a *prior* is a function $\pi_i: \Omega \rightarrow (0, 1]$ satisfying $\Sigma_{\omega \in \Omega} \pi_i(\omega) = 1$; $\pi_i(\omega)$ is interpreted as a subjective probability. We extend this notation to events by defining, for each event E, $\pi_i(E) = \Sigma_{\omega \in E} \pi_i(\omega)$. A prior π_i is *independent* if, for all players j, k (with $j \neq k$), for all strategies $s_j \in S_j^*$, $s_k \in S_k^*$: $\pi_i(E[s_j] \cap E[s_k]) = \pi_i(E[s_j])\pi_i(E[s_k])$. A profile $\pi = (\pi_1, ..., \pi_n)$ of priors is independent if each component π_i is independent. Posterior probabilities, conditional on events, are defined from priors by means of Bayes's rule. The requirement that, for each player i and for each state ω, $\pi_i(\omega) > 0$ guarantees that posterior probabilities are well-defined and that the priors of different players have common support. Common support is a much weaker condition than that of common priors (i.e. that, for all distinct players i and j, $\pi_i = \pi_j$).[9] We allow but do not impose the latter, as it is not required for the paradoxes that we present.

We define a *choice function* for 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 $E \in \mathtt{I}_i$, for all ω, $\omega' \in E$: $\chi_i(\omega) = \chi_i(\omega')$. The interpretation is that a choice function encapsulates some normative standard of practical rationality; $\chi_i(\omega)$ is the set of strategies which, according to that standard, 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 can be conditioned only on events that he observes.

A choice function is a device for representing the implications of whatever decision principles are taken as "rational". However, a Bayesian analysis requires a particular standard of rationality, namely that of maximizing subjective expected utility. 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 $E \in \mathtt{I}_i$, 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 $\mathtt{I}_i$ 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 $\omega \in E \in \mathtt{I}_i$. 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.[10] The choice function χ_i is *SEU-rational* with respect to $\mathtt{I}_i$ and π_i if, for all $\omega \in \Omega$, $\chi_i(\omega)$ is the set of strategies that are SEU-rational for i at ω with respect to $\mathtt{I}_i$ and π_i.

We define a *Bayesian model* of a particular game as an ordered quintuple $<\Omega, b(.), \mathtt{I}, \pi, \chi>$, where Ω is a finite, nonempty set of states and b(.), $\mathtt{I} = (\mathtt{I}_1, ..., \mathtt{I}_n)$, $\pi = (\pi_1, ..., \pi_n)$ and $\chi = (\chi_1, ..., \chi_n)$ are, respectively a behaviour function, an information structure, a profile of priors and a profile of choice 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-Maximization. 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 $\mathtt{I}_i$ 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-Maximization stipulates that the standard of rationality is the maximization 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.

The following result is a precursor to our discussion of paradoxes:

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

Proposition 1 is implied by the analysis of Aumann (1987).[11] It 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.

Given Proposition 1, it is natural to ask whether further conditions can be imposed on Bayesian models. We consider two additional requirements:

Independence (of Priors). The profile π of priors is independent.

Privacy (of Tie-Breaking). For all distinct $i, j \in N$, for all $\omega \in \Omega$, for all $s_i \in S_i$: $s_i \in \chi_i(\omega) \Rightarrow \omega \notin K_j(\Omega \backslash E[s_i])$.

Independence rules out the possibility that some player i believes that the strategy choices of any two distinct players are correlated with one another. 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. If the representation of CKR is to apply to such cases, it must be possible to impose Independence on the model.

Privacy requires that if some strategy s_i is choiceworthy for player i at some state ω, then it is not the case that some other player j knows at ω that s_i is not chosen. Given that Choice Rationality holds, if s_i is choiceworthy for player i at state ω, to suppose that, at the same state, another player j could know that s_i is not chosen would be to suppose that $\chi_i(\omega)$ is not a singleton and that j can replicate the tie-breaking mechanism that i uses to discriminate between options which, according to the standard of rationality, are equally choiceworthy. 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. Hence, whether tie-breaking mechanisms are private or not is an empirical question, not one that can resolved by a priori considerations of rationality and common knowledge. If the representation of CKR is to apply to cases in which tie-breaking rules are private, it must be possible to impose Privacy on the model.

Privacy can also be interpreted as a principle of *caution* with respect to posteriors. As prior probabilities are constrained to be nonzero, the proposition $\omega \notin K_j(\Omega \backslash E[s_i])$ implies that, at ω, j's posterior probability for $E(s_i)$ is nonzero. Thus, Privacy requires that, if a strategy s_i is choiceworthy for player i at some state, then, at that state, other players assign nonzero probability to its being chosen. Consequently, if it is choiceworthy at any state, s_i is an element of S_i*.

We define an *independent-priors cautious expected utility (ICEU) Bayesian model* as a Bayesian model which satisfies Independence and Privacy. Such a model represents in the Bayesian framework a situation where there is CKR, with rationality governed by the *ICEU standard*. According to this standard, each player's beliefs assign independent probabilities to other players' strategies, zero probability to strategies regarded as not rationally playable, and non-zero probability to all strategies regarded as rationally playable; and each player maximizes expected utility relative to these beliefs. Clearly the ICEU standard is more restrictive than expected utility maximisation alone, but it is attractive for certain contexts

for the reasons given above. More importantly, it would be paradoxical if an otherwise coherent representation of CKR could not accommodate the view of rationality embedded in the ICEU standard without giving rise to puzzles or impossibility.

However, that is how matters turn out. We use two games to illustrate this.

Our first exhibit, illustrating the *Proving Too Much Paradox,* is Game 1. (Although we work entirely with normal-form games, it is worth noting that Game 1 is the normal form of a Centipede game.)

Game 1

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

Proposition 2: In every ICEU Bayesian model of Game 1, $S_1^* = \{first\}$ and $S_2^* = \{left, right\}$.

Proposition 2 is paradoxical as it implies that, in every ICEU Bayesian model of Game 1, player 1 must assign a prior probability greater than 2/3 to player 2's choosing *left* (since, otherwise, *second* would be SEU-rational at every state), when player 1 knows that player 2 is indifferent between *left* and *right*. If player 2 is indifferent between her strategies, which of them she chooses must be determined by a non-rational tie-breaking mechanism. The properties of this mechanism cannot be determined by assumptions about rationality and common knowledge. So, why must player 1 believe that player 2's tie-breaking mechanism selects *left* with probability greater than 2/3? In more general terms, the paradox is that a particular belief, held by a particular player, is common to *all* ICEU Bayesian models, with the apparent implication that the existence of this belief is implied *merely* by the assumption that the ICEU standard of rationality is common knowledge, when there seems to be no way in which the player could reason her way to that belief, given only the knowledge that is attributed to her by that assumption. In this sense, we seem to have proved too much.

Game 1 does not yield an outright inconsistency in the conditions that define an ICEU Bayesian model. In fact, these conditions are mutually consistent for every 2-player game in G.[12] However, an inconsistency can be shown using the following 3-player game introduced by Cubitt and Sugden (1994):

Game 2 (Tom, Dick and Harry)[13]

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

The paradox is contained in the following proposition:

Proposition 3. There is no ICEU Bayesian model of Game 2.

Thus, there are games for which the idea that each player has an ICEU standard of rationality, and that this is common knowledge, cannot be represented in a Bayesian model.

Proposition 1 shows that *some* standards of rationality – in particular, subjective expected utility maximization *without* caution – can be represented in Bayesian models without contradiction. However, 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. Whether or not one thinks rationality really requires players to obey the ICEU standard, it is puzzling that common knowledge of ICEU cannot be represented in a Bayesian model.[14]

We present an alternative approach to modelling CKR in Sections 3-7.

3. Iterative categorisation

In this section, we define a class of iterative procedures, each of which can be interpreted as a summary of a reasoning process. (At this stage, the idea of "summarising a reasoning process" will be presented intuitively; a full analysis will be given in Sections 4-7.) Each member of this class of procedures is based on a different profile of "categorisation functions". A categorisation function for a given player can be interpreted as mapping from statements about what her opponents might in fact do, to statements about what is permissible for that player. As we will show in later sections, a very wide range of

conceptions of practical rationality can be represented by categorisation functions. In this section, however, we focus on the ICEU standard of rationality and on the corresponding iterative procedure, the "ICEU procedure". This procedure generates a new solution concept, the "ICEU solution".

For some games, there are similarities between the ICEU procedure and iterative deletion of weakly dominated strategies (IDWDS). Although IDWDS has a long history, it is well-known that it has some unattractive features, especially when viewed from a CKR perspective (Samuelson, 1992). We argue in this section that, despite embodying a requirement of caution that is a close relative of weak dominance, the ICEU procedure avoids the analogous problems and is quite distinct from IDWDS. A fuller comparison of the ICEU procedure with existing deletion procedures is given in Appendix 1.

Consider any game in G. (Our analysis applies to every such game but, to avoid clutter, we hold the game fixed and, except when stating formal results, suppress clauses of the form "for all games in G" where possible.) For any player i, we define a *categorisation* of S_i as an ordered pair $<S_i^+, S_i^->$ of disjoint subsets of S_i, satisfying the conditions that $S_i^- \subset S_i$ and, if $S_i \backslash S_i^- = \{s_i\}$ for any $s_i \in S_i$, then $S_i^+ = \{s_i\}$.[15] S_i^+ is the *positive component* of the categorisation and S_i^- the *negative component*. The set of all possible categorisations of S_i is denoted $\Phi(S_i)$. We use two interpretations, on each of which a categorisation can be read as summarising a statement. On one interpretation, this statement asserts of the strategies in S_i^+ that they are rationally permissible for i and of the strategies in S_i^- that they are not rationally permissible. On the other interpretation, the statement asserts of the strategies in S_i^+ that i might possibly play them and of the strategies in S_i^- that i will not play them. The requirement that S_i^+ and S_i^- are disjoint reflects the mutual incompatibility of permissibility and impermissibility (resp. possibility and impossibility). The further conditions imposed by the definition require that not all of i's strategies are categorised as impermissible (resp. impossible) and that, if all but one are categorised as impermissible (resp. impossible), the remaining strategy is categorised as permissible (resp. possible). In general, a categorisation of S_i defines a trinary partition of S_i, whose elements are S_i^+, S_i^-, and the set $S_i \backslash (S_i^+ \cup S_i^-)$ of strategies that are uncategorised. If this third set is empty, the categorisation is *exhaustive*.

We now introduce a notation which allows us to combine categorisations of the strategy sets of two or more players, while maintaining the distinction between positive and negative components. 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. We define an "addition" or "union" relation $\cup^*$ between such categorisations such that $\cup^*_{i\in N'} <S_i^+, S_i^-> \equiv <\cup_{i\in N'} S_i^+, \cup_{i\in N'} S_i^->$. We will say that each $\cup^*_{i\in N'} <S_i^+, S_i^->$ is a *categorisation* of the set $\cup_{i\in N'} S_i$. The set of all categorisations of the latter set is denoted $\Phi(\cup_{i\in N'} S_i)$. This notation uses unions of sets of strategies, some of which "belong" to one player and some to another. However, the labelling convention defined in Section 2 enables us, as analysts, to keep track of which strategies belong to whom.

Two kinds of combination of categorisation are particularly significant. Members of the first kind arise when $N' = N$; they combine categorisations of all players' strategy sets to produce categorisations of $\cup_{i\in N} S_i$. We use $\mathbb{S}$ as a shorthand notation for $\cup_{i\in N} S_i$, so that $\Phi(\mathbb{S})$ denotes the set of all categorisations of this first kind. The positive and negative components of such categorisations will typically be denoted $\mathbb{S}^+$ and $\mathbb{S}^-$. Members of the second kind arise when $N' = N\backslash\{i\}$; they combine categorisations of the strategy sets of all players except some player i to produce categorisations of $\cup_{i\in N\backslash\{i\}} S_i$. We use $\mathbb{S}_{-i}$ as a shorthand notation for $\cup_{i\in N\backslash\{i\}} S_i$, so that $\Phi(\mathbb{S}_{-i})$ denotes the set of all categorisations of this second kind. The positive and negative components of such categorisations will typically be denoted $\mathbb{S}_{-i}^+$ and $\mathbb{S}_{-i}^-$. Where it is unnecessary to spell out the positive and negative components, we use C, C′ and so on as shorthand notation for particular categorisations of $\mathbb{S}$; C_{-i}, C_{-i}' and so on as notation for particular categorisations of $\mathbb{S}_{-i}$; and C_i, C_i' and so on as notation for particular categorisations of S_i.

