CEDEX

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham

Discussion Paper No. 2011-01

Robin P. Cubitt
and
Robert Sugden

January 2011

Common Reasoning in Games: A Lewisian Analysis of Common Knowledge of Rationality

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CEDEX

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

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

Sue Berry
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 15469
Fax: +44 (0) 115 95 14159
sue.berry@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:

a Lewisian analysis of common knowledge of rationality*

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

14th January 2011

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

Email:

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

* We are grateful for comments on earlier versions of the paper to Giacomo Bonnano, Adam Brandenburger, John Collins and participants in various seminars, conferences and workshops at which we have presented the paper.

Abstract

The game-theoretic assumption of 'common knowledge of rationality' leads to paradoxes when rationality is represented in a Bayesian framework as cautious expected utility maximisation with independent beliefs (ICEU). We diagnose and resolve these paradoxes by presenting a new class of formal models of players' reasoning, inspired by David Lewis's account of common knowledge, in which the analogue of common knowledge is *derivability in common reason*. We show that such models can consistently incorporate any of a wide range of standards of decision-theoretic practical rationality. We investigate the implications arising when the standard of decision-theoretic rationality so assumed is ICEU.

Short title

Common reasoning in games

Keywords

Common reasoning; common knowledge; common knowledge of rationality; David Lewis; Bayesian models of games.

1. Introduction

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

Intuitively, CKR seems a meaningful idealisation, in the same sense that perfect competition is a meaningful idealisation in economics or frictionless surfaces are in theoretical mechanics. However, attempts to formalise the assumption have sometimes generated paradoxical implications that appear to call into question the coherence of the concept. In this paper, we offer a diagnosis of these paradoxes and, by presenting a new class of 'common-reasoning models', show how the intuitive idea of CKR can be formulated without creating paradoxes.

Our approach to modelling CKR is inspired by Lewis (1969). Although Lewis is widely credited with the first precise definition of common knowledge, it is less well known among game theorists that this definition is only one component of a detailed analysis of interlocking processes of individual reasoning. Building on an analysis of Lewis's game theory by Cubitt and Sugden (2003), we formalise and extend Lewis's approach to represent how individual players may reason about the standards of *practical* – that is, decision-theoretic – rationality that other players endorse, and in this way reach conclusions about whether specific strategies are or are not rationally playable.

Although lip-service is often paid to Lewis's historic role, the approach to modelling CKR now seen by most game theorists as canonical is that due to Aumann (1987); and it contrasts markedly with that of Lewis.[1] Aumann offers a Bayesian modelling framework which can be used to represent CKR among the players of any given noncooperative game.

[1] Further developments of Aumann's Bayesian approach have been made, for example, by Tan and Werlang (1988), Dekel and Gul (1997) and Aumann (1999a, b). Cubitt and Sugden (2003) is one of several recent analyses of Lewis's argument. It contrasts that approach with that of Aumann. Other formalisations of Lewis's argument, based on different understandings of what is central to it, are offered by Vanderschraaf (1998), Sillari (2005), Gintis (2009) and Paternotte (2010). We discuss them further in Section 4.

Aumann sees this framework as providing formal foundations for a solution concept, correlated equilibrium, which generalises Nash equilibrium. The central assumption of the model is that 'it is common knowledge that all the players are Bayesian utility maximizers' (p. 2), which Aumann treats as synonymous with there being 'common knowledge of rationality' (p.12). Although Aumann's model is logically consistent, apparently natural extensions of it, intended to introduce different conceptions of practical rationality involving principles of 'caution' or weak dominance, turn out to generate puzzles and even contradictions in some games (Borgers and Samuelson, 1992; Samuelson, 1992; Cubitt and Sugden, 1994). Of course, 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 puzzles and contradictions arise are paradoxical and troubling exhibits for the Bayesian approach. In our view, there is also a deeper conception of paradox that does not require the extensions to define uniquely, or even especially, compelling conceptions of practical rationality, but only ones that are 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.

The underlying reason why the Bayesian approach runs into problems, we will argue, is that it seeks to represent a situation in which, in every state of the world, each player's choices are expected utility maximising relative to her beliefs; and this situation is represented as common knowledge. This implies a binary partition of the set of strategies: one element of this partition contains those strategies that are played in *some* state(s) of the world, while the other contains those that are played in *none*. It also implies common knowledge that a strategy in the first element is played. However, the Bayesian approach to modelling CKR does not attempt to describe the modes of reasoning by which the players might discover the partition for themselves.

In contrast, a central feature of a Lewisian approach is that it does describe players' reasoning. In our formulation, we capture the intuitive idea of CKR by assuming that players have access to specific modes of reasoning that constitute the common rationality being modelled and that this common rationality embeds, as one of its components, some standard of practical rationality. Our common-reasoning model, for a given such standard and a given game, gives an explicit representation of reasoning by which players can arrive at conclusions about the rational permissibility or impermissibility of strategies. For a given strategy, there are three possibilities: *either* the permissibility of the strategy can be established by reasoning; *or* its impermissibility can be so established; *or* neither its permissibility nor its impermissibility can be established. Thus, it is intrinsic to our approach that each common-reasoning model generates a trinary partition of the strategy set of a given game. The properties of these trinary partitions and the relationship between them and the binary partitions arising from the corresponding Bayesian models are at the heart of our resolution of the paradoxes faced by the latter. Moreover, we show that the relevant common-reasoning model provides a consistent rendition of CKR, for any game and any coherent standard of practical rationality. By doing so, we 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.

Of course, the Lewisian approach is not the *only* way of modelling 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. In our approach, in contrast, there are no assumptions about priors.

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. We see our work as, in some respects, 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.

One of the merits of the Lewisian approach is that it resolves the paradoxes that arise when the Bayesian approach is combined with principles of weak dominance or caution. We do not claim that these paradoxes can be resolved in no other way, but only that the Lewisian resolution is both natural and general. A different response to these paradoxes, discussed for example by Brandenburger (2007) and Brandenburger *et al* (2008), retains more features of Aumann's model but uses lexicographic probability systems in place of Bayesian probabilities. Whatever the merits of this rendition of CKR, the use of non-Bayesian probabilities is a major departure from Aumann's explicitly Bayesian approach. Thus, the viability of this modelling strategy does not compromise our claim that the Bayesian approach leads to paradoxes; nor does it demonstrate the possibility of modelling CKR in a way that is robust to different conceptions of practical rationality.

The remainder of the paper is organised as follows: Section 2 presents the Bayesian approach to modelling CKR and shows by means of three exhibit games how, though that approach is internally consistent, it gives rise to paradoxes when extended to capture a conception of practical rationality in which rational individuals maximise expected utility in relation to beliefs that are independent and cautious, in the sense of Pearce (1984) and Borgers and Samuelson (1992). By doing so, it motivates the development of our alternative approach to modelling CKR.

Sections 3–5 introduce the major ingredients of our Lewisian approach. Section 6 presents that approach itself, by first defining the class of common-reasoning models and then establishing the consistency of every such model. Section 7 introduces a sense in which a given common-reasoning model defines a 'solution' to the game and defines a 'recommendation algorithm' which can be used to identify that solution, and which is interpretable as tracking specific steps of reasoning that lead 'common reason' to it.

As the primitives of our common-reasoning models are very different from those of the Bayesian models introduced in Section 2, it helps to define a framework within which they can be compared. We present such a framework in Section 8, exploiting concepts introduced in Cubitt and Sugden (2010). Section 9 then specialises the common-reasoning framework to the case where common rationality embodies the conception of practical rationality that Section 2 showed to give rise to paradoxes within the Bayesian approach. Using the framework of Section 8 as a bridge, Section 9 also establishes precise relationships between the corresponding Bayesian and common-reasoning models. These relationships provide the ingredients for a final resolution of the paradoxes, presented in Section 10.

Between them, two appendices provide proofs of all the formal results. They do so in part by exploiting relationships between concepts presented in this paper and the concept of a ‘categorisation procedure’ introduced in Cubitt and Sugden (2010). These relationships are demonstrated in Appendix 1; Appendix 2 gives proofs of all results from the main text, drawing in some cases on preliminaries established in Appendix 1. The results of Appendix 1 are also of independent interest because they demonstrate that the common-reasoning models presented in Section 6 can be interpreted as formal foundations for Cubitt and Sugden (2010)’s categorisation procedures.

2. Common knowledge of rationality in a Bayesian model: three paradoxes

In this section, we present three 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[2] of strategies $s = (s_1, ..., s_n)$, there is a profile $u(s) = (u_1[s], \ldots, u_n[s])$ of real-valued and finite *utilities*. The set $S_1 \times ... \times S_\text{n}$ 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. This convention allows a conveniently compact notation in later sections, in common with that used by Cubitt and Sugden (2010).

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

Uncertainty is represented 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*.

[2] 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.

