CBESS Discussion Paper 09-04

# The reasoning-based expected utility procedure

**by Robin P Cubitt* and Robert Sugden****

* CeDEx, University of Nottingham,
Nottingham NR7 2RD, robin.cubitt@nottingham.ac.uk
** CBESS, University of East Anglia,
Norwich NR4 7TJ, r.sugden@uea.ac.uk

7 September 2009

**Abstract**
This paper presents a new iterative procedure for solving finite noncooperative games, the reasoning-based expected utility procedure (RBEU), and compares this with existing iterative procedures. RBEU deletes more strategies than iterated deletion of strictly dominated strategies, while avoiding the conceptual problems associated with iterated deletion of weakly dominated strategies. It uses a sequence of “accumulation” and “deletion” operations to categorise strategies as permissible and impermissible; strategies may remain uncategorised when the procedure halts. RBEU and related “categorisation procedures” can be interpreted as tracking successive steps in players’ own reasoning.

**JEL classification codes**
C72

**Short title**
The reasoning-based EU procedure

**Acknowledgements**
We are grateful for comments on earlier versions to a referee and an associate editor; to Giacomo Bonnano, Adam Brandenburger and Michael Mandler; and to participants in various seminars, conferences and workshops at which we have presented the paper. Sugden’s work was supported by the Economic and Social Research Council (award no. RES 051 27 0146).

University of East Anglia

CBESS

Centre for
Behavioural and
Experimental
Social
Science

CBESS
University of East Anglia
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/ssf/cbess

## 1. Introduction

In this paper, we introduce a new iterative solution procedure for noncooperative games, called the *reasoning-based expected utility (RBEU) procedure*. In general, the RBEU procedure deletes more strategies than iterative deletion of strictly dominated strategies (IDSDS), while not coinciding with any of the family of procedures based on iterated deletion of weakly dominated strategies (IDWDS).

The puzzling features of IDWDS are well known. One difficulty is the *order-sensitivity problem*: the conclusions that can be derived by IDWDS are sensitive to the order in which deletions are made. Another, closely related, difficulty is the *undercutting problem*. IDWDS can delete a weakly dominated strategy for some player at one stage of the procedure, only for it to transpire, when further stages of deletion have been undertaken, that that strategy is no longer weakly dominated against the non-deleted strategies of other players.

The RBEU procedure is not vulnerable to any analogous problems. Many of its distinctive properties flow from the fact that, at each stage, it has an operation of *accumulation* of strategies, as well as the more familiar one of deletion. At each stage, deletion and accumulation are sensitive to previous accumulations, as well as to previous deletions. The essential idea is that, at each stage, previously-deleted strategies are assumed to have zero probability, and previously-accumulated strategies are assumed to have strictly positive probability. If a strategy has not yet been deleted or accumulated, no restrictions are imposed on its probability. A strategy is accumulated if, given the assumptions applicable at the relevant stage, it can be shown to maximise expected utility; it is deleted if, given the same assumptions, it can be shown *not* to maximise expected utility. Because RBEU has these two operations, it induces a trinary partition of each player's strategy set at each stage, corresponding to the fact that not being able to establish the falsity of a proposition is not the same thing as being able to establish its truth.

In using the term "reasoning-based" to describe this procedure, we are signalling a particular orientation towards game theory. Many branches of game theory start from the pre-theoretic idea that players have mutual understanding of each other's rationality, and then proceed to represent and develop this idea in different ways, such as the various formal concepts of common knowledge. More specifically, game theory uses two different types of solution concept: equilibrium concepts and iterative procedures. Each type can be motivated in terms of mutual understanding of rationality, but the nature of the motivation can be quite different in the two cases.

For a given game, an equilibrium solution concept defines a set of equilibria, each of which specifies a particular configuration of players' strategy choices and/or beliefs. When such a solution concept is interpreted as embodying mutual understanding of rationality, the implicit claim is that each equilibrium *could be* common knowledge among the players, consistently with the players' rationality also being common knowledge. A game may have more than one equilibrium, in which case the equilibrium approach does not explain how players come to know what other players choose or believe. One way of providing conceptual foundations for an equilibrium solution concept is to show that the relevant equilibrium properties are implied by an epistemic model in which some form of mutual understanding of rationality is represented explicitly; Aumann's (1987) derivation of correlated equilibrium is a classic example. Within the equilibrium-based approach, iterative procedures are sometimes used as devices that help to narrow or assist in a search for particular types of equilibrium. One example is the long-established use of IDSDS to narrow a search for Nash equilibria, exploiting the fact that only strategies which survive IDSDS can have strictly positive probability in any Nash equilibrium.[1] When iterative procedures are used in relation to some kinds of epistemically-grounded solution concepts, the successive stages of strategy deletion may correspond to different levels of belief in a lexicographic probability system, as in the approach to IDWDS analysed by Stahl, 1995; Brandenburger, Friedenberg and Kiesler, 2008; and Asheim and Perea, 2009.