A further piece of notation allows us to say that one categorisation of a given set of strategies is unambiguously "larger" than, or "has more content than", another. Consider any categorisations $C_i' = <S_i^{+\prime}, S_i^{-\prime}>$, $C_i'' = <S_i^{+\prime\prime}, S_i^{-\prime\prime}>$ that are elements of $\Phi(S_i)$, for some player i. We define a binary relation $\supset^*$ between such categorisations as follows: $C_i'' \supset^* C_i'$ if *either* (i) $S_i^{+\prime\prime} \supset S_i^{-\prime}$ and $S_i^{+\prime\prime} \supseteq S_i^{-\prime}$ *or* (ii) $S_i^{+\prime\prime} \supseteq S_i^{-\prime}$ and $S_i^{+\prime\prime} \supset S_i^{-\prime}$. On the reading of categorisations as statements about permissibility (resp. possibility), $C_i'' \supset^* C_i'$ has a natural interpretation: it indicates that the statement represented by C_i'' is strictly stronger than that represented by C_i'. We also define a weaker relation $\supseteq^*$ such that $C_i'' \supseteq^* C_i'$ if *either* $C_i'' \supset^* C_i'$ *or* $C_i'' = C_i'$. This notation is extended in an obvious way to categorisations of $\Phi(\mathbb{S})$ and $\Phi(\mathbb{S}_{-i})$. For example, consider categorisations $C' = <\mathbb{S}^{+\prime}, \mathbb{S}^{-\prime}>$, $C'' = <\mathbb{S}^{+\prime\prime}, \mathbb{S}^{-\prime\prime}>$ in $\Phi(\mathbb{S})$. In this case, $C'' \supset^* C'$ if *either* (i) $\mathbb{S}^{+\prime\prime} \supset \mathbb{S}^{-\prime}$ and $\mathbb{S}^{+\prime\prime} \supseteq \mathbb{S}^{-\prime}$ *or* (ii) $\mathbb{S}^{+\prime\prime} \supseteq \mathbb{S}^{-\prime}$ and $\mathbb{S}^{+\prime\prime} \supset \mathbb{S}^{-\prime}$.

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 $C_{-i}'' \supset^* C_{-i}'$ then $f_i(C_{-i}'')$

$\supseteq^*$ $f_i(C_{-i}')$. A categorisation function may be interpreted as summarising reasoning which produces categorisations of S_i, conditional on categorisations of $\mathbb{S}_{-i}$. On this reading, the categorisations of S_i attribute permissibility and impermissibility to strategies available to i; and the categorisations of $\mathbb{S}_{-i}$ on which they are conditioned attribute possibility and impossibility to strategies available to players other than i. Thus, different categorisation functions for player i correspond to different ways of conditioning statements about what is rationally permissible for i on statements about what other players might do. The Monotonicity requirement formalises an obvious principle of reasoning, namely that any conclusions that can be derived from a given set of premises can also be derived from any strictly stronger set of premises.[16]

We will be concerned with profiles of categorisation functions. It is convenient to express the content of a given profile $f = (f_1, \ldots, f_n)$ of categorisation functions for individual players as a single function $\zeta: \Phi(\mathbb{S}) \rightarrow \Phi(\mathbb{S})$, constructed as follows. Let $C = \langle \mathbb{S}^+, \mathbb{S}^- \rangle$ be any categorisation of $\mathbb{S}$. For each player i, define $C_{-i} = \langle \mathbb{S}^+ \backslash S_i, \mathbb{S}^- \backslash S_i \rangle \in \Phi(\mathbb{S}_{-i})$. Next, define $S_i^{+\prime}$ and $S_i^{-\prime}$ as, respectively, the positive and negative components of $f_i(C_{-i})$. Finally, define $\zeta(C) = \cup^*_{i \in N} \langle S_i^{+\prime}, S_i^{-\prime} \rangle$. We will say that ζ *summarises* f. A function $\zeta: \Phi(\mathbb{S}) \rightarrow \Phi(\mathbb{S})$ that summarises some profile f of categorisation functions is an *aggregate categorisation function*. Notice that, for any given profile f, there is one and only one function ζ which summarises it.

We can now introduce the central concept of this section. For any aggregate categorisation function ζ, the *categorisation procedure* is defined by the following pair of instructions, which generate a sequence of categorisations $C(k) \equiv \langle \mathbb{S}^+(k), \mathbb{S}^-(k) \rangle$ of $\mathbb{S}$, for successive stages $k \in \{0, 1, 2, \ldots.\}$, inductively, as follows:

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

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

If there exists $k^* \in \{1, 2, \ldots..\}$ such that $C(k^*) = C(k^*-1)$ then, for all $k' > k^*$, $C(k') = C(k^*)$. Since this renders further application of the continuation rule redundant, we will say that the procedure *halts* at the lowest value of k^* for which $C(k^*) = C(k^*-1)$. Then, $C(k^*)$ is the *categorisation solution* of the game, relative to ζ. As the procedure halts if and when a categorisation is mapped to itself, the categorisation solution is a *fixed point* of ζ.

The categorisation procedure, for a given ζ, can be interpreted as shorthand for a process of reasoning in phases. The first phase starts from premises represented by the *null categorisation* $C(0) = \langle \varnothing, \varnothing \rangle$ and draws from them conclusions about the permissibility or

otherwise of strategies, represented by the categorisation C(1). For the second phase, the strategies whose permissibility (resp. impermissibility) was established in the first phase are taken as possible (resp. as impossible), so allowing further conclusions to be drawn about permissibility, captured by the categorisation C(2); and so on, with each phase taking as given the possibility (resp. impossibility) of strategies whose permissibility (resp. impermissibility) was established at the previous phase.

Different categorisation procedures are obtained by specifying the aggregate categorisation function in different ways. Before focussing on the case where ζ incorporates the ICEU standard of rationality, we establish a result that, in virtue of our definition, applies to all categorisation procedures.

> *Theorem 1*: 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.

Part (ii) of Theorem 1 guarantees that a unique categorisation solution exists, for any aggregate categorisation function for any game. Part (i) is a step in the proof of part (ii), but is also significant in its own right, in terms of the interpretation of a categorisation procedure as shorthand for a reasoning process whereby strategies are classified as permissible or impermissible. On that reading, it shows that the reasoning process has the attractive property that each of its phases re-affirms the classifications made by previous phases.

The *reaffirmation property* of a categorisation procedure, expressed by part (i) of Theorem 1, is connected to the fact that, at each stage k, a categorisation procedure induces a trinary partition $C(k) = \langle \mathbb{S}^+(k), \mathbb{S}^-(k) \rangle$ of $\mathbb{S}$. The three elements of this partition are: $\mathbb{S}^+(k)$, $\mathbb{S}^-(k)$ and the residual set $\mathbb{S} \backslash (\mathbb{S}^+(k) \cup \mathbb{S}^-(k))$. For any $k > 0$, we will say that strategies in $\mathbb{S}^+(k) \backslash \mathbb{S}^+(k-1)$ are *accumulated at stage k*, and that strategies in $\mathbb{S}^-(k) \backslash \mathbb{S}^-(k-1)$ are *deleted at stage k*. By virtue of part (i) of Theorem 1, $\mathbb{S}^+(k)$ contains the strategies accumulated at stages 1, ..., k, while $\mathbb{S}^-(k)$ contains the strategies deleted at stages 1, .., k. The residual set $\mathbb{S} \backslash (\mathbb{S}^+(k) \cup \mathbb{S}^-(k))$ contains those strategies not accumulated or deleted at any stage 1, …, k. As the procedure may halt at a stage at which the residual set is non-empty, the categorisation solution is not necessarily exhaustive.

These properties are common to all categorisation procedures, as we have defined them. However, we now present the particular procedure which embeds the ICEU standard of rationality. For any player i, we define a probability distribution over S_{-i} as *IC-consistent*

with a categorisation C_{-i} of $\mathbb{S}_{-i}$ if it satisfies the following three conditions. First, probabilities are independent in the sense that, for any $s_{-i} \in S_{-i}$, the probability of s_{-i} is the product of the marginal probabilities of the individual strategies appearing in s_{-i}. Second, every strategy in the positive component of C_{-i} has strictly positive marginal probability. Third, every strategy in the negative component of C_{-i} has zero marginal probability. The second condition can be interpreted as a requirement of caution. For each player i, there is a unique *ICEU categorisation function*, defined as the categorisation function f_i such that, for all $C_{-i} \in \Phi(\mathbb{S}_{-i})$, $f_i(C_{-i}) = \langle S_i^+, S_i^- \rangle$ where S_i^+ is the set of strategies in S_i that are expected utility maximising for every probability distribution over S_{-i} that is IC-consistent with C_{-i}; and S_i^- is the set of strategies in S_i that are not expected utility maximising for any probability distribution over S_{-i} that is IC-consistent with C_{-i}. This construction guarantees that the Monotonicity property required for a categorisation function is satisfied.[17]

The profile that associates with each player her ICEU categorisation function is summarised by the *ICEU aggregate categorisation function*. The *ICEU procedure* and the *ICEU solution* are, respectively, the categorisation procedure and the categorisation solution, defined by the ICEU aggregate categorisation function. Each of these concepts is unique, for every game in G. In the remainder of this section, we illustrate the ICEU solution by comparing it briefly to IDWDS.

The most obvious difference between the ICEU procedure and IDWDS is that the former has operations of accumulation and deletion, while the latter has only a deletion operation. As a result, the ICEU procedure induces a trinary partition of each S_i, whereas IDWDS induces a binary partition. Further, there are significant differences between the deletions made by the ICEU procedure and those made by IDWDS.

A strategy for player i which is weakly dominated, given some specification of the strategies available to other players, may no longer be weakly dominated if some of those strategies are removed from consideration. This is the source of two well-known peculiarities of IDWDS: sensitivity of the outcome to the order of deletions and the "re-entry problem" that the deletion of a particular strategy for one player at a given stage of IDWDS may lose its justification at a later stage, because the justification for the original deletion depended on the possibility of a strategy for another player that is subsequently deleted. In contrast, the specification of the ICEU procedure guarantees that the order of deletions and accumulations is determined uniquely and, more fundamentally, that the case for a given deletion or accumulation at one stage is never undercut at a later one. At a given stage of the

ICEU procedure, players are required to assign zero probability to previously deleted strategies and strictly positive probability to *previously accumulated* strategies. In contrast, at a given stage of IDWDS, players are implicitly required to assign zero probability to previously deleted strategies and strictly positive probability to strategies *that have not yet been deleted*. This difference is crucial. Strategies which have been accumulated in the ICEU procedure are never deleted later, so the cautious rationale for requiring strictly positive probability on them is secure. But, strategies which, at some stage of IDWDS, have not yet been deleted may still be deleted later, so the case for requiring strictly positive probability on them can be undercut.

We illustrate the importance of the accumulation operation for the difference between the ICEU procedure and IDWDS with two examples, the first of which is the following game:

Game 3:

		Player 2	
		left	*right*
Player 1	*first*	1,1	0,0
	second	0,0	0,0

For this game, the ICEU procedure runs as follows: C(0) = <∅, ∅>; C(1) = <{*first, left*}, ∅> (i.e. *first* and *left* are accumulated at stage 1); C(2) = <{*first, left*}, {*second*, *right*}> (i.e. *second* and *right* are deleted at stage 2); C(3) = C(2) (i.e. the procedure halts at stage 3). In this example, there is an obvious similarity between the ICEU procedure and IDWDS, namely that each deletes both (and only) *second* and *right*. However, the two procedures arrive at this result in different ways. IDWDS deletes *second* and *right* immediately, on grounds of weak dominance; *first* and *left* remain as an undeleted residual. For the ICEU procedure, *second* and *right* are only deleted after *first* and *left* have been accumulated. Interpreting the ICEU procedure as summarising a reasoning process, we can say, for example, that it is only once player 1 has established that *left* is permissible for player 2 that she can conclude that she must assign strictly positive probability to that strategy, and hence that *second* is impermissible for her.

Further differences between the ICEU procedure and IDWDS emerge in the next example:

Game 4

		Player 2	
		left	*right*
Player 1	*first*	1, 1	1, 1
	second	0, 0	1, 0
	third	2, 0	0, 0
	fourth	0, 2	0, 0

For this game, the ICEU procedure runs as follows: $C(0) = <\varnothing, \varnothing>$; $C(1) = <\{left\}, \{fourth\}>$; $C(2) = <\{left, right\}, \{second, fourth\}>$; $C(3) = C(2)$. Thus, for Game 4, the ICEU procedure distinguishes between two types of undeleted strategy: those accumulated (*left* and *right*) and those neither accumulated nor deleted (*first* and *third*). This contrasts with IDWDS, which makes no distinction between undeleted strategies. To see the basis for the distinction, note that, in the absence of any restrictions on beliefs, all strategies in Game 4 could be justified except for strictly dominated *fourth*. Each of *left* and *right* is optimal for *all* beliefs that satisfy the resulting requirement to assign zero probability to *fourth*. In contrast, the feature that *first* and *third* share is merely that, for each of them, there are *some* beliefs, satisfying the relevant requirements (i.e. to assign strictly positive probability to both *left* and *right*), that would justify it and some that would justify the other.

Another difference is that the ICEU solution is uniquely defined, whereas IDWDS can yield different outcomes depending on the order of deletions. In Game 4, *second*, *fourth* and *right* are all weakly dominated. If the first deletion made by IDWDS is *fourth*, the only other deletion is *second*; but, if the first deletion by IDWDS is *right*, then *first*, *second* and *fourth* are all deleted. The latter case illustrates the re-entry problem: the initial deletion of *right* is justified because of the possibility that *fourth* might be played, but this justification is undercut by the subsequent deletion of *fourth*. No analogous problem arises for the ICEU procedure, as it would never delete *right* unless *fourth* had been accumulated at an earlier stage; and that can never happen, since *fourth* is strictly dominated.