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 (such as mixed strategies) is represented as choice that is conditioned on random events. For each profile s of strategies and each strategy s_i, we define the events $E(s) = \{\omega \in \Omega \mid b(\omega) = s\}$ and $E(s_i) = \{\omega \in \Omega \mid b_i(\omega) = s_i\}$. Let $S^* = \{s \in S \mid E(s) \neq \varnothing\}$ and $S_i^* = \{s_i \in S_i \mid E(s_i) \neq \varnothing\}$. S^* (respectively S_i^*) is the set of strategy profiles (respectively strategies for i) *included* in the Bayesian model. Thus, a Bayesian model specifies a binary partition of each player's strategy set S_i, the elements of which are the set of included strategies S_i^* and the set of *excluded* strategies $S_i \backslash S_i^*$. By construction, each S_i^* is non-empty.

Players' knowledge is represented by an *information structure* $\mathscr{I} = (\mathscr{I}_1, ..., \mathscr{I}_n)$. For each player i, $\mathscr{I}_i$ is an information partition of Ω, representing what i knows at each state. $K_i(E)$, where E is an event, is the event $\{\omega \in \Omega \mid \exists E' \in \mathscr{I}_i: (\omega \in E') \wedge (E' \subseteq E)\}$.[3] 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 $s_j \in S_j^*$, $s_k \in S_k^*$: $\pi_i(E[s_j] \cap \mathrm{E}[s_k]) = \pi_i(E[s_j])\pi_i(E[s_k])$. A profile $\pi = (\pi_1, ..., \pi_\mathrm{n})$ 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.[4]

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

[3] We use the connectives $\neg$, $\wedge$, and $\Rightarrow$ for negation, conjunction and material implication, respectively. We use $\subset$ (resp. $\subseteq$) to denote 'is a strict (resp. weak) subset of'.

[4] Common support is a much weaker condition than that of common priors (i.e. that, for all distinct players i and j, $\pi_i = \pi_j$). The latter assumption, made by Aumann (1987), has proved controversial. See, for example, Morris (1995), Gul (1998) and, for a response, Aumann (1998). We allow but do not impose the common priors assumption, as it is not needed for the paradoxes we present.

are *choiceworthy* for i at ω, is nonempty for all ω. Second, for all $E \in \mathcal{I}_i$, 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'. It is conventional to treat the maximisation of subjective expected utility as one of the defining characteristics of a Bayesian model, and we follow that convention here. Consider any player i. For any $s \in S$, for any $s_i' \in S_i$, let $\sigma_i(s, s_i')$ denote the strategy profile created by substituting s_i' for s_i in s (i.e. $\sigma_i[s, s_i'] = [s_1, ..., s_{i-1}, s_i', s_{i+1}, ..., s_n]$). For any prior π_i, for any state ω', for any $E \in \mathcal{I}_i$, let $\pi_i(\omega'|E)$ denote the posterior probability of ω', given E. For each player i, for each state ω, a strategy s_i is *SEU-rational* for i at ω with respect to the information partition $\mathcal{I}_i$ and prior π_i if, for each strategy $s_i' \in S_i$, $\sum_{\omega' \in E} \pi_i(\omega'|E)\, (u_i[\sigma_i(b[\omega'], s_i)] - u_i[\sigma_i(b[\omega'], s_i')]) \geq 0$, where E is the event such that $\omega \in E \in \mathcal{I}_i$. Thus, s_i is SEU-rational for i at ω if it maximizes expected utility for i, conditional on his prior beliefs updated by his information at ω.[5] The choice function χ_i is *SEU-rational* with respect to $\mathcal{I}_i$ 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 $\mathcal{I}_i$ and π_i.

We define a *Bayesian model* of a particular game as an ordered quintuple $<\Omega, b(.), \mathcal{I}, \pi, \chi>$, where Ω is a finite, nonempty set of states and $b(\omega) = (b_1[\omega], ..., b_n[\omega])$, $\mathcal{I} = (\mathcal{I}_1, ..., \mathcal{I}_n)$, $\pi = (\pi_1, ..., \pi_n)$ and $\chi = (\chi_1, ..., \chi_n)$ are, respectively a behaviour function, an information structure, a profile of priors and a profile of 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-Maximisation. For all $i \in N$, for all $\omega \in \Omega$: $\chi_i(\omega) = \{s_i \in S_i |\ s_i$ is SEU-rational at ω with respect to $\mathcal{I}_i$ and $\pi_i\}$.

Knowledge of Own Choice. For all $i \in N$, for all $\omega \in \Omega$: $\omega \in K_i[E(b_i[\omega])]$.

Choice Rationality requires that, at each state, each player's actions are consistent with whatever standard of decision-theoretic rationality is being modelled. SEU-Maximisation

[5] 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*.

stipulates that the standard of rationality is the maximisation of subjective expected utility. Knowledge of Own Choice imposes the obvious restriction that, at each state, every player knows the pure strategy that he chooses.[6]

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

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

Theorem 1 is implied by the analysis of Aumann (1987).[7] 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 Theorem 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 choices of any two distinct players from among their included strategies 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 (or grounds for believing in it). If the representation of CKR is to apply to such cases, it must be possible to impose Independence on the model.

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

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

[7] 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. The proof shows how, for any game in G, the different components of a Bayesian model may be assembled from some correlated equilibrium for the game.

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* (and so played at some state).[8]

We interpret Bayesian models which satisfy both Independence and Privacy as attempting to represent common knowledge of the following standard of practical rationality: each player's beliefs assign independent probabilities to other players' strategies, zero probability to strategies regarded as not rationally playable, and strictly positive probability to all strategies regarded as rationally playable; and each player maximises expected utility relative to these beliefs. We call this standard that of *independent cautious expected utility maximization* (or the *ICEU standard*, for short).[9] Thus, a Bayesian model which satisfies Independence and Privacy is an *ICEU Bayesian model*.

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 three games to illustrate this.

[8] This conception of caution is distinct from that used by some others in the literature (e.g. Asheim and Dufwenberg, 2003), for whom caution requires no strategy to be regarded as entirely impossible. It conforms more closely to that of Börgers and Samuelson (1992) and Pearce (1984).

[9] The ICEU standard is very closely related to that described in Section 5 of Cubitt and Sugden (2010), where an independence condition is added to the 'reasoning-based expected utility' conception of practical rationality introduced in Section 3 of that paper. We use a different name here, to avoid associating the standard with any particular approach to modelling CKR.

Our first exhibit, illustrating the *Proving Too Much Paradox,* is Game 1.[10]

Game 1:

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

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

Proposition 1 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.

Our second exhibit illustrates another way in which a Bayesian modelling approach can seem to prove too much. We call it the *Three-lane Road Game* in memory of a method of marking lanes on single-carriageway roads which was once (but fortunately is no longer) common in Britain. Each curbside lane was designated for slow traffic in the direction consistent with the 'keep left' rule, while a single central lane was designated for overtaking in both directions. If two drivers travelling in opposite directions in their respective slow lanes had simultaneous overtaking opportunities, overtaking would be safe for either of them if and only if the other chose not to overtake. This can be represented as follows, where the

[10] Game 1 is the normal-form of a simple extensive-form 'Centipede' game in which the initial move belongs to Player 1, the second move to Player 2 and the third and final move to Player 3. Although Centipede games have most often been discussed in the literature using the extensive form, our analysis here uses the normal form only.

two strategies available to a player *i* correspond to staying in their curbside lane (in_i) and pulling out into the central one (out_i), and each player is indifferent between all strategy profiles that do not expose them to the risk of simultaneously pulling out.

Game 2: (Three-lane Road)

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

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

Proposition 2 implies that, in every ICEU Bayesian model of Game 2, one of the two players plays the 'risky' strategy (*out*) in some states, while the other plays it in none. The structure of the game is entirely symmetrical with respect to the two players. But, if one player plays *out* in some states and the other plays it in none, there must be some asymmetry that tells one and only one player that they may pull out. The apparent implication of Proposition 2 is that the existence of such an asymmetry is implied merely by the assumption that the ICEU standard of rationality is common knowledge; but there seems to be no way in which the players could discover that asymmetry using only the knowledge attributed to them by that assumption. Of course, there is no paradox in the idea that there *could* be common knowledge of an asymmetry in what rationality requires of the players, grounded on information external to the formal description of the game (for example, information about previous play of the game in some population). The *Three-lane Road Paradox* is the apparent demonstration that a conception of CKR implies that there *must* be such knowledge. Again, we seem to have proved too much.

Neither Game 1 nor Game 2 yields an outright inconsistency in the conditions that define an ICEU Bayesian model. In fact, these conditions are mutually consistent for every two-player game in *G*.[11] However, an inconsistency can be shown using a game introduced by Cubitt and Sugden (1994), which can be thought of as a three-player extension of Game 2. We call the inconsistency shown by this game the *Tom, Dick and Harry Paradox* to match the

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

name given to the game by Cubitt and Sugden (1994), in view of the following story suggested to them 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.

Game 3 (Tom, Dick and Harry)

Player 3: in_3

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

Player 3: out_3

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

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

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

As Theorem 1 shows, *some* standards of rationality can be represented in Bayesian models without contradiction. However, Theorem 2 is troubling for anyone who thinks that the Bayesian modelling framework should be able to represent common knowledge of *any* specific standard of practical rationality. 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 always be represented in a Bayesian model.

As indicated in Section 1, the key to our resolution of these paradoxes is development of a model in which the modes of reasoning that players might use are represented explicitly. We now turn to that task.

3. Reasoning schemes

As a first step, 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. Until then, 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. We impose throughout that every proposition in P can be made up from some set of atomic propositions using a finite number of logical connectives.

We will say, for any finite subset $Q = \{q_1, ..., q_m\}$ of P and any $p \in P$, that p is *logically entailed* by Q if $q_1 \wedge ... \wedge q_m \wedge \neg p$ is a contradiction; and that a set of propositions is *consistent* if no conjunction of its elements is a contradiction. The conjunction of an empty set of propositions is the *null proposition*, which we treat as a tautology, denoted #.[12] We impose that P contains #.

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 (whose elements are the *premises* of the rule) and the second place by an element of P (the *conclusion* of the rule). An inference rule is *valid* if its conclusion is logically entailed by the set of its premises.

An *inference structure* is a triple $R = <P, A(R), I(R)>$, 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

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

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.

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. 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 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. For this reason, we 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, the Lewisian approach does not model what is really true in the world; instead, it models what players have reason to take to be true. In line with this interpretation, we will say that a person *endorses* a reasoning scheme R if he takes its axioms to be true and accepts the authority of its inference rules; a person who endorses R has *reason to believe* each of its theorems.

For any proposition p and for any reasoning scheme R, 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.

4. Common reasoning in a population

Our approach is to model CKR among a population of agents as the existence of a core of shared reasoning which is endorsed by each agent in the population and is commonly attributed to other such agents. In a metaphorical sense, this core of shared reasoning can be thought of as a subroutine of each agent's individual reasoning. It allows each individual to maintain a distinction between (on the one hand) propositions which *everyone* has reason to believe, given the axioms and inference rules that everyone endorses and (on the other hand) propositions which *he* has reason to believe, given the axioms and inference rules that he endorses. Thus, given a finite, non-empty, *population* $N = \{1, \ldots, n\}$ of agents, we postulate the existence, for each agent i, of a reasoning scheme R_i of *private reason* which i endorses, and the existence of a reasoning scheme R^* of *common reason*.

We take as given 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 now define the following concept as a representation of the links between private and common reason. An *interactive reasoning system* among the population $N = \{1, \ldots, n\}$ is a triple $< P_0, R^*, (R_1, \ldots, R_n)>$, where P_0 is a set of primitive 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. The following theorem establishes that, in an interactive reasoning system, there is iterated reason to believe all propositions that are derivable in R^*:

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

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

Our concept of an interactive reasoning scheme is in the spirit of Lewis's (1969, pp. 52–60) analysis of common knowledge, as understood by Cubitt and Sugden (2003). Lewis defines a proposition p to be 'common knowledge' in a population N if some 'state of affairs' A holds, such that (i) everyone in N has reason to believe that A holds, (ii) A 'indicates' to everyone in N that everyone in N has reason to believe that A holds, and (iii) A 'indicates' to everyone in N that p. (A is then the 'basis' for common knowledge that p.) Lewis defines 'A indicates to person i that p' as 'if i has reason to believe that A holds, i thereby has reason to believe that p'. He sketches a proof of the theorem that if p is common knowledge in this sense, and given (not fully specified) premises to the effect that individuals share, and have reason to believe that they share, certain principles of rationality, inductive standards and background information, there is iterated reason to believe p in N.

Cubitt and Sugden (2003) reconstruct the theorem and its proof, using an explicit specification of the key properties of 'indication'. This specification is motivated by the ideas that 'i has reason to believe that p' can be interpreted as saying that p is treated as true in a mode of reasoning that i endorses; and that 'A indicates to i that p' can be interpreted as saying that, in that mode of reasoning, there is an inference from 'A holds' to p. Our concept of an interactive reasoning system embodies a more direct representation of the same ideas. It also allows us to represent Lewis's postulate that certain items of information and modes of reasoning are common to the members of a population by specifying axioms and inference rules of R^*. That some state of affairs A is such that, if it occurs, its occurrence is public and self-evident can be represented by (A holds) $\Rightarrow$ (A holds) $\in A(R^*)$. That there are common standards of background knowledge and inductive inference such that, if there is common reason to believe that A holds, there is thereby common reason to believe p can be represented by «from (A holds), infer p» $\in I(R^*)$, which might be expressed in Lewisian language as 'A indicates p in common reason'. Given these two conditions, it follows immediately from Theorem 3 that if A holds, then there is iterated reason to believe p in N. This result expresses the close affinity between Theorem 3 and Lewis's common knowledge theorem.

Other authors have represented Lewis's concept of common knowledge in different ways, more akin to Bayesian models. Some theorists have represented 'i has reason to believe p' as the Bayesian event $K_i(p)$, and have translated the statement 'if i has reason to believe that A holds, i thereby has reason to believe that p' (i.e. Lewis's definition of indication) as the Bayesian statement $K_i(A \text{ holds}) \subseteq K_i(p)$ (Vanderschraaf, 1998; Gintis, 2009, pp. 141–143). Others have recognised a distinction between reason to believe and knowledge, while still using a set-theoretic framework in which 'i has reason to believe that p' is an event (Sillari, 2005; Paternotte, 2010). Such frameworks are not conducive to the representation of processes of reasoning. Whether or not such representation is (as we believe) fundamental to Lewis's original analysis, it is central to our approach to resolving the paradoxes presented in Section 2. But before we can proceed with this resolution, we must complete our rendition of common knowledge of practical rationality among the players of a game.

5. 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’. This concept does not presuppose any particular principles of practical rationality, but uses a purely formal notion of ‘permissibility’.

Here, and throughout Sections 5–8, we fix a given game in *G*. Our analysis applies to any such game but we suppress phrases of the form ‘for all games in *G*’ except in formal results. Differences between games become important again only in Sections 9 and 10.

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.

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.

The definition of a correlate is extended to recommendations and predictions, so that for each recommendation there is a unique correlate prediction and vice versa. The correlate of a recommendation (resp. prediction) is the conjunction of the correlates of its component permissibility (resp. possibility) propositions.[13] For any non-empty set of players $N' \subseteq N$, a *collective prediction* about N' is a conjunction of the elements 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}: F_i contains at most one maxim whose antecedent is logically equivalent to x_{-i}; and (ii) (*Deductive Closure*) for all x_{-i}', for all non-null y_i': if the material implication $x_{-i}' \Rightarrow y_i'$ is logically entailed by a conjunction of all elements of F_i, then F_i contains a maxim $x_{-i}'' \Rightarrow y_i''$ such that x_{-i}'' is logically equivalent to x_{-i}' and y_i'' logically entails y_i'. By virtue of Distinct Antecedents, a decision rule for i makes a set of recommendations to her that are conditional on logically distinct predictions about the other players. In view of this, the Deductive Closure condition implies that, for any collective prediction, all the permissibility 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.

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

6. Common practical reasoning in a game

We now use the concepts of an interactive reasoning system and of a decision rule, developed in Sections 4 and 5 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 $D = (D_1, ..., D_n)$. We then construct reasoning schemes $R^* = <\varphi(P_0), A(R^*), I(R^*)>$, $R_1 = <\varphi(P_0), A(R_1), I(R_1)>, ... , R_n = <\varphi(P_0), A(R_n), I(R_n)>$ 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 some recommendation that applies to him, it can be inferred that he will act on that recommendation. In this sense, common reason attributes practical rationality to each player.

An interactive reasoning system $<P_0, R^*, (R_1, \ldots, R_n)>$ defined in relation to a profile D of decision rules and constructed according to rules (1) to (4) is a *common-reasoning model* of the game; D is its *common standard of practical rationality*.

It is immediate that, for any profile D of decision rules for any game in G, a corresponding (and unique) common-reasoning model exists: the model is constructed by following rules (1) to (4). What is not so obvious (since rules (1) to (4) attribute substantive axioms, as well as some inference rules besides those of valid inference, to the component reasoning schemes) is whether the model so constructed is consistent. The following theorem establishes this property:

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

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

7. The recommendation algorithm

We now focus on the content of common reason in the common-reasoning model defined by a given profile D of decision rules, in so far as that content relates to permissibility and impermissibility of strategies.[14]

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

[14] From rules (3) and (4) and Theorem 4, each R_i will have the same content as common reason in relation to permissibility and impermissibility of strategies.