We conclude that the ICEU procedure has attractive features and that the presence of an operation of accumulation, as well as one of deletion, is central to how the ICEU procedure avoids well-known peculiarities of IDWDS. Because it has two operations rather than one, the ICEU procedure can induce a trinary partition of strategies at a given stage; and it can halt at a stage where the partition induced is still trinary. We claim that this feature is consistent with the view of the ICEU procedure as a summary of a process of reasoning, as there is nothing puzzling about the idea that reasoning from given premises may fail to

establish whether a particular strategy is permissible or not. Fortunately, however, we do not need to rely on this intuitive argument to support our claim that the ICEU procedure is a shorthand representation of a reasoning process. We can present the reasoning process itself.

4. Reasoning schemes

As a first step in developing our model of players' reasoning, we introduce our representation of a mode of reasoning.

We define a mode of reasoning in relation to some domain P of *propositions* within which reasoning takes place. This domain may be interpreted as the class of propositions defined within some formal structure or language. Initially, we impose only minimal conditions on P, which we will state as the concepts required to do so are defined. Later, when we apply our model to games, we will specify P precisely, using a particular formal language for game-relevant propositions that satisfies these conditions. We use p, q, r, to denote particular propositions in P and use the logical connectives $\neg$, $\wedge$, and $\Rightarrow$ to make up complex propositions from atomic ones, where those connectives are defined by the usual semantic rules.[18] We impose that every proposition in P can be made up from some set of atomic propositions using a finite number of logical connectives.

For any finite subset $Q = \{q_1, ..., q_m\}$ of P and any $p \in P$, p is *logically entailed* by Q if $\neg p \wedge q_1 \wedge ... \wedge q_m$ is a contradiction. Two propositions p and q are *logically equivalent* if each is logically entailed by the singleton containing the other. The conjunction of an empty set of propositions is the *null proposition*, which we treat as a tautology, denoted #.[19] We impose that P contains #. A set of propositions is *consistent* if no conjunction of its elements is a contradiction.

An *inference rule* in domain P is a two-place instruction of the form "from …, infer ...", where the first place is filled by a finite subset of P and the second place by an element of P. The elements of the set that fills the first place are the *premises* of the inference rule; the proposition that fills the second place is its *conclusion*. An inference rule is *valid* if its conclusion is logically entailed by the set of its premises.

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, with $\# \in A(R)$, 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 $T(R) = T_0(R) \cup T_1(R) \cup \ldots$. Each proposition in T(R) is *derivable* in R. Intuitively, a proposition is a theorem of R if it can be derived using only axioms and inference rules of R.

We will say that I(R) *includes the rules of valid inference* if, for every finite $Q \subseteq P$ and every $p \in P$, if p is logically entailed by Q then "from Q, infer p" $\in$ I(R). An inference structure R such that I(R) includes the rules of valid inference is a *reasoning scheme*. Thus, if R is a reasoning scheme, every proposition in P that is logically entailed by the theorems of R is itself a theorem of R. Since every tautology in P is logically entailed by {#}, every such tautology is a theorem of R. Note, however, that I(R) can contain inference rules other than those of valid inference. This allows us to represent forms of inference which, although not licensed by deductive logic, are used in game-theoretic and practical reasoning.

We will say that a reasoning scheme R is *consistent* if T(R) is consistent. If A(R) is consistent and all the inference rules of R are valid, it is immediate that R is consistent; but, in general, there are inconsistent reasoning schemes, as well as consistent ones. Our ultimate 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; such a framework must allow the possibility of inconsistency. We therefore do not impose consistency as part of the definition of a reasoning scheme.

Before proceeding, it is worth highlighting an important difference between the Lewisian approach that we follow and the Bayesian one. The Bayesian framework may be seen as a model of the world which incorporates a specification, captured by an information partition, of what each player knows in each state. The conception of knowledge is objective, relative to what the modeller has deemed to be true. Thus, within the Bayesian framework, it is true by definition that, for any event E and any player i, $K_i(E) \subseteq E$, so that "knowledge implies truth". In contrast, our approach is subjectivist. Rather than modelling how things are, we model only what players *take to be true*. We interpret a reasoning scheme R as representing a mode of reasoning that a person might use when forming his beliefs. The axioms of R represent propositions which, within this mode of reasoning, are accepted without further need for justification. The inference rules of R represent the steps by which propositions are inferred from (sets of) other propositions. The theorems of R exhaust the conclusions that can be reached by this mode of reasoning. A person *endorses* a reasoning scheme R if he takes its axioms to be true and accepts the authority of its inference rules. We will say that a person who endorses R has *reason to believe* each of its theorems.

For any proposition p, we use the notation R(p) as shorthand for the proposition that p is a theorem of R. This notation allows us to represent reasoning schemes which interact, in the sense of having theorems about what is derivable in other reasoning schemes, or indeed in themselves. For example, $R_1[R_2(p)]$ denotes the proposition that "p is a theorem of R_2" is a theorem of R_1, where R_1 and R_2 are (possibly distinct) reasoning schemes.[20]

5. Common reasoning in a population

Our approach, following Lewis (1969), 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 commonly attributed to other such agents. This core of shared reasoning may or may not exhaust the reasoning endorsed by each agent separately. 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*.

We now postulate a set P_0 of *primitive propositions*, such that $\# \in P_0$, and such that no proposition in P_0 can be expressed by any formula containing any of the terms R*(.), $R_1(.),..., R_n(.)$. For each $k \geq 1$, we define P_k to contain all of the following propositions (and no others): (i) every proposition which can be constructed from the elements of P_{k-1} using a finite number of logical connectives from the set $\{\neg, \wedge, \Rightarrow\}$, (ii) every proposition of the form R*(p) where $p \in P_{k-1}$; and (iii) every proposition 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 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 propositions, 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 $p \in \varphi(P_0)$: $R^*(p) \Rightarrow [R^*(p) \in A(R_i)]$.
>
> *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 represents the idea that agents know of common reason in the sense that, if some proposition p is a theorem of common reason, the fact that it is such a theorem is treated as self-evident by each agent's private reason. The Authority condition requires that each agent accepts the authority of common reason in the following sense: from the premise that some proposition p is a theorem of common reason, the private reason of each agent infers p as a conclusion. The Attribution condition requires that, from any premise p, common reason infers the conclusion $R_i(p)$ in relation to each agent i. In this sense, common reason attributes its own theorems to the private reason of each agent.

We will say that, in population N, there is *iterated reason to believe* some proposition p if all finitely nested propositions of the form $R_i(R_j(... R_k(p)...))$ for $i, j, ..., k \in N$ are true. Iterated reason to believe is a core concept in Lewis's game theory, analogous with Bayesian common knowledge in Aumann's. More precisely, it is a consequence of the property that Lewis defines as "common knowledge" (Cubitt and Sugden, 2003). The following theorem establishes that, in an interactive reasoning system, there is iterated reason to believe all propositions that are derivable in R*:

> *Theorem 2*: 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 proposition $p \in T(R^*)$, there is iterated reason to believe p in population N.

Our method of modelling CKR will be to represent game-theoretic rationality in terms of axioms and inference rules, and to attribute these to common reason in an interactive reasoning system. By virtue of Theorem 2, any propositions that are derivable 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 propositions

In this section, we develop a general method of representing principles of practical rationality in the form of a particular kind of proposition, which we call a "decision rule". The concept of a decision rule does not presuppose particular substantive principles of practical rationality, but uses a purely formal notion of "permissibility".

Fix any game in G. For any player i and any $s_i \in S_i$, $p_i(s_i)$ denotes the proposition "s_i *is permissible for* i", by which is meant that, normatively, i may choose s_i (but not that he

must, since two or more strategies might be permissible for him). The formula $m_i(s_i)$ denotes the descriptive proposition "s_i *might in fact be chosen by* i" or, for short, "s_i *is possible for* i". Propositions of the form $p_i(s_i)$ or $\neg p_i(s_i)$ are *permissibility propositions*. For each permissibility proposition $p_i(s_i)$ or $\neg p_i(s_i)$, the corresponding *possibility proposition* $m_i(s_i)$ or $\neg m_i(s_i)$ is its *correlate*, and vice versa; the possibility proposition states what may in fact happen if the correlate permissibility proposition is acted on.

We will say of any conjunction of propositions that it *asserts* each of its conjuncts. A *recommendation* to a player i is a conjunction of the elements of a consistent set of permissibility propositions referring to the strategies available to i, satisfying the conditions that not every strategy in S_i is asserted to be impermissible and that, if every strategy but one in S_i is asserted to be impermissible, the remaining strategy is asserted to be permissible. Analogously, a *prediction* about a player i is a conjunction of the elements of a consistent set of possibility propositions referring to i's strategies, satisfying the conditions that not every strategy in S_i is asserted to be impossible and that, if every strategy but one in S_i is asserted to be impossible, the remaining strategy is asserted to be possible. The definition of a prediction rests on 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 suffices to establish that the remaining one is possible (indeed, certain). Given these points, the definition of a recommendation reflects the principle that normative requirements must be logically capable of being satisfied ("ought implies can").

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 is the conjunction of the correlates of its component permissibility propositions; the correlate of a prediction is the conjunction of the correlates of its component possibility propositions.[21] For any non-empty set of players $N' \subseteq N$, a *collective prediction* about N' is a conjunction of some set of predictions about individual players, where that set contains no more than one non-null prediction about each member of N'. For each player i, the null proposition is both a recommendation to i and a prediction about i; and, for every $N' \subseteq N$, the null proposition is also a collective prediction about N.

Recommendations to a player i, collective predictions about the set of players $N\backslash\{i\}$, and predictions about i are propositions that have special roles to play in what follows. To distinguish them from other propositions, we use y_i to denote a recommendation to i, x_{-i} to

denote a collective prediction about $N\backslash\{i\}$, and z_i to denote a prediction 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 propositions asserted 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}: if x_{-i} is a collective prediction about $N\backslash\{i\}$, then 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 y_i': if x_{-i}' is a collective prediction about $N/\{i\}$, if y_i' is a non-null permissibility proposition for i, and 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 propositions 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.

The concept of a decision rule allows us to represent principles of practical rationality *as propositions*. That is required by our strategy of representing players' reasoning explicitly, within a domain of propositions. However, the *content* of a decision rule can also be represented, or *encoded*, by a categorisation function, as defined in Section 3. We now explain the nature of this encoding, so forging the key link between the analyses of Section 3 and Sections 4-7.

Consider any recommendation y_i to any player i. The content of the proposition y_i can be expressed by specifying two subsets of S_i: the set S_i^+ of strategies which are asserted to be permissible for i, and the set S_i^- of strategies which are asserted to be impermissible. It follows from the definition of a recommendation that $<S_i^+, S_i^->$ is a categorisation of S_i. We will say that $<S_i^+, S_i^->$ *encodes* y_i. For every recommendation, there is a unique

categorisation that encodes it. Note that the null recommendation # is encoded by $<\varnothing, \varnothing>$. Conversely, every categorisation $<S_i^+, S_i^->$ of S_i encodes each member of a set of logically equivalent recommendations to i.[22]

Analogous concepts of encoding can be defined for collective predictions, for profiles of predictions, and for profiles of recommendations. Consider any player i, and any collective prediction x_{-i} about the set of players $N\backslash\{i\}$. The content of the proposition x_{-i} can be encoded in a categorisation C_{-i} of $\mathbb{S}_{-i}$, the positive (resp. negative) component of which contains all strategies asserted to be possible (resp. impossible). Similarly, consider any profile $(z_1, \ldots, z_n)$ of predictions about players 1, ..., n. The content of this profile can be encoded in a categorisation C of $\mathbb{S}$, the positive (resp. negative) component of which contains all strategies asserted to be possible (resp. impossible) by any of the predictions z_1, …, z_n. In the analogous way, the content of any profile $(y_1, \ldots, y_n)$ of recommendations can be encoded in a categorisation C′ of $\mathbb{S}$.

We now extend the concept of "encoding" to decision rules. Consider any decision rule D_i for player i. This is a conjunction of a set F_i of maxims of the form $x_{-i} \Rightarrow y_i$. The content of each such maxim can be expressed by the ordered pair $<C_{-i}, C_i>$ where C_{-i} is the categorisation of $\mathbb{S}_{-i}$ that encodes x_{-i} and C_i is the categorisation of S_i that encodes y_i. Because of Distinct Antecedents, no two maxims in F_i have antecedents encoded by the same categorisation of $\mathbb{S}_{-i}$. However, as a decision rule for a player need not contain maxims covering all collective predictions for other players, there may be categorisations of $\mathbb{S}_{-i}$ which do not encode the antecedent of any maxim in F_i. Such cases can be represented as follows. Let C_{-i}' be a categorisation of $\mathbb{S}_{-i}$ which does not encode the antecedent of any maxim in F_i. This fact can be represented by the ordered pair $<C_{-i}', <\varnothing, \varnothing>>$, signifying that the decision rule makes no recommendation conditioned on a collective prediction encoded by C_{-i}'. The set of ordered pairs constructed in this paragraph together define a unique function f_i: $\Phi(\mathbb{S}_{-i}) \rightarrow \Phi(S_i)$; this function *encodes* the decision rule D_i.

The following result establishes a formal equivalence between decision rules and categorisation functions which will be used later.

> *Proposition 4*: 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 is a categorisation function for i.