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

These arguments indicate that the common-reasoning model, for a given profile of decision rules, defines a 'solution' of the game that is interpretable as indicating which strategies are asserted by common reason to be permissible (resp. impermissible). But, Theorem 4 does not in itself show how we, as analysts, can discover that solution; nor does it indicate a specific line of reasoning whereby common reason can reach the conclusions about permissibility (resp. impermissibility) of strategies that are summarised by the profile of common-reasoning partitions (except, of course, to the extent of indicating that any such line uses the axioms and inference rules of R^*). Each of these gaps can be filled by defining a particular algorithm, as we now explain.

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

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

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

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

[15] Put briefly, the reason is as follows. From rules (2i) and (2ii), for every permissibility proposition that is a theorem of R^*, the correlate of that proposition is also such a theorem. The converse is also true, since R^* has no non-null possibility proposition as an axiom, and (given Theorem 4) no inference rules which allow such a proposition to be derived, unless the correlate permissibility theorem is a theorem of R^*.

(a) s_i is asserted by y_i^{k*} to be permissible if, and only if, $R^*[p_i(s_i)]$; and

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

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

The recommendation algorithm can also be interpreted as indicating a specific line of reasoning by which R^* can establish the conclusions captured by the players' common-reasoning partitions. To see this, consider any profile $D = (D_1, ..., D_n)$ of decision rules. The initiation rule of the recommendation algorithm for D and rule (1) for constructing the corresponding common-reasoning model guarantee that, for each player i, y_i^0 is an axiom (and, therefore, a theorem) of R^*. Now consider any stage $k > 0$ of the recommendation algorithm and suppose that, for each player i, y_i^{k-1} is a theorem of R^*. Then, for any player i, the recommendation z_i^k, defined by operation 1 of the recommendation algorithm, is also a theorem of R^*; and it is obtainable from those just supposed by using the inference rules of R^* specified by rule (2i), followed by those specified by rule (2ii). Then, for each player i, the collective prediction x_{-i}^k, defined by operation 2 of the recommendation algorithm, is also a theorem of R^*; and it is obtainable from those just described by application of rules of valid inference attributed to R^* by rule (2). Finally, for each player i, the recommendation y_i^k, defined by operation 3 of the recommendation algorithm is also a theorem of R^*; and it is obtainable from those just described together with D_i (which, by rule (1), is an axiom of R^*) by application of rules of valid inference attributed to R^* by rule (2). Thus, by induction, for each player i, y_i^{k*} is a theorem of R^*; and one can read off from the recommendation algorithm a specific sequence in which the various inference rules of R^* can be invoked to infer y_i^{k*} from the axioms of R^*.

8. Categorisations

In Section 9 below, we compare our Lewisian common-reasoning approach with the Bayesian approach described in Section 2, for the cases where they are used to model common knowledge of the ICEU standard of practical rationality. But first we define some concepts introduced by Cubitt and Sugden (2010) that are useful in making the comparison, because they provide a convenient way to summarise binary and trinary partitions of sets of strategies. (Recall, from Section 2, that each Bayesian model defines a binary partition of each player's strategy set whereas, from Section 7, each common-reasoning model defines a trinary partition.)

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

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

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

Although the Bayesian modelling framework and the framework of common-reasoning models have very different primitives, we can relate them to each other using the concepts just defined as a bridge.

Consider any Bayesian model M of the game, as defined in Section 2. For each player i, the model specifies a set $S_i^*(M) \subseteq S_i$ of included strategies. Equivalently, for each i, M specifies a categorisation $C_i^M = <S_i^+(M), S_i^-(M)>$ of S_i, where $S_i^+(M) = S_i^*(M)$ and $S_i^-(M) = S_i \backslash S_i^*(M)$.[16] Thus, aggregating across all players, M specifies a single categorisation $C^M =$

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

$\cup^*_{i\in N}$ $<S_i^+(M), S_i^-(M)>$ of $\mathbb{S}$. We will say that C^M is the *inclusion categorisation* with respect to Bayesian model M. By construction, C^M is exhaustive.

Now consider the common-reasoning model of the game, for a given profile D of decision rules, as defined in Section 6. As Section 7 showed, for any player i, this model defines a trinary common-reasoning partition of S_i, two of whose elements are $\{s_i \in S_i| R^*[p_i(s_i)]\}$ and $\{s_i \in S_i| R^*[\neg p_i(s_i)]\}$, where R^* is common reason. Thus, the model defines, for each i, a categorisation $<S_i^+(D), S_i^-(D)>$ of S_i, where $S_i^+(D)= \{s_i \in S_i| R^*[p_i(s_i)]\}$ and $S_i^-(D) = \{s_i \in S_i| R^*[\neg p_i(s_i)]\}$.[17] Again aggregating across players, the common-reasoning model for the profile D specifies a single categorisation $C^D = \cup^*_{i\in N} < S_i^+(D), S_i^-(D)>$ of $\mathbb{S}$. Unlike C^M, C^D may or may not be exhaustive, depending on the common-reasoning partitions resulting from profile D. As the positive (resp. negative) component of C^D is the set of strategies whose permissibility (resp. impermissibility) is established in common reason in the common-reasoning model, we will say that C^D is the *common-reasoning solution* of the game, with respect to the profile D of decision rules.

9. ICEU Bayesian models revisited

In this Section, we compare our approach to modelling CKR, set out in Sections 3–7, to the Bayesian approach, set out in Section 2. We focus on the cases in which each approach is adapted to a conception of practical rationality provided by the ICEU standard.

We have already defined the concept of a Bayesian model which incorporates the ICEU standard – the ICEU Bayesian model. For some games, as Theorem 2 shows, no such model exists. Alternatively, as is evident from Game 2, a given game may have more than one such model.

In contrast, the common-reasoning model is uniquely defined by the rules set out in Section 6, for any game and any profile D of decision rules. But, to relate our approach to ICEU Bayesian models, we still need to specify a class of common-reasoning models in which the conception of practical rationality is ICEU. That is, we need to define a profile D of ICEU decision rules. We do this in the following way.

[17] Since the common-reasoning partition for player i is a partition of S_i, condition (i) of the definition of a categorisation is satisfied. That conditions (ii) and (iii) are satisfied too follows from the facts that decision rules are defined in terms of predictions and recommendations, and conditions analogous with (ii) and (iii) are embedded in the definitions of 'prediction' and 'recommendation'.

For any player i and for any collective prediction x_{-i} about $N\backslash\{i\}$, we define a probability distribution over S_{-i} as *ICEU-consistent* with x_{-i} if it satisfies the following three conditions. First, probabilities are independent in the sense that, for each $s_{-i} \in S_{-i}$, the probability of s_{-i} is the product of the marginal probabilities of the individual strategies appearing in s_{-i}. Second, every strategy that x_{-i} asserts to be impermissible has zero marginal probability. Third, every strategy that x_{-i} asserts to be permissible has strictly positive marginal probability. An *ICEU maxim* for i is a maxim $x_{-i} \Rightarrow y_i$ such that (i) y_i asserts $p_i(s_{\mathrm{i}})$ if, and only if, s_i maximises i's expected utility relative to *all* probability distributions that are ICEU-consistent with x_{-i}, and (ii) y_i asserts $\neg p_i(s_{\mathrm{i}})$ if, and only if, s_i does *not* maximise i's expected utility relative to *any* probability distribution that is ICEU-consistent with x_{-i}. Because collective predictions and recommendations are conjunctions of more basic propositions, and because the elements of a given set of propositions can be conjoined in different orders, there may be collective predictions (resp. recommendations) that are formally distinct from, but logically equivalent to x_{-i} (resp. y_i), and so there may be more than one ICEU maxim with the logical content of $x_{-i} \Rightarrow y_i$. By taking exactly one maxim from every set of logically equivalent ICEU maxims for i, we can construct a *non-redundant* set F_i of ICEU maxims for each player i.

Consider any two ICEU maxims $x_{-i}' \Rightarrow y_i'$ and $x_{-i}'' \Rightarrow y_i''$. It follows from the definition of an ICEU maxim that if x_{-i}' and x_{-i}'' are logically equivalent, then so too are y_i' and y_i''. Hence, given the definition of non-redundancy, F_i satisfies Distinct Antecedents. It also follows from the definition of an ICEU maxim that if x_{-i}' logically entails x_{-i}'', then y_i' logically entails y_i''.[18] Thus, F_i satisfies Deductive Closure. So any conjunction of the elements of a non-redundant set F_i of ICEU maxims for i is a decision rule for that player. Since all such conjunctions are logically equivalent, we can fix on any one of these as 'the' ICEU decision rule $D_{i.}$ In this way, we can construct 'the' profile D of ICEU decision rules and hence 'the' *ICEU common-reasoning model*. This model implies a unique common-reasoning solution, which we may unambiguously take as the *ICEU common-reasoning solution* (since every profile of ICEU decision rules for the game yields the same common-

[18] Intuitively: as collective predictions about $N\backslash\{i\}$ become stronger, the recommendations licensed by the ICEU standard of rationality cannot become weaker. The reason is that, as collective predictions become (strictly) stronger, the restrictions on probabilities required by ICEU-consistency with such predictions tighten (that is, more strategies must have zero marginal probability and/or more strategies must have strictly positive marginal probability), so making it 'easier' for a strategy to be expected utility maximising for all probability distributions satisfying the restrictions and also easier for it to be expected utility maximising for no such probability distribution.

reasoning solution). In view of Section 7, the solution can be found using the recommendation algorithm for a profile of ICEU decision rules.

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. 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 specifying what players have reason to believe about the game, and 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. The following theorem establishes that this conjecture is indeed correct.

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

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

Theorem 7 establishes that, in the special case in which the ICEU common-reasoning solution is exhaustive, and with respect to the resulting categorisations, the Bayesian and common-reasoning approaches are equivalent. In this case, as is evident from the proof of Theorem 7, an ICEU Bayesian model can be constructed using *any* profile of independent priors that respects the common-reasoning solution.

Summing up the conclusions of this Section, we have found that, for games in which the ICEU common-reasoning 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 describes 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 common-reasoning solution is *not* exhaustive. We turn to such cases in the next Section.

10. Resolving the paradoxes

In Section 2, we presented three paradoxes, using Games 1, 2 and 3. The results of Section 9 imply that, for those games to be genuinely troubling exhibits for the Bayesian approach to modelling CKR, it would have to be the case that the ICEU common-reasoning solutions to these games are not exhaustive. It is straightforward to show that this is indeed the case.

Theorem 5 implies that, in each game, the ICEU common-reasoning solution is identifiable from the final output of the recommendation algorithm for 'the' profile of ICEU decision rules. In Game 1, the outputs of this recommendation algorithm are: $y_1^0 = y_2^0 = \#$; $y_1^1 = \neg p_1(third)$, $y_2^1 = \#$; $y_1^2 = \neg p_1(third)$, $y_2^2 = p_2(left)$; $y_1^3 = y_1^2$, $y_2^3 = y_2^2$. Thus, the ICEU common-reasoning solution for Game 1 is $<\{left\}, \{third\}>$. In Games 2 and 3, the outputs of the algorithm are (for $i = 1, 2$ in the case of Game 2, and for $i = 1, 2, 3$ in the case of Game 3): $y_i^0 = \#$; $y_i^1 = p_i(in_i)$; $y_i^2 = y_i^1$. Thus the ICEU common-reasoning solutions for Games 2 and 3 are $<\{in_1, in_2\}, \varnothing>$ and $<\{in_1, in_2, in_3\}, \varnothing>$ respectively. In each case, some strategies (*first*, *second* and *right* in Game 1; out_1 and out_2 in Game 2; and out_1, out_2 and out_3 in Game 3) are neither shown to be permissible nor shown to be impermissible.

It is an immediate corollary of Theorem 6 (as C^M is exhaustive by definition) that, whenever the ICEU common-reasoning solution to a game is *not* exhaustive, one of two cases must hold. The first case, the discussion of which we postpone, is that no ICEU Bayesian model exists. The Tom, Dick and Harry Paradox of Game 3 illustrates this case.

The second case is that at least one ICEU Bayesian model exists, and that for each such model M, the inclusion categorisation has the property $C^M \supset^* C^*$. This implies that, for any such M, there is at least one player i and strategy s_i such that *either* s_i is included in M but is not shown to be permissible by common reason in the ICEU common-reasoning model, *or*

s_i is excluded from *M* but is not shown to be impermissible by common reason in the common-reasoning model. In such cases we will say that the inclusion of s_i in (resp. the exclusion of s_i from) the Bayesian model is *ungrounded.* In some games, specific ungrounded inclusions (resp. exclusions) are common to *every* ICEU Bayesian model, with the apparent implication that the corresponding propositions about rational playability 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 source of the Proving Too Much Paradox is that the apparent implication is invalid. We take it that a Bayesian model is to be interpreted as a formal representation of what individuals know and believe in some *conceivable 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 conceivable world in a way that is external to the Bayesian modelling framework: conceivable worlds are what Bayesian models are models *of*. On our understanding, common knowledge of a specific standard of practical rationality is a property of a putative conceivable world; in attempting to construct a Bayesian model which incorporates that standard, one is attempting to represent that world in a formal model. In setting out to do this, one is not entitled to presuppose the success of one's modelling strategy. Thus, from the result (Proposition 1) that every ICEU Bayesian model of Game 1 includes *first*, *left* and *right* and excludes *second* and *third*, one is not entitled to infer that the corresponding permissibility propositions are common knowledge in the world that is being modelled.

This analysis of the Proving Too Much Paradox is supported by the Tom, Dick and Harry Paradox of Game 3. That game provides an exhibit of a conceivable world in which there is common knowledge of a particular standard of practical rationality, namely ICEU, but of which no Bayesian model can be constructed. We submit that the Tom, Dick and Harry Paradox does not reveal the incoherence of supposing such a world to be possible (and hence the incoherence of the concept of common knowledge of the ICEU standard of rationality). Rather, it reveals a limitation of the Bayesian modelling approach, as formulated in Section 2.

To make this argument convincing, one needs to show that there can be a coherent understanding of a world in which the players of Game 3 have common knowledge that each of them endorses and acts on the ICEU standard of rationality. We submit that the ICEU common-reasoning model of this game provides just such an understanding. In this model, for each player i = 1, 2, 3, neither the possibility nor the impossibility of out_i is a theorem of common reason, but the possibility 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. These conclusions seem entirely coherent, despite the fact that they cannot be expressed in the Bayesian modelling framework.

The Three-lane Road Paradox illustrates a further difference between the Bayesian and Lewisian approaches. Proposition 2 establishes that ICEU Bayesian models of Game 2 can be partitioned into two classes – those models M' for which the inclusion categorisation is $C^{M'}$ = <{in_1, in_2, out_2}, {out_1}>, and those models M'' for which the inclusion categorisation is $C^{M''}$ = <{in_1, out_1, in_2}, {out_2}>. The paradox is that every one of these models seems to represent a world in which the players have common knowledge of an asymmetry between what rationality requires of one of them and what it requires of the other. Since the formal structure of the game is symmetrical between the two players, and since no other information has been used in deriving Proposition 2, it is puzzling that the existence of some such asymmetry appears to be an implication of common knowledge of the ICEU standard of rationality. The paradox is resolved by understanding that there *can* be worlds in which Game 2 is played, there is common knowledge of the ICEU standard, and the implications of that standard are symmetrical between the players. The ICEU common-reasoning model represents precisely such a world. Proposition 2 tells us only that such worlds cannot be represented by ICEU Bayesian models.[19]

So we suggest that the three paradoxes of Section 2 stem from the same source, namely the presumption that, for any game, an ICEU Bayesian model can be constructed of every conceivable world in which the players of that game have common knowledge of the ICEU standard of rationality. Our diagnosis of all three paradoxes is the same: it is that, when the ICEU common-reasoning 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 consistent 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

[19] The Lewisian approach also allows one to model interactive reasoning systems in which common reason has axioms and inference rules in addition to those that define common-reasoning models. Such systems can represent worlds in which Game 2 is played and common reason generates asymmetric recommendations. This, essentially, is how Lewis models conventions (Lewis, 1969; Cubitt and Sugden, 2003).

taken to be common knowledge. Since Bayesian models induce exhaustive categorisations of $\mathbb{S}$, no such model can represent a conceivable 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 conceivable 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 explicit analysis of reasoning itself. We believe that, by building on the foundations of Lewis's account of common knowledge, we have been able to show the feasibility and usefulness of such an analysis.

Appendix 1: Categorisation procedures and recommendation algorithms

In this appendix, we introduce the concept of a 'categorisation procedure', as defined and analysed by Cubitt and Sugden (2010) (henceforth CS10). This enables us to demonstrate a relationship between that concept and that of the recommendation algorithm, introduced in Section 7. The fruits of this demonstration are twofold.

First, it allows us to use a result from CS10, together with new results presented here, as ingredients for proofs presented in Appendix 2. In relation to this objective, note that the ingredients which this appendix provides for Appendix 2 (in particular, Propositions A1–A3 below) are free-standing, in the sense that their proofs draw only on definitions from CS10 (repeated here for convenience) and from the main text of the current paper; they do not presuppose any of the *results* from the main text. Where not drawn directly from CS10, the proofs of Propositions A1–A3 are presented at the end of this appendix.

Second, by proving Proposition A3 below we establish a precise sense in which, for a given standard of practical rationality, the categorisation procedure generates the same sequence of outputs as the recommendation algorithm. This allows a substantiation of CS10's claim that the categorisation solution can be interpreted as the result of reasoning that the players can undertake.