This proposition, combined with the definition of an aggregate categorisation function (given in Section 3), implies the following: for any profile $D = (D_1, \ldots, D_n)$ of decision rules, there exists a unique profile $f = (f_1, ,,, f_n)$ of categorisation functions, and 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.

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 propositions, so that it contains # and, for each $i \in N$ and for each $s_i \in S_i$, the propositions $m_i(s_i)$ and $p_i(s_i)$ (and no other propositions). This specification implies that all decision rules are in $\varphi(P_0)$. Next, we specify a particular profile of decision rules $D^* = (D_1^*, ..., D_n^*)$, which is to be built into the model as the standard of practical rationality. We then construct reasoning schemes $R^* = <\varphi(P_0), A(R^*), I(R^*)>$, $R_1 = <\varphi(P_0), A(R_1), I(R_1)>$, ... , $R_n = <\varphi(P_0), A(R_1), I(R_n)>$ 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 nothing else:

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

(4) $I(R_i)$ contains the rules of valid inference and those specified below, and nothing else:

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, $< P_0, R^*, (R_1, \ldots, R_n)>$ is an interactive reasoning system. 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 practical rationality. This inference rule embeds in common reason the following principle: from the proposition that i has reason to believe that some recommendation applies to him, it can be inferred that he will act on it. In this sense, common reason attributes practical rationality to each player.

An interactive reasoning system $<P_0, R^*, (R_1, \ldots, R_n)>$ defined in relation to a profile of decision rules D* and constructed according to rules (1) to (4) is a *common-reasoning model* of the game. We will say that D* is the *common standard of practical rationality* in that model. Our Lewisian rendition of CKR is a consistent common-reasoning model.

It is immediate that, for any game in G and for any profile of decision rules D* for its players, a corresponding (and unique) common-reasoning model exists: the model is constructed by following rules (1) to (4). What is not so obvious is whether the model so constructed is consistent. If we can show that, for every game in G, and for every profile of decision rules for that game, the resulting common-reasoning model is consistent, we will have shown that our framework can represent without contradiction common knowledge of *any* conception of rationality that can be formulated as a profile of decision rules. The following theorem establishes this property:

> *Theorem 3*: 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.

We prove this result jointly with Theorem 4 (below), using the fact, established in Section 6, that every profile of decision rules can be encoded as an aggregate categorisation function.

We are now able to formalise the claim of Section 3 that a categorisation procedure may be interpreted as a summary of a reasoning process. The following proposition establishes, for any given common-reasoning model and its corresponding aggregate categorisation function, a link between that function and certain kinds of theorems of common reason.

> *Proposition 5*: Consider any profile D* of decision rules for any game in G. Let R* be common reason in the common-reasoning model in which D* is the common standard of practical rationality and let ζ be the aggregate categorisation function that encodes D*. For any categorisation C of $\mathbb{S}$: if C encodes a profile of recommendations $(y_1, \ldots, y_n)$ such that, for each player i, $y_i \in T(R^*)$, then
>
> (i) C encodes a profile $(z_1, \ldots, z_n)$ of predictions such that, for each player i, z_i is the correlate of y_i and $z_i \in T(R^*)$; and
>
> (ii) $\zeta(C)$ encodes a profile of recommendations $(y_1', \ldots, y_n')$ such that, for each player i, $y_i' \in T(R^*)$.

To see the significance of this result, consider any profile D* of decision rules. As well as defining a common-reasoning model, D* defines an aggregate categorisation function ζ and, thus, a particular categorisation procedure. This procedure tracks inferences made in common reason R* in the common reasoning model. The procedure is initiated by setting $C(0) = <\varnothing, \varnothing>$. This categorisation encodes the null recommendation # which,

trivially, is a theorem of R*. By part (i) of Proposition 5, it also encodes the null prediction, which is also a theorem of R*. By part (ii) of the Proposition, $\zeta[C(0)]$ encodes a profile of recommendations that are theorems of R*. But, $\zeta[C(0)]$ is C(1), the output of the first stage of the categorisation procedure. As it encodes a profile of recommendations that are theorems of R*, part (i) of Proposition 5 implies that it also encodes a profile of predictions that are theorems of R*. And so on. At each subsequent stage, the process takes as its input a profile of recommendations encoded by the output of the previous stage, converts them into their correlate predictions and then obtains a further profile of recommendations from them. Thus, for each stage $k \in \{0, 1, \ldots. k^*\}$, C(k) encodes a profile of recommendations that (together with their correlate predictions) are theorems of R*.

We have established that the categorisation solution, interpreted as a statement about recommendations, asserts permissibility propositions that are theorems of R*. The following theorem encapsulates this result and adds the converse principle that every permissibility proposition that is a theorem of R* is asserted by the categorisation solution.

> *Theorem 4*: Consider any profile D* of decision rules for any game in G. Let R* be common reason in the common-reasoning model in which D* is the common standard of practical rationality and let ζ be the aggregate categorisation function that encodes D*. Let C(k*) be the categorisation solution relative to ζ. For each $i \in N$ and for each $s_i \in S_i$:
>
> (i) s_i is in the positive component of C(k*) if, and only if, $R^*[p_i(s_i)]$; and
>
> (ii) s_i is in the negative component of C(k*) if, and only if, $R^*[\neg p_i(s_i)]$.

It is time to take stock of the development of our approach. For any game in G, and for any principles of practical rationality for its players that can be stated as decision rules, the corresponding common-reasoning model represents a situation in which those principles are taken as axiomatic by common reason (therefore, by Theorem 2, being the object of iterated reason to believe among the players) and in which reasoning satisfies the requirements of deductive and game-theoretic inference embedded in rules (2) and (4). Theorems 3 and 4 are the results that establish the credentials of this approach.

Theorem 3 shows that, for any given profile D* of decision rules, common reason and every player's private reason are consistent in the common-reasoning model of the game in which D* is the common standard of practical rationality. This demonstrates the coherence of our approach to modelling the reasoning of rational players.

Theorem 4 shows that, for any given profile D* of decision rules, the categorisation solution of the game defined by the corresponding aggregate categorisation function encodes

exactly those permissibility propositions that are theorems of common reason in the common-reasoning model in which D* is the common standard of practical rationality. Thus, given Theorem 3, Theorem 4 provides a formal foundation for the general concept of a categorisation solution presented in Section 3.

It is important to stress that all of the results of this section hold for *any* profile of decision rules and so for any conception of individual rationality that can be expressed as such a profile. Thus, Theorems 2, 3 and 4 deliver on our opening promise to separate the content of a conception of practical rationality for a game from the representation of what it is for some such conception to be common knowledge.

In order to apply these general results to the special case in which the conception of practical rationality is provided by the ICEU standard, all that is needed is to represent that standard as a profile D* of decision rules. We do so as follows. Consider any player i, and any collective prediction x_{-i} about $N\backslash\{i\}$. Consider the categorisation of $\mathbb{S}_{-i}$ whose positive component consists of those strategies whose possibility is asserted by x_{-i} and whose negative component consists of those strategies whose impossibility is asserted by x_{-i}. We will say that a probability distribution is IC-consistent with x_{-i} if it is IC-consistent with this categorisation. (IC-consistency in relation to categorisations is defined in Section 3.) Define $S_i^+(x_{-i}) = \{s_i \in S_i |\ s_i$ is EU-maximising for every probability distribution that is IC-consistent with $x_{-i}\}$ and $S_i^-(x_{-i}) = \{s_i \in S_i |\ s_i$ is not EU-maximising for any probability distribution that is IC-consistent with $x_{-i}\}$. A proposition y_i is an *ICEU response* for i to x_{-i} if it is a conjunction of permissibility propositions that (i) asserts the permissibility of a strategy if, and only if, it is in $S_i^+(x_{-i})$; and (ii) asserts the impermissibility of a strategy if, and only if, it is in $S_i^-(x_{-i})$. By construction, an ICEU response is a recommendation to i. Thus, we may define an *ICEU maxim* as a maxim whose consequent is an ICEU response to its antecedent. For every player i, and for every collective prediction x_{-i} about $N\backslash\{i\}$, there is a set of logically equivalent ICEU maxims for i. An *ICEU decision rule* for i is constructed by conjoining all elements of a set F_i of ICEU maxims for i, such that, for each possible collective prediction x_{-i}, F_i contains exactly one maxim whose antecedent is logically equivalent to x_{-i}.

This terminology presupposes that F_i satisfies the Distinct Antecedents and Deductive Closure conditions (otherwise a conjunction of its elements would not be a decision rule). That Distinct Antecedents is satisfied is obvious from the construction of F_i. That Deductive Closure is also satisfied follows from the definition of an ICEU response.

Intuitively, an ICEU response to a given collective prediction treats that prediction as providing information about probabilities – namely that each strategy that is asserted to be possible (resp. impossible) has non-zero (resp. zero) probability. Applying the criterion of expected utility maximisation, an ICEU response identifies as permissible (resp. impermissible) those strategies that are unambiguously optimal (resp. sub-optimal), conditional on (and only on) the information contained in the relevant collective prediction. Thus, each ICEU maxim provides a complete statement of the implications of the expected utility criterion, conditional on the relevant information, and provides no other information.

From any profile of ICEU decision rules, a consistent *ICEU common-reasoning model* can be constructed. (That is an immediate implication of Theorem 3.) It is straightforward to verify that every profile of ICEU decision rules is encoded by the ICEU aggregate categorisation function, as defined in Section 3. Thus, by Proposition 5 and Theorem 4, the permissibility propositions that are theorems of common reason in the ICEU common-reasoning model are summarised by the ICEU solution, and the process whereby common reason arrives at them is summarised by the ICEU procedure. Our approach has grounded the ICEU solution in a model of common reasoning.

8. ICEU Bayesian models revisited

In this Section, we compare our common-reasoning approach with the more conventional Bayesian approach, introduced in Section 2. For this comparison, we focus on the case where the conception of practical rationality is provided by the ICEU standard. In broad terms, ICEU common-reasoning models and ICEU Bayesian models can be interpreted as alternative ways of representing a shared idea: that players are rational in the ICEU sense, and that this is common knowledge. What is the relationship between those representations? And does the common-reasoning approach illuminate the paradoxes of the Bayesian one?

Intuitively, one might expect to find certain general relationships between Bayesian concepts and common-reasoning concepts. For a given game and on the (non-innocuous) assumption that an ICEU Bayesian model exists, each ICEU Bayesian model can be interpreted as describing players' knowledge in some world in which there is common knowledge that players' beliefs and strategy choices are consistent with the ICEU standard. An ICEU common-reasoning model can be interpreted as a description of what players have reason to believe about the game, and of the steps of reasoning by which those beliefs can be

derived, given that the ICEU standard is axiomatic in common reason. Given these interpretations, it is natural to conjecture that if some possibility proposition is a theorem of common reason in an ICEU common-reasoning model, there is common knowledge of its content in every ICEU Bayesian model; or, more precisely, that if the possibility (resp. impossibility) of some strategy is a theorem of common reason, then every ICEU Bayesian model includes (resp. does not include) that strategy. We now show that this *inclusion conjecture* is indeed correct.

For the remainder of this Section, we again fix some game in G. We have shown that a consistent ICEU common-reasoning model of the game exists and that, in all such models, the set of permissibility propositions that are theorems of common reason is the same. From the specification of a common-reasoning model, the same is true of possibility propositions. Thus, there is no loss of generality in fixing a specific ICEU common-reasoning model, which from now on we call "the" ICEU common-reasoning model. In this Section, we use R* to denote common reason in this model. There is a unique ICEU aggregate categorisation function ζ and a unique ICEU solution for the game; the ICEU solution is a categorisation of $\mathbb{S}$, which we denote C*. From Theorem 4, for every strategy in the positive component of C*, it is a theorem of R* that that strategy is possible (likewise permissible); and, for every strategy in the negative component of C*, it is a theorem of R* that that strategy is impossible (likewise impermissible). However, as C* is not necessarily exhaustive, there may be strategies such that neither their possibility nor their impossibility is a theorem of R*.

The concept of an ICEU Bayesian model is well-defined, but such a model does not necessarily exist: in general, there may be more than one such model, exactly one model, or none. Suppose, however, that some ICEU Bayesian model M of the game exists. This model specifies, for each player, a set S_i* of included strategies. Thus, it induces an *inclusion categorisation* $C^M = <\mathbb{S}^+, \mathbb{S}^->$ of $\mathbb{S}$, such that $\mathbb{S}^+ = \cup_{i \in N} S_i$* and $\mathbb{S}^- = \cup_{i \in N} S_i \backslash S_i$*. Notice that C^M is exhaustive, by construction. On the supposition that some model M exists, we investigate the relationship between C^M and C*.

The following theorem provides the basis for a proof of the inclusion conjecture:

> *Theorem 5*: 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 ζ be the ICEU aggregate categorisation function for the game. Then, for every categorisation C of $\mathbb{S}$: $[C^M \supseteq^* C] \Rightarrow [C^M \supseteq^* \zeta(C)]$.

This result allows us to link ICEU Bayesian models of the game with the ICEU solution. Trivially, $C^M \supseteq^* <\varnothing, \varnothing>$. Thus, by repeated application of Theorem 5, $C^M \supseteq^* \zeta(<\varnothing, \varnothing>)$, $C^M \supseteq^* \zeta[\zeta(<\varnothing, \varnothing>)]$, and so on. But $\zeta(<\varnothing, \varnothing>)$, $\zeta[\zeta(<\varnothing, \varnothing>)]$, ... are respectively the outputs C(1), C(2), ... of stages 1, 2, ... of the ICEU categorisation procedure. At some finite stage, this procedure halts, and its output is the ICEU solution C*. Hence:

> *Corollary*: 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 solution of the game. Then, $C^M \supseteq^* C^*$.

This establishes the truth of the inclusion conjecture.

So far in this section, we have left open the possibility that no ICEU Bayesian model exists. However, the following theorem establishes a sufficient condition for the existence of such a model.

> *Theorem 6*: Consider any game in G. Let ζ be the ICEU aggregate categorisation function for the game. If there exists an exhaustive categorisation C' of $\mathbb{S}$ such that $C' = \zeta(C')$, then there exists an ICEU Bayesian model of the game with inclusion categorisation C'.

Since the ICEU solution is a fixed point of ζ, the following is a consequence of Theorems 5 and 6:

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

Theorem 7 establishes that, *if the ICEU solution of the game is exhaustive*, then an ICEU Bayesian model exists and, for every such model, its inclusion categorisation coincides with the ICEU solution. Thus, in this special case, and with respect to categorisations, the Bayesian and common-reasoning approaches are mutually consistent. Of course, an ICEU Bayesian model consists of more than its inclusion categorisation. In particular, each ICEU Bayesian model has its own profile π of priors. We now consider the status of these priors in relation to the common-reasoning approach.

The idea that each player assigns *some* probability to each state of the world is fundamental to subjective expected utility theory. Thus, if an ICEU Bayesian model is to describe a world in which ICEU is the standard of practical rationality, it must include a full specification of players' subjective beliefs in that world. There need be no presumption that these beliefs are uniquely determined by considerations of rationality. However, if the Bayesian and common-reasoning approaches are to be consistent, the priors of an ICEU Bayesian model should not contradict theorems of R*. We will say that a profile $\pi = (\pi_1, ...,$

π_n) of priors *respects* a categorisation C of $\mathbb{S}$ if, for every pair of distinct players i and j, for every strategy $s_j \in S_j$: $\pi_i[E(s_j)] > 0$ if s_j is in the positive component of C, and $\pi_i[E(s_j)] = 0$ if s_j is in the negative component of C. It follows from the Corollary to Theorem 5 that, for every game in G for which an ICEU Bayesian model exists, the priors of each such model respect C*. In fact, *for a game in which the ICEU solution is exhaustive*, this condition is the only constraint on priors. In this case, as is evident from the proof of Theorem 6, an ICEU Bayesian model can be constructed using *any* profile of priors that respects C*. In this sense, *the set of* ICEU Bayesian models of the game is consistent with the ICEU solution.

Applying the conclusions of this Section, we have found that, for games in which the ICEU solution is exhaustive, there is nothing paradoxical in the concept of an ICEU Bayesian model. On the contrary, for such games, the ICEU common-reasoning model may be seen as *justifying* ICEU Bayesian models, in the sense that it illustrates explicit steps of reasoning whereby the players could establish the rational permissibility (resp. impermissibility) of those strategies included in (resp. excluded from) each ICEU Bayesian model. It follows that games which pose genuine paradoxes for the Bayesian approach to modelling common knowledge of ICEU rationality must be ones for which the ICEU solution is *not* exhaustive. We turn to such cases in the next Section.

9. Resolving the paradoxes

In Section 2, we presented two paradoxes, using Games 1 and 2 respectively. The argument of Section 8 implies that, for these games to be genuinely troubling exhibits for the Bayesian approach to modelling CKR, it would have to be the case that the ICEU solutions to these games are not exhaustive. That the ICEU solutions are, in fact, not exhaustive is easy to show. Applied to Game 1, the ICEU procedure deletes *third* at stage 1, accumulates *left* at stage 2, then halts to yield the non-exhaustive ICEU solution $<\{left\}, \{third\}>$. Applied to Game 2, the ICEU procedure accumulates in_1, in_2 and in_3 at stage 1, then halts to yield the non-exhaustive ICEU solution $<\{in_1, in_2, in_3\}, \varnothing>$.

Our concept of an ICEU common-reasoning model provides a foundation for ICEU Bayesian models which extends *only* to games for which the ICEU solution is exhaustive. The fact that Games 1 and 2 lie outside that class is what gives rise to the potential for paradox. However, it is also instructive to delve further. Whenever, for the game under analysis, the ICEU solution C* is *not* exhaustive, it follows from the Corollary to Theorem 5

that one of two cases must hold. The first case is that no ICEU Bayesian model exists. The Tom, Dick and Harry Paradox of Game 2 illustrates this case. The second case is that there exists at least one ICEU Bayesian model M, with inclusion categorisation $C^M \supset^* C^*$. In this case, C^M encodes at least one possibility proposition that is not a theorem of R*. We will say that such a proposition is *ungrounded*. In some games, specific ungrounded propositions are encoded by the inclusion categorisation of *every* ICEU Bayesian model, with the apparent implication that they are implied by the common knowledge and rationality assumptions that are common to all ICEU Bayesian models. The Proving Too Much Paradox of Game 1 illustrates this case.

The concept of a model of reasoning gives us a new perspective from which to interpret these puzzling features of the Bayesian approach. To do so, we must go beyond the formal definition of Bayesian and common-reasoning models to consider matters of interpretation. We take it that a Bayesian model is to be interpreted as a formal representation of what individuals know in some *possible world* – that is, in some world that could *conceivably* exist. It is important to understand that, in saying this, we are using the concept of a possible world in a way that is external to the Bayesian modelling framework: possible worlds are what Bayesian models are models *of*.[23] To express this idea, we fix the game and suppose that there exists a set of possible worlds in which this game is played. For any ICEU Bayesian model M of the game, we suppose that there exists a possible world which M represents. Finally, we interpret common knowledge of ICEU rationality as a putative property of possible worlds.

Now consider the following three postulates:

(P1) For every possible world w: if common knowledge of ICEU rationality holds in w, there is some ICEU Bayesian model M′ such that w is represented by M′.

(P2) There is some possible world w such that common knowledge of ICEU rationality holds in w.

(P3) For every ICEU Bayesian model M of the game: if M has property X, then "X" holds in the possible world represented by M.

P1 and P2 are claims about the set of possible worlds. P1 asserts the generality of the Bayesian modelling strategy in relation to this set: it states that every possible world in which the game is played under conditions of common knowledge of ICEU rationality can be represented by some ICEU Bayesian model. P2 states that the concept of common knowledge of ICEU rationality is coherent (that is, it holds in *some* possible world). P3

relates properties of ICEU Bayesian models to properties of the possible worlds represented by them. Its interpretation will depend on the game but, in general, the idea is that X is some property that an ICEU Bayesian model of the game might have and "X" is a corresponding property that might hold in a given possible world.

We first consider Game 1 and the associated Proving Too Much paradox. In the case of Game 1, we interpret X as the property that the inclusion categorisation is <{*first*, *left*, *right*}, {*second*, *third*}> and "X" as the corresponding property in the context of possible worlds, that is, common knowledge of the possibility of *first*, *left* and *right*, and of the impossibility of *second* and *third*. Then, by Proposition 2, the following is a true theorem about the game:

(T1) For every ICEU Bayesian model M: M has property X.

Note that P1, P3 and T1 jointly imply:

(P4) For every possible world w: if common knowledge of ICEU rationality holds in w, then "X" holds in w.

P4 attributes a particular material implication to *every* possible world. The paradox can be stated as follows. It seems that P4 can be true only if *either* "X" holds in all possible worlds *or* there is no possible world in which common knowledge of ICEU rationality holds *or* "X" is derivable from the assumption of common knowledge of ICEU rationality itself, together with any necessary truths, since those are the only features that are common to *all* possible worlds in which common knowledge of ICEU rationality holds. Yet, it is clear that there are possible worlds in which "X" does not hold (for example, worlds in which the game is played by irrational players); and the supposition that there is no possible world in which common knowledge of ICEU rationality holds is the negation of P2. Thus, when T1 is a true theorem, P1, P2, and P3 jointly imply that "X" can be derived from the assumption of common knowledge of ICEU rationality itself (together with necessary truths).

However, in our view, it cannot be so derived. Rather, the appearance that it can be rests on a so far unsupported presumption that P1, P2 and P3 all hold, as applied to Game 1. It seems that, to escape the paradox, we must reject at least one of those three postulates.

The choices become starker when we consider the Tom, Dick and Harry Paradox of Game 2. By Proposition 3, the following is a true theorem about Game 2.

(T2) No ICEU Bayesian model of the game exists.

The paradox is that P1, P2 and T2 are logically inconsistent. Since T2 is a theorem, the inconsistency can be resolved *only* by denying at least one of P1 or P2. In other words, we *must* deny either the generality of the Bayesian modelling strategy *or* the coherence of the concept of common knowledge of ICEU rationality.

There may be readers who are inclined to respond to a choice between these options by denying the coherence of the concept of ICEU rationality. But, we submit that our analysis in earlier sections provides strong arguments in favour of P2 and, for the case where the ICEU solution of the game is not exhaustive, against P1.

Theorem 3 establishes a precise sense in which, for any internally consistent standard of rationality, including ICEU, the concept of CKR can be represented in a common-reasoning model that is consistent. As explained in Section 7, this provides a coherent rendition of the concept of common knowledge of ICEU rationality, for any game; and so provides support for P2.

By Theorem 4, when the ICEU solution C* of the game is not exhaustive, there must be at least one strategy in $\mathbb{S}$, say s_j for player j, such that neither its possibility nor its impossibility is a theorem of R* (where R* is common reason in the ICEU common-reasoning model). The fact that we have presented a formal model of such a case supports the intuition that there is no incoherence in conceiving a possible world in which the ICEU standard of rationality is common knowledge and yet, for every strategy (such as s_j) which is in not in either component of C*, neither its possibility nor its impossibility is common knowledge. Such a possible world cannot be represented by an ICEU Bayesian model. Any ICEU Bayesian model must *either* include s_j, with the ungrounded implication that the possibility of that strategy is common knowledge, *or* exclude it, with the ungrounded implication that its impossibility is common knowledge. We conclude that P1 is not compelling, when the ICEU solution is not exhaustive.

This argument can be illustrated using Game 2, whose ICEU solution is the non-exhaustive categorisation $<\{in_1, in_2, in_3\}, \varnothing>$. For each player i = 1, 2, 3, neither the possibility nor the impossibility of out_i is a theorem of R*, in the ICEU common-reasoning model, but the permissibility of in_i is such a theorem. Thus, consistently with common reason, each player i may attach any non-zero probability to each of the strategies in_j and in_k of his co-players; depending on these probabilities, he may strictly prefer in_i to out_i, or be indifferent between the two. The existence of our common-reasoning model suggests that these conclusions are coherent but, as there is no ICEU Bayesian model of the game, they

cannot be expressed in that framework. Thus, the game seems to identify a limitation of the Bayesian modelling approach, as formulated in Section 2.

Our argument that, if C* is not exhaustive, P1 is not compelling applies generally. It is not conditional on the non-existence of an ICEU Bayesian model, but only on the non-exhaustiveness of the ICEU solution. Thus, it applies also to Game 1, whose ICEU solution is the non-exhaustive categorisation <{*left*}, {*third*}>. If P1 is rejected in relation to Game 1, then the case for P4 disappears. We submit that, given the interpretations we have put on X and "X" for Game 1, P4 is false and P2 holds. In Game 1, neither the possibility nor the impossibility of any of *first*, *second* or *right* is a theorem of R*. Consistently with common reason, player 1 may attach any non-zero probability to *left*; depending on the value of this probability, he may strictly prefer *first* to *second*, strictly prefer *second* to *first*, or be indifferent between the two. Player 2 may attach any probabilities to *first* and *second*, provided that they sum to 1; depending on these probabilities, she may strictly prefer *left* to *right*, or be indifferent between the two. Thus, we have described a family of possible worlds in which common knowledge of ICEU rationality holds but "X" does not, so contravening P4 and supporting P2.

Our analysis suggests that the Proving Too Much paradox and the Tom, Dick and Harry paradox both stem from the same source, namely the presumption that, for any game, P1 is true. Our diagnosis of both paradoxes is the same: it is that, when the ICEU solution of the game is not exhaustive, this presumption is unwarranted.

We can also solve the more general puzzle of why, within the Bayesian modelling approach formulated in Section 2, CKR is a coherent concept for some internally coherent conceptions of rationality (for example, SEU-maximisation without caution) but not for others (for example, ICEU). The source of the problem is that the Bayesian approach to the modelling of CKR has a *general* limitation, irrespective of the conception of rationality that is taken to be common knowledge. Since Bayesian models induce exhaustive categorisations of $\mathbb{S}$, no such model can represent a possible world in which some strategy has the property that neither its possibility nor its impossibility is common knowledge. For some conceptions of rationality and for some games, the effect of this limitation is that none of the possible worlds in which CKR holds can be given Bayesian representations.