Throughout this appendix, our analysis applies to any given game in G. We begin by extending the concepts introduced in Section 8, in a way that follows CS10. Recall that Section 8 defined the concept of a categorisation of S_i; and the concept of a categorisation of $\cup_{i\in N'} S_i$, for any non-empty set $N' \subseteq N$ of players. In what follows, we require the case where $N' = N\backslash\{i\}$, for any given player i, as well as the case (already introduced) where $N' = N$. We use $\mathbb{S}_{-i}$ as a shorthand for $\cup_{i\in N\backslash\{i\}} S_i$; the positive and negative components of a categorisation of the latter set will typically be denoted $\mathbb{S}_{-i}^{+}$ and $\mathbb{S}_{-i}^{-}$.

We denote the set of categorisations of S_i, the set of categorisations of $\mathbb{S}_{-i}$ and the set of categorisations of $\mathbb{S}$ by, respectively, $\Phi(S_i)$, $\Phi(\mathbb{S}_{-i})$ and $\Phi(\mathbb{S})$. The *null categorisation* $<\varnothing, \varnothing>$ is an element of each of these sets. Where convenient, we use C_i, C_i', and so on, to denote particular categorisations in $\Phi(S_i)$; C_{-i}, C_{-i}', and so on, to denote particular categorisations in $\Phi(\mathbb{S}_{-i})$; and C, C', and so on, to denote particular categorisations in $\Phi(\mathbb{S})$. We extend to categorisations in $\Phi(S_i)$ and $\Phi(\mathbb{S}_{-i})$, in the obvious way, the definitions of the relations $\sqsupseteq^*$ ('has weakly more content than') and of $\sqsupset^*$ ('has strictly more content than'), introduced in Section 8 for categorisations in $\Phi(\mathbb{S})$. (See CS10, Section 2, for details.)

CS10 defines 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 $f_i(C_{-i}'') \supseteq^* f_i(C_{-i}')$.

The content of a given profile $f = (f_1, \ldots, f_n)$ of categorisation functions can be expressed as a single function ζ: $\Phi(\mathbb{S}) \rightarrow \Phi(\mathbb{S})$, constructed as follows. Let $C = <\mathbb{S}^+, \mathbb{S}^->$ be any categorisation of $\mathbb{S}$. For each player i, define $C_{-i} = <\mathbb{S}^+ \backslash S_i, \mathbb{S}^- \backslash S_i>$. Next, define $S_i^{+\prime}$ and $S_i^{-\prime}$ as, respectively, the positive and negative components of $f_i(C_{-i})$. Finally, define $\zeta(C) = \cup^*_{i \in N} <S_i^{+\prime}, S_i^{-\prime}>$. We will say that ζ *summarises* f. A function ζ: $\Phi(\mathbb{S}) \rightarrow \Phi(\mathbb{S})$ that summarises some profile f of categorisation functions is an *aggregate categorisation function*.

For any aggregate categorisation function ζ, the *categorisation procedure* is defined by CS10 by the following pair of instructions, which generate a sequence of categorisations $C(k) \equiv <\mathbb{S}^+(k), \mathbb{S}^-(k)>$ of $\mathbb{S}$, for successive stages $k \in \{0, 1, 2, \ldots.\}$, inductively, as follows:

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

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

The procedure *halts* at the lowest value of k' for which $C(k') = C(k'–1)$; this value of k' will be denoted by k^*. $C(k^*)$ is the *categorisation solution* of the game, relative to ζ. CS10 proves the following result (their Proposition 1):

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

We are now in a position to relate these concepts from CS10 to those introduced in Sections 5–7. The key to this step is a correspondence between decision rules and categorisation functions.

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

The following result, the proof of which makes important use of Deductive Closure, establishes the correspondence between decision rules and categorisation functions.

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

Together with the definition of an aggregate categorisation function, Proposition A2 implies the following: for any profile $D = (D_1, ..., D_n)$ of decision rules, there exists a unique profile $f = (f_1, \ldots, f_n)$ of categorisation functions, and 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*. It is then immediate that any profile D of decision rules defines a unique categorisation procedure (the categorisation procedure for the ζ that encodes D).

Recall, from Section 7, that any profile D of decision rules also defines a recommendation algorithm. The latter algorithm generates, for each player i, an output y_i^k, for each of its stages $k = 0, 1, 2, ...$, where each such output is a recommendation to i. Since any such recommendation is encoded by a categorisation of S_i, and such categorisations can be aggregated across players, the combined output of each stage k of the recommendation algorithm is *encoded* by a categorisation of $\mathbb{S}$, defined as $\cup^*_{i \in N} <S_i^+(k), S_i^-(k)>$ where, for each i, $<S_i^+(k), S_i^-(k)>$ encodes y_i^k.

Thus, any profile D of decision rules defines two sequences of categorisations of $\mathbb{S}$: one consisting of the outputs of successive stages of the categorisation procedure for the aggregate categorisation function that encodes D; and the other consisting of the categorisations that encode the outputs of successive stages of the recommendation algorithm. The following result establishes that these sequences are one and the same:

> *Proposition A3*: Consider any game in G, and any profile D of decision rules for its players. Let ζ be the aggregate categorisation function that encodes D. Let $C(0)$, $C(1)$, …. be the sequence of categorisations generated by the categorisation procedure

> for ζ; and $C'(0)$, $C'(1)$, … be the sequence of categorisations that encode the combined outputs of successive stages of the recommendation algorithm for D. Then, for each $k \in \{0, 1, 2, \ldots\}$, $C(k) = C'(k)$.

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

Proposition A3 is also of independent interest, in allowing us to demonstrate a relationship between CS10's concept of a categorisation solution and the concept of a common-reasoning solution. Consider any profile D of decision rules. By Theorem 4, the common-reasoning model with respect to D is consistent, and hence (as explained in Section 8) induces a unique common-reasoning solution. Theorem 5 establishes that the content of this solution is given by the final output of the recommendation algorithm for D. Proposition A3 establishes that the sequence of categorisations generated by the recommendation algorithm is identical to the sequence generated by the corresponding categorisation procedure, and hence that the categorisation induced by the final output of the recommendation algorithm is identical to the categorisation solution. Thus, the following result is a corollary of Proposition A3 and Theorems 4 and 5:

> *Corollary*: Consider any game in G and any profile D of decision rules for that game. Let ζ be the aggregate categorisation function that encodes D. Then the categorisation solution relative to ζ is identical to the common-reasoning solution with respect to D.

This result shows that CS10's concept of a categorisation solution can be justified as the implication of a Lewisian understanding of CKR, as represented in the concept of a common-reasoning model.

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

Proof of Proposition A2: To show that the function f_i which encodes a decision rule D_i is a categorisation function, it suffices (since, by construction, f_i has the appropriate range and domain) to show that f_i satisfies Monotonicity. That this condition is satisfied follows from the fact that D_i is a conjunction of all elements of a set F_i of maxims for i, which satisfies Distinct Antecedents and Deductive Closure. To see this, suppose f_i does not satisfy Monotonicity. Then there are C_{-i}', $\mathrm{C}_{-i}'' \in \Phi(\mathbb{S}_{-i})$ such that $C_{-i}'' \supset^* C_{-i}'$ and *not* $f_i(C_{-i}'') \sqsupseteq^* f_i(C_{-i}')$. So F_i contains maxims $x_{-i}' \Rightarrow y_i'$ and $x_{-i}'' \Rightarrow y_i''$, such that (i) x_{-i}'' entails x_{-i}' and (ii) y_i'' does not entail y_i'. Notice that (ii) implies that y_i' is non-null. Because of (i), the conjunction of these maxims entails $x_{-i}'' \Rightarrow y_i'$. So, by Deductive Closure, F_i contains a maxim $x_{-i}^* \Rightarrow y_i^*$ where x_{-i}^* is logically equivalent to x_{-i}'' and y_i^* entails y_i'. But because of Distinct

Antecedents, this requires $x_{-i}* = x_{-i}''$ and hence $y_i* = y_i''$. Thus y_i'' entails y_i', contradicting (ii). □

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

Appendix 2: Proofs of results from main text

Proof of Theorem 1: For any game in G, let ρ: $S \to [0, 1]$ 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: $S_i \to S_i$, $\sum_{s\in S} \rho(s)\ (u_i[s] - u_i[\sigma_i(s, g_i[s_i])]) \geq 0$. 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 $\mathcal{I}$ such that, for each player i, for each strategy $s_i \in S_i^*$: $E(s_i) \in \mathcal{I}_i$. 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 $\mathcal{I}_i$ and π_i. By construction, the Bayesian model $<\Omega, b(.), \mathcal{I}, \pi, \chi>$ 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. □

Preliminaries for proofs relating to ICEU Bayesian models (Propositions 1 and 2 and Theorems 2, 6 and 7): 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 $\mathcal{I}_i$. Define $U_i(s_i| E)$ as the expected value of $u_i(s)$, given that player i

chooses s_i and that the probability distribution over S_{-i} is determined by conditioning i's prior π_i on the event E. 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 A1*: 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 $E \in \mathcal{I}_i$. Since this inequality holds for each such E, it must also hold for their union. By Knowledge of Own Choice, the union of all such events E is $E(s_i)$. Thus, for all $s_i' \in S_i$, $U_i(s_i | E[s_i]) \geq U_i(s_i' | E[s_i])$. By Independence, 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 $\mathcal{I}_i$. 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 $E' \in \mathcal{I}_i$ 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 1. Consider any ICEU Bayesian model of Game 1. For player 1, *third* is not marginally EU-maximising with respect to any probability distribution over player 2's strategies. Thus, by Lemma A1, *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 A1, *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 A1, *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 A1, $S_2^* = \{left, right\}$. □