The Bayesian approach rests on the implicit assumption that, by some unmodelled process of reasoning, players are able to arrive at common knowledge of a binary partition of the set of strategies into the possible and the impossible. Our analysis of common-reasoning

models shows that this assumption is not justified in general. We conclude that, in the investigation of the implications of rationality and common knowledge in games, there is no substitute for an explicit analysis of reasoning itself.

Appendix 1: Comparison of the ICEU procedure with existing deletion procedures

We begin by comparing the ICEU procedure to the following well-known operations which also iteratively delete strategies: iterative deletion of strictly dominated strategies (IDSDS), iterative deletion of weakly dominated strategies (IDWDS), maximal iterative deletion of weakly dominated strategies (MIDWDS), and the operation (DF) introduced by Dekel and Fudenberg (1990), which has elements of both IDSDS and MIDWDS. Throughout, we rely on the standard definitions of dominance in which a pure strategy is dominated if there exists either another pure strategy or a mixed strategy which dominates it. In terms explained in Section 3, a notable difference between the ICEU procedure and each of these operations is that they *only* delete strategies, whereas the ICEU procedure deletes *and* accumulates. It follows trivially that the ICEU procedure is different in respect of accumulations, but this leaves open the comparison in terms of deletions on which we now report.

IDSDS: It is well known that IDSDS has the attractive properties of insensitivity to order of deletion and the absence of a re-entry problem.[24] As we have shown, the ICEU procedure has these same properties. However, the ICEU procedure deletes more than IDSDS. For any game in G, any strategy that is deleted by IDSDS is also deleted by the ICEU procedure.[25] But, the converse does not hold (see, for example, Game 3).

IDWDS: In Game 4, there are strategies that IDWDS deletes under some orders of deletion but which the ICEU procedure does not delete. Although the order of deletion by IDWDS is ambiguous in that game, the possibility that IDWDS deletes a strategy that the ICEU procedure does not delete does not depend on such ambiguity, as Game 5 shows.

Game 5:

		Player 2	
		left	*right*
Player 1	*first*	1,1	1,0
	second	1,0	0,1

Here, the order of deletion under IDWDS is unique: *second* is deleted first; then, *right*; leaving *first* and *left* undeleted. But, there is still a re-entry problem, as deletion of *right* undercuts the reason for deletion of *second*. In contrast, the ICEU procedure accumulates *first*, but does not delete any strategy. (The rationale is that deletion of *second* is not required unless *right* is accumulated. But, *right* is never accumulated as it is not a best reply to *first,* which is accumulated at the first stage).

Whether every strategy deleted by the ICEU procedure can be deleted by IDWDS depends on the number of players, because the independence condition on beliefs used in the ICEU procedure only bites when n > 2.

For n > 2, there are games where the ICEU procedure deletes strategies that are not deleted by IDWDS under any sequence of deletions. As this possibility arises from the independence component of the concept of IC-consistency, it parallels the disanology between IDSDS and the original definitions of rationalisability (Bernheim, 1984; Pearce, 1984) for games with n > 2.

For any 2-player game, the deletions made by the ICEU procedure can always be made by IDWDS under *some* sequence of deletion. Specifically, deletions made in the same order as they are under the ICEU procedure are always consistent with IDWDS.[26] Thus, the following holds for 2-player games: the sequence of deletions made by the ICEU procedure coincides with *one possible* sequence of IDWDS *up to the stage at which the ICEU procedure halts.* But there may be no sequence of IDWDS that stops deleting when the ICEU procedure does (and there may be other, quite different, sequences of IDWDS).

MIDWDS: This is a variant of IDWDS in which the order of deletion is fixed: at each stage k, MIDWS deletes *every* strategy that is weakly dominated in the game obtained from the original by removing all strategies deleted at previous stages. Although each can be seen as a solution to the order-sensitivity problem of IDWDS, there are marked differences between MIDWDS and the ICEU procedure. Consider Game 6:

Game 6

		Player 2		
		left	*centre*	*right*
Player 1	*first*	1,1	1,1	0,0
	second	1,1	0,1	1,0
	third	0,1	0,0	2,0

In Game 6, MIDWDS deletes *centre* and *right*, then *third*, leaving *first*, *second*, and *left* undeleted. The ICEU procedure accumulates *left* and deletes *right*; then deletes *third*; then accumulates *centre*, and finally deletes *second*. Thus, *second* is deleted by the ICEU procedure, but not by MIDWDS; whereas, *centre* is deleted by MIDWDS, but not by the ICEU procedure. In fact, the ICEU procedure does not just not delete *centre*; it actually accumulates it! Note how the deletions undertaken by MIDWDS give rise to a re-entry problem, posed by the question of why, if *first* and *second* are rationally playable and *third* is not, is *centre* not rationally playable? The ICEU procedure avoids this. Its accumulation of

centre is justified by *centre* being optimal against all strategies except *third*, which is deleted by iterative deletion of *strictly* dominated strategies; and its deletion of *second* is justified by it not being optimal, given the deletion of *right*, once strictly positive probability on *centre* is required.

DF: In DF, there is one stage of maximal deletion of weakly dominated strategies, followed by IDSDS on the game that remains. In some games, DF deletes strategies that the ICEU procedure does not; for example, this is true of *second* in Game 5. In other games, the reverse is true. For example, in Game 6, *second* is deleted by the ICEU procedure but not by DF.

Finally, each of the operations with which we have just compared the ICEU procedure starts, for each player i, with S_i and proceeds to delete strategies from it. In contrast, Asheim and Dufwenberg (2003) present an operation of *iterative deletion of choice sets* which, for each player i, starts with the set of non-empty subsets of S_i and then deletes elements from that set, to obtain a set of undeleted choice sets for player i. As it has no analogue of accumulation, the Asheim-Dufwenberg operation specifies, for each player i, a set of binary partitions of S_i; whereas the ICEU procedure specifies a unique, but trinary, partition.[27]

Appendix 2: Proofs

Proof of Proposition 1: For any game in G, let ρ: $S \rightarrow [0, 1]$ be a probability distribution over the set S of strategy profiles. The probability distribution 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 $\mathtt{I}$ such that, for each player i, for each strategy $s_i \in S_i^*$: $E(s_i) \in \mathtt{I}_i$. Define a prior π^* such that, for each $s \in S^*$: $\pi^*(E[s]) = \rho^*(s)$; notice that this implies $\pi^*(\omega) > 0$ for all $\omega \in \Omega$. Define the profile π of priors such that, for each player i: $\pi_i = \pi^*$. Define the profile χ of choice 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 $\mathtt{I}_i$ and π_i. By construction, the Bayesian model $<\Omega, b(.), \mathtt{I}, \pi, \chi>$ 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. □

Ingredients: For results concerning ICEU Bayesian models, it is convenient to begin by establishing terminology and a lemma used in several subsequent proofs. For any game in G, consider a Bayesian model of the game in which the profile of priors is $\pi = (\pi_1, ..., \pi_n)$. For any distinct players i and j, and for any $s_j \in S_j$, i's *marginal prior* on the event $E(s_j)$ is given by $\pi_i[E(s_j)]$. A strategy s_i for player i is *expected utility maximising with respect to products of marginal priors* if it maximises the expected value of $u_i(s)$ under the assumption that, for each $s_{-i} \in S_{-i}$, the probability of s_{-i} is the product of i's marginal priors on the strategies comprising s_{-i}. For the case of Bayesian models satisfying Independence, we formalise an equivalent conception of expected utility maximisation as follows: Consider 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 $\mathtt{I}_i$. 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. Given that π_i satisfies Independence, we will say that $s_i \in S_i$ is *marginally EU-maximising*, if, for all $s_i' \in S_i$, $U_i(s_i|\ \Omega) \geq U_i(s_i'|\ \Omega)$.

> *Lemma 1*: For any game in G, for any ICEU Bayesian model of that game, for any player i, for any strategy $s_i \in S_i$: $s_i \in S_i^*$ if, and only if, s_i is marginally EU-maximising.

Proof: Consider any game in G, any ICEU Bayesian model of that game, any player i, and any strategy $s_i \in S_i$. To prove the "if" component of the lemma, suppose $s_i \in S_i^*$. By 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 $E \in \mathtt{I}_i$. 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, 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 marginally EU-maximising. To prove the "only if" component, suppose that s_i is marginally EU-maximising, but $s_i \notin S_i^*$. 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 elements of $\mathtt{I}_i$. By Independence, and the fact that s_i is marginally 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 $E' \in \mathtt{I}_i$ 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. By Privacy, it cannot be the case that, at any such state ω, some player $j \neq i$ knows that s_i will not be played. Thus, $s_i \in S_i^*$, contradicting the original supposition. □

Proof of Proposition 2. Consider any ICEU Bayesian model of Game 1. For player 1, *third* is not EU-maximizing with respect to any probability distribution over player 2's strategies. Thus, by Lemma 1, *third* $\notin S_1^*$. Suppose (this is *Supposition 1*) that *second* $\in S_1^*$ and *right* $\in S_2^*$. This implies that E(*second*) and E(*right*) both have strictly positive marginal prior probability. Then *right* is not marginally EU-maximising, and so by Lemma 1, *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 marginally EU-maximising, and so by Lemma 1, *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 marginally EU-maximising and hence, by Lemma 1, $S_2^* = \{$*left*, *right*$\}$. □

Proof of Proposition 3. Suppose there exists an ICEU Bayesian model of Game 2. 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 that E(out_2) has strictly positive marginal prior probability for player 1, and hence that out_1 is not marginally EU-maximising. By Lemma 1, $out_1 \notin S_1^*$, a contradiction. So Supposition 1 is false. Since there are three players, this entails that there are two distinct players i, j such that $out_i \notin S_i^*$ and $out_j \notin S_j^*$. Without loss of generality, set i = 1 and j = 2. Then out_1 is marginally EU-maximising and so, by Lemma 1, $out_1 \in S_i^*$, a contradiction. □

Proof of Theorem 1: Fix any game in G and any aggregate categorisation function ζ for the game. Let f be the profile of categorisation functions that ζ summarises. (If follows from the definition of an aggregate categorisation function that there is exactly one such f.) Let C(0), C(1), C(2), ... be the sequence of categorisations generated by the categorisation procedure. We will say that this procedure has the property of *weak expansion* at stage k if $C(k) \supseteq^* C(k-1)$. Since each categorisation function comprising the profile f satisfies Monotonicity, the continuation rule of the procedure implies that, if the weak expansion property holds at any stage $k' \geq 1$, it also holds at stage $k' + 1$. Since $C(0) = <\varnothing, \varnothing>$, the property holds at stage 1. Thus, by induction, it holds at every stage $k \geq 1$. This establishes part (i) of the theorem. To establish part (ii), note that since weak expansion holds at every stage $k \geq 1$, one of the following must hold: *either* ('Possibility 1') at every stage $k \geq 1$, $C(k) \supset^* C(k-1)$, *or* ('Possibility 2') there is some stage $k' \geq 1$ such that $C(k') = C(k'-1)$. Suppose Possibility 1 is the case. Then, for all $k \geq 1$, $|\mathbb{S}^+(k)| + |\mathbb{S}^-(k)| \geq |\mathbb{S}^+(k-1)| + |\mathbb{S}^-(k-1)| + 1$. Thus, $|\mathbb{S}^+(k)| + |\mathbb{S}^-(k)| \rightarrow \infty$ as $k \rightarrow \infty$. But, by the definition of a categorisation, $|\mathbb{S}^+(k)| + |\mathbb{S}^-(k)| \leq |S_i| + \ldots + |S_n|$. Since the game is finite, this implies a finite upper bound to $|\mathbb{S}^+(k)| + |\mathbb{S}^-(k)|$: a contradiction. Thus, Possibility 2 is the case, so that the procedure halts at stage k*, with k* equal to the lowest value of k' at which the equality defining Possibility 2 holds. □

Proof of Theorem 2. Consider any interactive reasoning system $<P_0, R^*, (R_1, \ldots, R_n)>$ among the population N. Suppose that, for some $p \in \varphi(P_0)$, R*(p) holds. 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:

(L1)	$R^*(p)$	(by supposition)
(L2)	for all $i \in N$: $R_i[R^*(p)]$	(from (L1), using Awareness)
(L3)	for all $i \in N$: $R_i(p)$	(from (L2), using Authority)
(L4)	for all $j \in N$: $R^*[R_j(p)]$	(from (L1), using Attribution)

(L5) for all i, j ∈ N: $R_i[R^*(R_j[p])]$ (from (L4), using Awareness)
(L6) for all i, j ∈ N: $R_i[R_j(p)]$ (from (L5), using Authority)
(L7) for all i, j ∈ N: $R^*[R_i(R_j[p])]$ (from (L4), using Attribution)
… and so on, indefinitely.

The role played by p in (L1), (L2), (L3) is played by $R_j(p)$ in (L4), (L5), (L6), by $R_i[R_j(p)]$ in (L7), (L8), (L9), … and so on. (L3), (L6), (L9), … establish that there is iterated reason to believe p in N. □

Proof of Proposition 4. Consider any game in G and any player i. Fix any decision rule D_i for i, let F_i be the set of maxims which D_i asserts and let f_i be the function that encodes D_i. It is sufficient to show that f_i satisfies Monotonicity. Consider any categorisations C_{-i}', $C_{-i}'' \in \Phi(\mathbb{S}_{-i})$ such that $C_{-i}'' \supset^* C_{-i}'$. We need to show that $f_i(C_{-i}'') \supseteq^* f_i(C_{-i}')$. Suppose (Case 1) that F_i contains no maxim whose antecedent is encoded by C_{-i}'. Then, $f_i(C_{-i}') = <\varnothing, \varnothing>$, trivially implying that $f_i(C_{-i}'') \supseteq^* f_i(C_{-i}')$. Now suppose (Case 2) that F_i contains a maxim $x_{-i}' \Rightarrow y_i'$ such that C_{-i}' encodes x_{-i}' but y_i' is null. As in Case 1, $f_i(C_{-i}') = <\varnothing, \varnothing>$, implying $f_i(C_{-i}'') \supseteq^* f_i(C_{-i}')$. Now consider the only remaining possibility (Case 3): F_i contains a maxim $x_{-i}' \Rightarrow y_i'$ such that C_{-i}' encodes x_{-i}' and y_i' is non-null. As $C_{-i}'' \supset^* C_{-i}'$, every collective prediction that is encoded by C_{-i}'' logically entails y_i'. Thus, by Deductive Closure, F_i contains a maxim $x_{-i}" \Rightarrow y_i"$, such that $x_{-i}"$ is encoded by C_{-i}'' and $y_i"$ logically entails y_i'. By construction, $f_i(C_{-i}') = y_i'$ and $f_i(C_{-i}'') = y_i"$. Since $y_i"$ logically entails y_i', we must have $f_i(C_{-i}'') \supseteq^* f_i(C_{-i}')$. □

Proof of Theorems 3 and 4: Consider any game in G, and any profile D* of decision rules for its players. Let $< P_0, R^*, (R_1, \ldots, R_n)>$ be the common-reasoning model of the game, defined in relation to D*. By Proposition 4, 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 guaranteed by Theorem 1. For each i, let $S_i^{+}* = \mathbb{S}^{+}* \cap S_i$ and $S_i^{-}* = \mathbb{S}^{+}* \cap S_i$.

We now define an inference structure $R_{-}*$ which has the same domain and axioms as R* but whose inference rules are, in a sense to be defined, "weaker" than those of R*. We define the following sets of inference rules. I_1 consists of the rules of 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 ∈ N. I_3 is the set of inference rules of the form "from $\{R_i(y_i)\}$, infer z_i", where i ∈ 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 ∈ 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 component maxim of D_i^*, and y_i is the consequent of that maxim.

Notice that R* is fully specified by its domain $\varphi(P_0)$, axiom set A(R*) and by the condition that it has the inference rules contained in $I_1 \cup I_2 \cup I_3$. We define R_-^* as the inference structure that is fully specified by having the domain $\varphi(P_0)$, the axiom set $A(R_-^*) = A(R^*)$ and the inference rules contained in $I_4 \cup I_5 \cup I_6$. Note that this implies that R_-^* does not have all rules of valid inference. The proof uses the following lemmas:

> *Lemma 2*: For each $i \in N$ and for each $s_i \in S_i$: (i) $s_i \in S_i^{+}*$ if, and only if, $p_i(s_i)$ is asserted by some theorem in $T(R_-^*)$; and (ii) $s_i \in S_i^{-}*$ if, and only if, $\neg p_i(s_i)$ is asserted by some theorem in $T(R_-^*)$.

Proof: For the purposes of this proof, we extend the definitions of "encoding", given in Section 6, to allow sets of permissibility propositions for any player i, for any set of players $N\backslash\{i\}$, and for the entire set of players N, to be encoded by categorisations of S_i, $\mathbb{S}_{-i}$ and $\mathbb{S}$ respectively. In each case, a strategy is assigned to the positive (resp. negative) component of the encoding categorisation if, and only if, its permissibility (resp. impermissibility) is asserted by some permissibility proposition in the relevant encoded set. All sets to which we apply the concept below have the properties necessary to permit encoding by a categorisation in this way.

It is convenient to group the steps of reasoning by which theorems can be derived in the inference structure R_-^* into "phases" of three, which successively use the inference rules in I_4, I_5 and I_6. We treat "phase 0" as generating the set of axioms $A(R_-^*) = \{\#, D_1^*, ..., D_n^*\}$. The only permissibility proposition in $A(R_-^*)$ is #, which is encoded by the categorisation $<\varnothing, \varnothing>$. We denote this categorisation C(0) to signify that it encodes permissibility propositions proved at the end of phase 0.

The only inference rules of R_-^* that can use subsets of $A(R_-^*)$ as premises and generate conclusions that are not themselves elements of $A(R_-^*)$ are those in I_6. Thus, the first active step in deriving any such theorem must use inference rules of the form "from $\{D_i^*, \#\}$, infer y_i", where $\# \Rightarrow y_i$ is a maxim of D_i^*; by this step, theorems of the form y_i, i.e. recommendations, may be derived. For consistency with later phases, we call this "step 1.3". Let $T_{1.3}(R_-^*)$ be the set which contains the axioms of R_-^* and all theorems that can be

proved using step 1.3. This is the end of phase 1. From an examination of this reasoning, it is evident that the permissibility propositions asserted by theorems in $T_{1.3}(R\_^*)$ are encoded by the categorisation $C(1) = \zeta[C(0)]$.

The only inference rules of R_* which can use subsets of $T_{1.3}(R\_^*)$ as premises and generate conclusions that are not themselves elements of $T_{1.3}(R\_^*)$ are those in I_4. Thus, step 2.1 (the first step of phase 2) must use inference rules of the form "from $\{y_i\}$, infer z_i", where $\{y_i\} \subseteq T_{1.3}(R\_^*)$; by this step, propositions of the form z_i, i.e. predictions, may be derived. Let $T_{2.1}(R\_^*)$ be the set which contains the axioms of R_* and all theorems that can be proved using steps 1.3 and 2.1. The only inference rules which can use subsets of $T_{2.1}(R\_^*)$ as premises and generate conclusions that are not themselves elements of $T_{2.1}(R\_^*)$ are those in I_5. Thus, step 2.2 must use inferences rule 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 $\{z_1, ..., z_{i-1}, z_{i+1}, ... z_n\} \subseteq T_{2.1}(R\_^*)$; by this step, propositions of the form $z_1 \wedge ... \wedge z_{i-1} \wedge z_{i+1} \wedge ... \wedge z_n$, i.e. collective predictions, may be derived. Let $T_{2.2}(R\_^*)$ be the set which contains the axioms of R_* and all theorems that can be proved using steps 1.3 to 2.2. Step 2.3 follows the model of step 1.3, using inference rules in I_6; this leads to $T_{2.3}(R\_^*)$ which contains all theorems that can be proved using steps 1.3 to 2.3. This completes phase 2. An examination of these steps of reasoning shows that the permissibility propositions asserted by theorems in $T_{1.3}(R\_^*)$ are encoded by the categorisation $C(2) = \zeta[C(1)]$.

Phases 3, 4, ... follow the same pattern. By Theorem 1, there is some finite k* such that $C(k^*) = C(k^*-1) \supset^* ... \supset^* C(1) \supset^* C(0)$. Thus, $T_{k^*.3}(R\_^*) = T_{(k^*-1).3}(R\_^*) \supset ... \supset T_{1.3}(R\_^*) \supset A(R\_^*)$. This implies that no new conclusions can be derived from $T_{k^*.3}(R\_^*)$ by using any of the inference rules of R_*. Thus, the categorisation solution C(k*) encodes all (and only) those permissibility propositions that are asserted by theorems of R_*.

Lemma 3: $T(R\_^*) \subseteq T(R^*)$.

Proof: By construction, R_* has the set of inference rules $I_4 \cup I_5 \cup I_6$, while R* has $I_1 \cup I_2 \cup I_3$. Note that $I_5 \cup I_6 \subseteq I_1$. Every inference that can be made using I_4 can also be made using I_2 followed by I_3. Thus, every inference that can be made in R_* can also be made in R*. Since both inference structures have the same axiom set, every theorem of R_* is also a theorem of R*.

Lemma 4: T(R_*) is consistent.

Proof: 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: $T^1 = A(R_{-}^*)$; $T^2 = \{p \in T(R_{-}^*) \mid p$ 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 2 implies that, for each player i, the set of strategies for i whose permissibility (resp. impermissibility) is asserted 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 $\mathbb{S}$, the construct obtained by removing from each of its components all strategies of players in $N\backslash\{i\}$ is a categorisation of S_i. It follows from the definition of a categorisation that the subset of T^3 containing recommendations to i is consistent. As this holds for each i, T^3 is consistent. Since each element of T^2 is either the correlate of some element of T^3 or a conjunction of a set of such correlates, and since T^3 is consistent, T^2 is consistent. Thus, since every proposition in T^2 is a conjunction of possibility propositions, whereas every proposition in T^3 is a recommendation, $T^2 \cup T^3$ 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. Finally, by the specification of I_6, every proposition in T^3 is logically entailed by $T^1 \cup T^2$. Thus, the union of the consistent sets T^1 and $T^2 \cup T^3$ is consistent.

Lemma 5: $T(R^*)$ is consistent.

Proof: By Lemma 4, $T(R_{-}^*)$ 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 propositions 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 $T(R^*)$.

Lemma 6: For each $i \in N$ and for each $s_i \in S_i$: $R^*[p_i(s_i)] \Rightarrow R_{-}^*[p_i(s_i)]$ and $R^*[\neg p_i(s_i)] \Rightarrow R_{-}^*[\neg p_i(s_i)]$.

Proof: By Lemma 5, $T(R^*)$ is consistent. Since all decision rules satisfy Deductive Closure, the specification of I_5 and I_6 ensures that each permissibility proposition that is derivable in

R* is a component of a recommendation that is derivable in R_* (compare the proof of Lemma 5).

Lemma 7: If T(R*) is consistent, then, for each $i \in N$, $T(R_i)$ is consistent.

Proof: Suppose T(R*) is consistent. Consider any $i \in N$. It follows from the definition of the common-reasoning model, and specifically from the use of rules (3) and (4), that $T(R_i)$ can be partitioned into the subsets T^1, T^2 and T^3, defined as follows: $T^1 = \{\#\} \cup \{p \in \varphi(P_0) \mid p = R^*(q)$ 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 # and propositions of the form R*(p), while T^2 is a consistent set which contains no proposition prefaced by $\neg R^*$, $T^1 \cup T^2$ is consistent. Since T^3 contains only propositions that are logically entailed by $T^1 \cup T^2$, $T^1 \cup T^2 \cup T^3$ is consistent.

Lemmas 5 and 7 together prove Theorem 3. Lemma 2 establishes that the positive (resp. negative) component of the categorisation solution contains those (and only those) strategies whose permissibility (resp. impermissibility) is asserted by theorems of R_*. Lemmas 3 and 6 together establish that strategies are asserted to be permissible (resp. impermissible) by theorems of R_* if, and only if, they are asserted to be permissible (resp. impermissible) by theorems of R*. These results prove Theorem 4. □

Proof of Proposition 5: Consider any profile $D^* = (D_1^*, \ldots, D_n^*)$ of decision rules for any game in G and let R* be common reason in the common-reasoning model in which D* is the common standard of practical rationality. Let ζ be the aggregate categorisation function that encodes D*. Consider any categorisation C of $\mathbb{S}$ that encodes a profile of recommendations $(y_1, \ldots, y_n)$ such that, for each player i, $y_i \in T(R^*)$. From the definitions of encoding, C also encodes the profile $(z_1, \ldots, z_n)$ such that, for each i, z_i is the correlate prediction of recommendation y_i. For each i, since $y_i \in T(R^*)$, rules (2)(i) and (2)(ii) for the construction of R* guarantee that $z_i \in T(R^*)$, so establishing part (i) of the Proposition. For each player j, define $x_{-j} = z_1 \wedge \ldots \wedge z_{j-1} \wedge z_{j+1} \wedge \ldots \wedge z_n$. As R* has the rules of valid inference, $x_{-j} \in T(R^*)$. Because of rule (1) for the construction of R*, $D_j^* \in T(R^*)$. By Distinct Antecedents, D_j^* asserts at most one maxim whose antecedent is logically equivalent to x_{-j}. If such a maxim exists, let y_j' be its consequent. If no such maxim exists, let $y_j' = \#$. Note that $x_{-j} \Rightarrow \#$ is a tautology. In either case, because R* has the rules of valid inference, $y_j' \in T(R^*)$. Since ζ encodes D*, the profile $(y_1', \ldots, y_n')$ so obtained is encoded by $\zeta(C)$. This establishes part (ii) of the Proposition. □

Proof of Theorem 5: To simplify exposition, if a probability distribution over S_{-i} for some player i is IC-consistent with a categorisation $C_{-i} = \langle \mathbb{S}_{-i}^{+}, \mathbb{S}_{-i}^{-} \rangle$ of $\mathbb{S}_{-i}$, we will say that it is

also IC-consistent with every categorisation $C = \langle \mathbb{S}^+, \mathbb{S}^- \rangle$ of $\mathbb{S}$ such that $\mathbb{S}_{-i}^{\ +} = \mathbb{S}^+ \backslash S_i$ and $\mathbb{S}_{-i}^{\ -} = \mathbb{S}^- \backslash S_i$. Consider any game in G, any ICEU Bayesian model M of that game, and any player i. Let C^M be the inclusion categorisation of that model. Let ζ be the ICEU aggregate categorisation function for the game. By Lemma 1, if some strategy $s_i \in S_i$ is in the positive component of C^M, it is marginally EU-maximising for some probability distribution over S_{-i} that is IC-consistent with C^M; if it is in the negative component of C^M, there is some such distribution for which it is *not* marginally 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 IC-consistent with C^M is also IC-consistent with C (this is *Result 2*). By the definition of the ICEU categorisation function for player i, if some strategy $s_i \in S_i$ is in the positive component of $\zeta(C)$, it is marginally EU-maximising for every probability distribution over S_{-i} that is IC-consistent with C; if it is in the negative component of $\zeta(C)$, it is marginally EU-maximising for no such distribution (this is *Result 3*).