Proof of Proposition 2. Suppose there exists an ICEU Bayesian model of Game 2. Using Lemma A1, it is straightforward to show that $S_1^* = \{in_1\} \Rightarrow S_2^* = \{in_2, out_2\}$, $S_1^* = \{out_1\} \Rightarrow S_2^* = \{in_2\}$, and $S_1^* = \{in_1, out_1\} \Rightarrow S_2^* = \{in_2\}$. Symmetrically, $S_2^* = \{in_2\} \Rightarrow S_1^* = \{in_1, out_1\}$, $S_2^* = \{out_2\} \Rightarrow S_1^* = \{in_1\}$, and $S_2^* = \{in_2, out_2\} \Rightarrow S_1^* = \{in_1\}$. Given that S_1^* and S_2^* are non-empty, these material implications can be satisfied simultaneously only if *either* (i) $S_1^* = \{in_1\}$ and $S_2^* = \{in_2, out_2\}$ *or* (ii) $S_1^* = \{in_1, out_1\}$ and $S_2^* = \{in_2\}$. □

Proof of Theorem 2. Consider Game 3 and suppose that an ICEU Bayesian model of this game exists. First, suppose (*Supposition 1*) that there are two distinct players i, j such that $out_i \in S_i^*$ and $out_j \in S_j^*$. Because of the symmetries of the game, there is no loss of generality in setting $i = 1$ and $j = 2$. This implies 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 A1, $out_1 \notin \mathrm{S}_1^*$, a contradiction. So Supposition 1 is false. Since there are three players, this entails that there are two distinct players i, j such that $out_i \notin \mathrm{S}_i^*$ and $out_j \notin \mathrm{S}_j^*$. Without loss of generality, set $i = 1$ and $j = 2$. Then out_1 is marginally EU-maximising and so, by Lemma A1, $out_1 \in S_i^*$, a contradiction. Thus, Game 3 has no ICEU Bayesian model. □

Proof of Theorem 3. Consider any interactive reasoning system $<P_0, R^*, (R_1, \ldots, R_n)>$ among the population N. Suppose that, for some $p \in \varphi(P_0)$, $R^*(p)$ 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 $i, j \in N$: $R_i[R^*(R_j[p])]$	(from (L4), using Awareness)
(L6)	for all $i, j \in N$: $R_i[R_j(p)]$	(from (L5), using Authority)
(L7)	for all $i, j \in N$: $R^*[R_i(R_j[p])]$	(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 Theorem 4: Consider any game in G, and any profile $D = (D_1, ..., D_n)$ of decision rules for its players. Let $< P_0, R^*, (R_1, \ldots, R_n)>$ be the common-reasoning model of the game, defined in relation to D.

We begin by defining, as a counterpart to R^*, an inference structure R_-^* which has the same domain and axioms as R^* but whose inference rules are, in a sense to be defined, 'weaker' than those of R^*. To do this, 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 $i \in N$. I_3 is the set of inference rules of the form «from $\{R_i(y_i)\}$, infer z_i», where $i \in N$, y_i is a recommendation to i, and z_i is the prediction about i that is the correlate of y_i. I_4 is the set of inference rules of the form «from $\{y_i\}$, infer z_i», where $i \in N$, y_i is a recommendation to i, and z_i is the prediction about i that is the correlate of y_i. I_5 is the set of inference rules of the form «from $\{z_1, ..., z_{i-1}, z_{i+1}, ..., z_n\}$, infer $z_1 \wedge ... \wedge z_{i-1} \wedge z_{i+1} \wedge ... \wedge z_n$», where $i \in N$ and each z_j is a prediction about the relevant player j. I_6 is the set of inference rules of the form «from $\{D_i, x_{-i}\}$, infer y_i», where $i \in N$, x_{-i} is a collective prediction about $N\backslash\{i\}$, x_{-i} is logically equivalent to the antecedent of some maxim asserted by D_i, and y_i is the consequent of that maxim.

R^* is fully specified by its domain $\varphi(P_0)$ and axiom set $A(R^*)$, and by the condition that it has the inference rules contained in $I_1 \cup I_2 \cup I_3$. We define R_-^* as the inference structure that has the domain $\varphi(P_0)$, the axiom set $A(R_-^*) = A(R^*)$ and the set of inference rules $I_4 \cup I_5 \cup I_6$. Note that this implies that R_-^* does not have all rules of valid inference. (Recall that the concept of an inference structure differs from that of a reasoning scheme by allowing this possibility. Intuitively, R_-^* is endowed with just the inference rules necessary for it to validate the operations of the recommendation algorithm, defined in Section 7. This algorithm does not feature in Theorem 4, but the relationship between it and R_-^* is important for the proof of Theorem 5.)

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

The proof of Theorem 4 uses the following lemmas:

> *Lemma A2*: For each $i \in N$ and for each $s_i \in S_i$: (i) $s_i \in \mathbb{S}^{+*}$ if, and only if, $p_i(s_i)$ is asserted by some theorem in $T(R_-^*)$; (ii) $s_i \in \mathbb{S}^{-*}$ if, and only if, $\neg p_i(s_i)$ is asserted by some theorem in $T(R_-^*)$; (iii) $s_i \in \mathbb{S}^{+*}$ if, and only if, $m_i(s_i)$ is asserted by some theorem in $T(R_-^*)$; (iv) $s_i \in \mathbb{S}^{-*}$ if, and only if, $\neg m_i(s_i)$ is asserted by some theorem in $T(R_-^*)$.

Proof: For the purposes of this proof, we extend the definitions of 'encoding', given in Section 8 and Appendix 1, to allow consistent sets of permissibility or possibility propositions for the set of players N to be encoded by categorisations of $\mathbb{S}$. In the case of permissibility, a strategy s_i is assigned to the positive (resp. negative) component of the encoding categorisation if, and only if, $p_i(s_i)$ (resp. $\neg p_i[s_i]$) is an element of the relevant encoded set. Similarly, in the case of possibility, s_i is assigned to the positive (resp. negative) component of the encoding categorisation if, and only if, $m_i(s_i)$ (resp. $\neg m_i[s_i]$) is an element of the relevant encoded set. Because of conditions (ii) and (iii) in the definition of a categorisation, it is not the case that all consistent sets of permissibility (resp. possibility) propositions can be encoded in this way. However, whenever we use this definition of encoding, those conditions are satisfied.

We now define a *proof algorithm* which progressively 'discovers' the content of the set $T(R\_*)$ by following a particular sequence of steps of reasoning that are licensed by the axioms and inference rules of R_*. R_* is specified so that no other reasoning is possible. The steps of the proof algorithm are grouped into 'phases' of three and numbered 1.1, 1.2, 1.3; 2.1, 2.2, 2.3; 3.1 , … . The set of theorems discovered up to the end of any step l is denoted $T_l(R\_*)$. At the end of each phase k (i.e. at the end of step k.3), the intersection of $T_{k.3}(R\_*)$ and the set of permissibility propositions is encoded as the categorisation $C(k)$. The algorithm is initiated by defining the set of already discovered theorems as $A(R\_*)$. Since the intersection of $A(R\_*)$ with the set of permissibility propositions is $\{\#\}$, this set is encoded as $C(0) = <\varnothing, \varnothing>$.

There are inference rules in each of the sets I_4, I_5 and I_6 that can use $\{\#\}$ as a premise. Since, for each player i, # is a both a (null) recommendation to i and a (null) prediction about i, inference rules in I_4 allow # to be inferred from $\{\#\}$; but that does not lead to new theorems. This is step 1.1; $T_{1.1} = A(R\_*)$. I_5 allows the proposition $\# \wedge \# \wedge ... \wedge \# \wedge$ which conjoins $N - 1$ null propositions (and which we denote $\#^{N-1}$) to be inferred from $\{\#\}$. This is step 1.2; $T_{1.2} = A(R\_*) \cup \{\#^{N-1}\}$. The only inference rules of R_* that can use subsets of $T_{1.2}$ as premises and generate conclusions that are not themselves elements of $T_{1.2}$ are those in I_6. Thus, the first step in deriving any non-null 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. This is step 1.3. $T_{1.3}(R\_*)$ contains the elements of $T_{1.2}$ and all theorems that can be proved in this way. This is the end of phase 1. From an examination of the reasoning in this phase, it is evident that the set of permissibility propositions asserted by theorems in $T_{1.3}(R\_*)$ is encoded by the categorisation $C(1) = \zeta[C(0)]$.