Suppose Theorem 5 is false. Then, using the fact that 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. □

Proof of Theorem 6: Consider any game in G. Let ζ be the ICEU aggregate categorisation function for the game. Suppose there exists an exhaustive categorisation $C' = \langle \mathbb{S}^+, \mathbb{S}^- \rangle$ of $\mathbb{S}$ such that $C' = \zeta(C')$. Let $\langle S_i^{\ +}, S_i^{\ -} \rangle$ and $\langle \mathbb{S}_{-i}^{\ +}, \mathbb{S}_{-i}^{\ -} \rangle$ be the corresponding categorisations of S_i and $\mathbb{S}_{-i}$ (i.e. categorisations satisfying $S_i^{\ +} \cup \mathbb{S}_{-i}^{\ +} = \mathbb{S}^+$ and $S_i^{\ -} \cup \mathbb{S}_{-i}^{\ -} = \mathbb{S}^-$).

We can construct an ICEU Bayesian model $M = \langle \Omega, b(.), \mathtt{I}, \pi, \chi \rangle$ of the game as follows: Set $S^* = S_1^* \times ... \times S_n^*$, with $S_i^* = S_i^{\ +}$ for each player i. 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. Define the behaviour function b(.) so that $b(\omega[s]) = s$. Define the information structure $\mathtt{I}$ such that, for each player i, for each strategy $s_i \in S_i^*$: $E(s_i) \in \mathtt{I}_i$. For each player i, fix any independent prior π_i such that, for each strategy s_j for each player $j \neq i$, the marginal probability of s_j is strictly positive if, and only if, $s_j \in S_j^*$. Define χ so that, for each player i, at every state ω, $\chi_i(\omega) = S_i^*$. By construction, M satisfies Independence, Knowledge of Own Choice and Privacy, and its inclusion categorisation is C'. Consider any player i and any strategy $s_i \in S_i^*$. Since $S_i^* = S_i^{\ +}$, it follows from the definition of the ICEU aggregate categorisation function that s_i is marginally EU-maximising with respect to all

probability distributions over S_{-i} which assign strictly positive probability to strategies in $\mathbb{S}_{-i}^{+}$ and zero probability to strategies in $\mathbb{S}_{-i}^{-}$. Hence, s_i is marginally EU-maximising with respect to π_i. Because π_i is independent, and because of the specification of $\mathtt{I}_i$, 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 $\omega \in \Omega$. Thus, at each state $\omega \in \Omega$, the set of strategies that are SEU-rational for i is S_i^*. Thus, the specification $\chi_i(\omega) = S_i^*$ ensures that Choice Rationality and SEU-Maximisation are satisfied. □

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 (1998). Common priors: a reply to Gul. *Econometrica* 66, 929-38.

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.

Bernheim, B. Douglas (1984). Rationalizable strategic behavior. *Econometrica* 52, 1007-1028.

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.

Borgers, 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.

Chen, Yi-Chun, Ngo Van Long, and Xiao Luo (2007). Iterated strict dominance in general games. *Games and Economic Behavior*, 61, 299-315.

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

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.

Dekel, Eddie and Drew Fudenberg (1990). Rational behaviour with payoff uncertainty. *Journal of Economic Theory* 52, 243-67.

Dekel, Eddie and Faruk Gul (1997). Rationality and knowledge in game theory. In D.M. Kreps and K.F. Wallis (eds.) *Advances in economics and econometrics: theory and applications Volume I.* Cambridge, UK: Cambridge University Press.

Dufwenberg, Martin and Mark Stegeman (2002). Existence and uniqueness of maximal reductions under iterated strict dominance. *Econometrica*, 70, 2007-24.

Ewerhart, Christian (2002). Ex-ante justifiable behaviour, common knowledge and iterated admissibility. Mimeo, University of Mannheim.

Gul, Faruk (1998). A comment on Aumann's Bayesian view. *Econometrica* 66, 923-8.

Harsanyi, John C. (1975). The tracing procedure. *International Journal of Game Theory* 4, 61-94.

Harsanyi, John C. and Reinhard Selten (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

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

Mandler, Michael (2007). Strategies as states. *Journal of Economic Theory* 135, 105-30.

Morris, Stephen (1995). The common prior assumption in economic theory. *Economics and Philosophy* 11, 227-54.

Myerson, Roger (1991). *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

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.

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

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

Samuelson, Larry (2004). Modeling knowledge in economic analysis. *Journal of Economic Literature* XLII, 369-403.

Skyrms, Brian (1989). Correlated equilibria and the dynamics of rational deliberation. *Erkenntnis* 31, 347-364.

Skyrms, Brian (1990). *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.

Tan, Tommy C.-C. and Sergio R. da C. Werlang (1988). The Bayesian foundations of solution concepts of games. *Journal of Economic Theory* 45, 370-391.

Notes

[1] Lewis is widely credited with the first formal definition of common knowledge, but it is less well known that this definition is only one component of a detailed analysis of a mechanism by which common knowledge can be generated through interlocking processes of individual reasoning (see Cubitt and Sugden, 2003).

[2] Lewis (1969) and Aumann (1987) are widely regarded as canonical treatments of common knowledge, but their formal definitions of the concept are very different. As Aumann (p. 10) recognises, it is necessary to use "common knowledge" in "its informal, everyday meaning" in order to interpret a formal model of the concept.

[3] The results we have in mind seem to call into question the coherence of CKR, as modelled in the Bayesian framework. They are distinct from implications of CKR that are simply surprising, in failing to correspond with expectations about how, in the real world, games are played by intelligent humans. They are also distinct from variants of Newcomb's problem which may arise in Bayesian models of games (Mandler, 2007).

[4] Bacharach (1987) presents a class of theories of rational play such that, for each theory T in the class, it is an axiom of T that if some proposition p is a theorem of T then it also a theorem of T that the players know p. Thus, although such a theory T is primarily a theory about *what* players know and do, it is also possible to interpret it as providing *the reasoning* through which players come to know p (p. 46).

[5] An analysis that initially seems similar to ours is that of Ewerhart (2002). Like us, Ewerhart distinguishes between truth and derivability within a given formal system; but his approach generates iterative deletion of weakly dominated strategies rather than a trinary partition of the strategy set. In Ewerhart's framework, it is "commonly assumed" among the players that each assigns strictly positive probability to a strategy if, and only if, it is provable that it is possibly chosen. In our model of ICEU reasoning, there is an analogue of the "if" part of this condition, but not of the "only if" part, which we see as abolishing an important distinction between "provably impossible" and "not provably possible".

[6] Throughout, we use the term "profile" of objects of a given type to denote a function which associates with each player $i \in N$ an object of that type that applies to i. For example, a strategy profile associates with each player i an element of S_i.

[7] Player indices are not always necessary. To avoid unnecessary subscripts, we use the convention that, for two-player games, *first*, *second*, …, are strategies for player 1 and *left, centre, right* are strategies for player 2.

[8] We use the connectives $\neg$, $\wedge$, and $\Rightarrow$ for negation, conjunction and material implication, respectively.

[9] The assumption, made by Aumann (1987), that players have common priors has proved controversial. See, for example, Morris (1995), Gul (1998) and, for a response, Aumann (1998).

[10] Note that the test of SEU-rationality of s_i at ω requires that s_i yields at least as high an expected utility as any other strategy in S_i, not just as those in S_i*.

[11] As the structure of our proof makes clear, Aumann's analysis implies a stronger result in which existence of a Bayesian model in which players have a common prior is established for every game in G.

[12] This can be proved by exploiting the existence proof for quasi-strict Nash equilibrium for 2-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 Proposition 1. The properties of quasi-strict Nash equilibrium ensure that Independence and Privacy are satisfied.

[13] We call the inconsistency shown by Game 2 the *Tom, Dick and Harry Paradox* because the game can be illustrated with the following story suggested to us by Michael Bacharach: 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. Each guest is indifferent between all outcomes, provided he does not meet the person he is trying to avoid.

[14] It is possible that cautious expected utility maximization could be represented in an epistemic model in which probabilities are non-Bayesian, for example by extending the analysis of Brandenburger *et al* (2008), which uses lexicographic probability systems in place of priors. However, we will offer a resolution of the paradoxes which does not require abandonment of the usual Bayesian concept of probability.

[15] Throughout, we use $\subset$ to denote 'is a strict subset of'.

[16] Monotonicity is not the *only* formal restriction on categorisation functions that might be justified by appeal to principles of reasoning; but it is sufficient for the results we prove in this section. In Section 6 we present the stronger condition that a categorisation function should "encode" some "decision rule".

[17] To see this, consider any C_{-i}, C_{-i}' such that $C_{-i}' \supset^* C_{-i}$. Since C_{-i}' has more content than C_{-i}, the requirement on a probability distribution over S_{-i} of being IC-consistent with C_{-i}' imposes more constraints than the requirement of being IC-consistent with C_{-i}. Thus, if some $s_i \in S_i$ is expected utility maximising for *every* probability distribution over S_{-i} that is IC-consistent with C_{-i}, it is also expected utility maximising for each such distribution than is IC-consistent with C_{-i}'; and, if some $s_i \in S_i$ is *not* expected utility maximising for

any probability distribution over S_{-i} that is IC-consistent with C_{-i}, it is also not expected utility maximising for any such distribution than is IC-consistent with C_{-i}'.

[18] Thus, $\neg p$ is true if, and only if, p is false; $p \wedge q$ is true if, and only if, both p and q are true; and $p \Rightarrow q$ is true if, and only if, either q is true or p false.

[19] It would be possible to formulate our model without the concept of the null proposition, but only at a cost of unnecessary cumbersomeness in the definitions of the components of a decision rule.

[20] A condition that, for a *given* reasoning scheme R, R(p) implies R(R(p)) would be analogous to a principle known in epistemic logic as *positive introspection*. It could be imposed on our framework by requiring R to have, for each proposition $p \in P$, the inference rule "from {p}, infer R(p)". It would make no substantive difference to our analysis if we were to add this requirement, but in fact we do not do so.

[21] As part of the definition of the correlate of a recommendation (resp: prediction), we require that the order of the component possibility (resp: permissibility) propositions in the correlate matches that of the component permissibility (resp: possibility) propositions in the recommendation (resp: prediction). This requirement has no substantive content, but simplifies the presentation and proof of our results.

[22] A recommendation is a conjunction of the elements of a set of permissibility propositions. The elements of a given set of (two or more) permissibility propositions can be conjoined in different orders to produce distinct but logically equivalent statements. For example, $p_i(s_i') \wedge \neg p_i(s_i'')$ and $\neg p_i(s_i'') \wedge p_i(s_i')$ are distinct recommendations, but share a common encoding.

[23] Thus, the concept of a "possible world" should not be confused with that of a "state", which is internal to the Bayesian modelling framework.

[24] Note that we have formulated our concept of a categorisation procedure only for finite games. The claim that IDSDS always has the attractive properties is also limited to this class of games. For discussion of IDSDS in infinite games, see Dufwenberg and Stegeman (2002) and Chen *et al* (2007).

[25] This follows from the relevant definitions, using Theorem 1.6 of Myerson (1991) to establish the equivalence of being strictly dominated against a given subset of S_{-i} and not being expected utility maximising for any probability distribution over that subset.

[26] To see this, note that (in a 2-player game) the strategies deleted at stage k of the ICEU procedure are not optimal for *any* beliefs about the other player's strategies that assign zero marginal probability to all previously deleted strategies and strictly positive marginal probability to all previously accumulated strategies. It follows that these strategies are also not optimal for any such beliefs that, *additionally*, assign strictly positive marginal probability to all strategies that are neither previously deleted nor previously accumulated. Consequently, by application of Theorem 1.7 of Myerson (1991), they are weakly dominated in the game obtained by removing from the original all strategies deleted at earlier stages.

[27] From the output of the Asheim-Dufwenberg procedure, one can infer, for each player i, an implied trinary partition of S_i. Its elements are the set of strategies for i that are in all subsets of S_i not deleted in the Asheim-Dufwenberg procedure; the set of strategies for i that are in no such subset; and the set of strategies for i that are in some, but not all, such subsets. One might conjecture that these sets would coincide, respectively, with $S_i \cap \mathbb{S}^+(k^*)$, $S_i \cap \mathbb{S}^-(k^*)$, and $S_i \setminus (\mathbb{S}^+(k^*) \cup \mathbb{S}^-(k^*))$, defined by the ICEU solution. But this is not so: Game 1 of Asheim and Dufwenberg (2003) is a counter-example.