The only inference rules of R_-* that 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 uses 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. $T_{2.1}(R_-*)$ contains the elements of $T_{1.3}(R_-*)$ and all theorems that can be proved in this way.

The only inference rules of R_-* that 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 and I_6. Step 2.2 uses 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 $\{z_1, ..., z_{i-1}, z_{i+1}, ... z_n\} \subseteq T_{2.1}(R_-*)$; 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. $T_{2.2}(R_-*)$ contains the elements of $T_{2.1}(R_-*)$ and all theorems that can be proved in this way.

Step 2.3 follows the model of step 1.3, using inference rules of the form «from $\{D_i, x_{-i}\}$, infer y_i», where $i \in N$, x_{-i} is a collective prediction in $T_{2.2}(R_-*)$, to arrive at $T_{2.3}(R_-*)$, defined to contain the elements of $T_{2.2}(R_-*)$ and all recommendations y_i that can be proved in this way. This is the end of phase 2. From an examination of the reasoning in this phase, it is evident that the set of permissibility propositions asserted by theorems in $T_{2.3}(R_-*)$ is encoded by the categorisation $C(2) = \zeta[C(1)]$.

Each succeeding phase follows the model of phase 1, using inference rules in I_4 (resp. I_5, I_6) in step k.1 (resp. k.2, k.3). For each $k > 0$, the set of permissibility propositions asserted by theorems in $T_{k.3}$ is encoded in the categorisation $C(k) = \zeta[C(k-1)]$.

Note that the sequence of categorisations $C(0)$, $C(1)$, …. defined by the proof algorithm is identical to the sequence generated by the categorisation procedure for ζ, defined in Appendix 1. Thus, by Proposition A1 and the definition of halting of the categorisation procedure, there is some finite $k*$ such that $C(k*) = C(k*-1) \supset* ... \supset* C(1) \supset* \mathrm{C}(0)$. This implies that no new theorems 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_-*. This proves parts (i) and (ii) of Lemma A2. The 'only if' implications of parts (iii) and (iv) follow from parts (i) and (ii), together with R_-* having the inference rules in I_4. The 'if' implications of parts (iii) and (iv) also follow from parts (i) and (ii) because $A(R_-*)$ contains no possibility propositions other than #, and $I(R_-*)$ contains no inference rules which have possibility propositions as conclusions, other than those in I_4. □

Lemma A3: $T(R_{-}*)$ 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: $T^1 = A(R_{-}*) \cup \{\#^{N-1}\}$; $T^2 = \{p \in T(R_{-}*) \mid p$ is a conjunction of one or more predictions about players, at least one of which is non-null$\}$; $T^3 = \{p \in T(R_{-}*) \mid p$ is a non-null recommendation to some $i\}$. From the definitions of these subsets, Lemma A2 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 $\mathbb{S}$, it follows from the definition of a categorisation that T^3 is consistent. Since each element of T^2 is a conjunction of a set of correlates of elements of T^3, and since T^3 is consistent, T^2 is consistent. The non-null elements of T^1 are decision rules for different players, so that, from the definition of a decision rule, T^1 is consistent. Since the elements of T^2 are conjunctions of predictions, since the non-null elements of T^1 are conjunctions of material implications whose consequents are recommendations, and since T^1 and T^2 are each consistent, $T^1 \cup T^2$ is consistent. Finally, by the specification of I_6 and the fact that every proposition in T^3 is the conclusion of an application of an inference rule in that set, every proposition in T^3 is logically entailed by $T^1 \cup T^2$. Thus, $T^1 \cup T^2 \cup T^3$, i.e. $T(R_{-}*)$, is consistent. □

Lemma A4: (i) $T(R*)$ is consistent. (ii) For each $i \in N$, and for each $s_i \in S_i$: (a) $R*[p_i(s_i)]$ if, and only if, $R_{-}*[p_i(s_i)]$; (b) $R*[\neg p_i(s_i)]$ if, and only if, $R_{-}*[\neg p_i(s_i)]$.

Proof: By Lemma A3, $T(R_{-}*)$ 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 $T(R*)$. This proves part (i) of the lemma. Given that $T(R*)$ and $T(R_{-}*)$ are consistent, that $A(R*) = A(R_{-}*)$, and that all decision rules satisfy Deductive

Closure, any permissibility proposition that can be derived from $A(R^*)$ using inference rules in $I_1 \cup I_2 \cup I_3$ can also be derived from $A(R\_^*)$ using inference rules in $I_4 \cup I_5 \cup I_{6,}$ and vice versa. This proves part (ii). □

Lemma A5: For each $i \in N$, $T(R_i)$ is consistent.

Proof: By part (i) of Lemma A4, $T(R^*)$ 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 $R^*(.)$, while T^2 is a consistent set which contains no proposition of the form $\neg R^*(.)$, $T^1 \cup T^2$ 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. □

Finally, Theorem 4 follows immediately from part (i) of Lemma A4 and Lemma A5. □

Proof of Theorem 5: Consider any game in G and any profile D of decision rules for the game. Since $\mathbb{S}$ is a finite set, part (i) of Theorem 5 follows from Proposition A3, together with part (i) of Proposition A1. Now, define the common-reasoning model with D as its common standard of practical rationality, as in Section 6, and let R^* be common reason in this model. To establish part (ii) of Theorem 5, we have to show that the propositions in the set $\{p \in T(R^*) \mid p$ is a permissibility proposition$\}$ are precisely those asserted by the final output of the recommendation algorithm for D.

To do this, we define the corresponding inference structure $R\_^*$, as in the proof of Theorem 4. By part (ii) of Lemma A4, the set $\{p \in T(R^*) \mid p$ is a permissibility proposition$\}$ is identical to the set $\{p \in T(R\_^*) \mid p$ is a permissibility proposition$\}$. By parts (i) and (ii) of Lemma A2, the propositions in the latter set are encoded by the categorisation solution for the game relative to ζ, where ζ is the aggregate categorisation function which encodes D. Finally, by Proposition A3, the categorisation solution is identical to the categorisation that encodes the combined final output of the recommendation algorithm. □

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

Lemma A6: For every categorisation C of $\mathbb{S}$: $[C^M \sqsupseteq^* C] \Rightarrow [C^M \sqsupseteq^* \zeta(C)]$.

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

By Lemma A1, 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 ICEU-consistent with C^M; if it is in the negative component of C^M, there is some such distribution for which it is *not* marginally EU-maximising (this is *Result 1*). Now consider any categorisation C of $\mathbb{S}$ such that $C^M \sqsupseteq^* C$. Since $C^M \sqsupseteq^* C$, every probability distribution over S_{-i} that is ICEU-consistent with C^M is also ICEU-consistent with C (this is *Result 2*). Because ζ encodes the profile D of ICEU decision rules, if some strategy $s_i \in S_i$ is in the positive component of $\zeta(C)$, it is marginally EU-maximising for every probability distribution over S_{-i} that is ICEU-consistent with C; if it is in the negative component of $\zeta(C)$, it is marginally EU-maximising for no such distribution (this is *Result 3*). Now suppose Lemma A6 is false. Then, using the fact that, by definition, C^M is exhaustive: *either* (i) for some player i, some strategy $s_i \in S_i$ is in the positive component of C^M and the negative component of $\zeta(C)$, *or* (ii) for some player i, some strategy $s_i \in S_i$ is in the negative component of C^M and the positive component of $\zeta(C)$. Using Results 1, 2 and 3, it can be shown that each of these possibilities implies a contradiction. □

We now complete the proof of the theorem. Trivially, $C^M \sqsupseteq^* <\varnothing, \varnothing>$. By repeated application of Lemma A6, $C^M \sqsupseteq^* \zeta(<\varnothing, \varnothing>)$, $C^M \sqsupseteq^* \zeta\ [\zeta(<\varnothing, \varnothing>)]$, and so on. But, by the initiation and continuation rules for categorisation procedures, $<\varnothing, \varnothing>$, $\zeta(<\varnothing, \varnothing>)$, $\zeta\ [\zeta(<\varnothing, \varnothing>)]$, ... are respectively the categorisations $C(0)$, $C(1)$, $C(2)$, ... induced by the categorisation procedure for ζ. By Proposition A1, this procedure halts at some finite stage k^*. By Proposition A3, Theorem 5 and the definition of the ICEU common-reasoning solution C^*, $C(k^*) = C^*$. Thus, $C^M \sqsupseteq^* C^*$. □

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

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

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

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

References

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

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

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

Aumann, Robert (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.

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.

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

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

Cubitt, Robin P. and Robert Sugden (2010). The reasoning-based expected utility procedure. *Games and Economic Behavior.* Forthcoming and available online as doi: 10.1016/j.geb.2010.04.002

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.

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

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.

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

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

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

Paternotte, Cedric (2010). Being realistic about common knowledge: a Lewisian approach. *Synthese*. Forthcoming and available online as doi: 10.1007/s11229-010-9770-y

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.

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

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.

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