However, an alternative interpretation of iterative procedures leads to a different type of motivation. The successive stages of an iterative procedure can be interpreted as tracking successive steps of reasoning that the players can perform; "rationality" is then interpreted as a property of the modes of reasoning that the players use, and which the procedure tracks. This approach does not merely purport to identify solutions that are consistent with rationality; it also explains how players can know that the solution is what it is. On this understanding, an iterative procedure is not an adjunct to a solution concept that has an independent rationalisation; rather it constitutes in summary form the rationalisation for the solution it generates. This approach is suggested by the understanding of "common knowledge" of rationality formulated by Lewis (1969), in which there is some mode of reasoning that is shared by the players and can be tracked by an iterative procedure.[2] It is in *this* sense that the RBEU procedure is "reasoning-based".

If one thinks of iterative procedures in this way, the order-sensitivity and undercutting problems of IDWDS are troubling. It is difficult to see how two equally valid paths of reasoning from a given set of (mutually consistent) premises, differing only in the order in which inferences were made, could produce mutually inconsistent conclusions. Similarly, it is difficult to see how a conclusion that is reached by valid reasoning from given premises could be undercut by other conclusions derived from the same premises. Thus, intuitively, one might expect that a procedure that tracked players' reasoning would not be subject to the order-sensitivity and undercutting problems. We will argue that each stage of the RBEU procedure can be interpreted as a step of reasoning which each player can make, and that the possibility of this interpretation is licensed by attractive properties of the procedure. Of course, the formal structure of the RBEU procedure is not dependent on this interpretation; we do not discount the possibility that it can also be motivated in other ways.

The remainder of the paper is structured as follows: Section 2 sets up a general framework in which iterative "categorisation procedures", capable of being interpreted as tracking players' reasoning, can be formulated. Section 3 uses this framework to define the RBEU procedure. Section 4 compares the RBEU procedure to IDSDS and IDWDS, and to related procedures in the literature. Section 5 concludes with some brief reflections on other applications of the concept of a categorisation procedure.

## 2. Framework

We consider the class $G$ of finite, normal-form games of complete information, interpreted as one-shot games. Our analysis applies to every such game but, to avoid clutter, we suppress clauses of the form "for all games in $G$" except when stating formal results, and proceed initially by fixing the game. The game is defined by a finite set $N = \{1, ..., n\}$ of *players*, with typical element $i$ and $n \geq 2$; for each player $i$, a finite, non-empty set of (pure) *strategies* $S_i$, with typical element $s_i$; and, for each profile[3] of strategies $s = (s_1, ..., s_n)$, a profile $u(s) = (u_1[s], …, u_n[s])$ of finite *utilities*. The sets $S_1 \times ... \times S_n$ and $S_1 \times ... \times S_{i-1} \times S_{i+1} \times ... \times S_n$ are denoted by $S$ and $S_{-i}$, respectively. We impose that, for all $i, j, \in N$, $S_i \cap S_j = \varnothing$. This condition allows us to use a conveniently compact notation below. It has no substantive significance, but merely imposes a labelling convention that the strategies available to different players are distinguished by player indices, if nothing else.[4]

Our fundamental building block is a formal object which we call a *categorisation* of a player's strategy set. 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 three conditions[5]: (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 $S_i^+ = \{s_i\}$. $S_i^+$ is the *positive component* of the categorisation and $S_i^-$ the *negative component*. The set of categorisations of $S_i$ is denoted $\Phi(S_i)$.

We use two interpretations of this formal object, on each of which a categorisation of $S_i$ represents the substantive content of some statement about player *i*'s strategies. On one interpretation, the statement asserts of the strategies in the positive component that they are rationally permissible for *i* and of the strategies in the negative component that they are not rationally permissible. On the other interpretation, the statement asserts of the strategies in the positive component that *i* might possibly play them and of the strategies in the negative component that that *i* will not play them. In terms of these interpretations, the requirement that $S_i^+$ and $S_i^-$ are disjoint reflects the mutual incompatibility of permissibility and impermissibility (resp. of possibility and impossibility). Condition (ii) requires that not all of *i*'s strategies are categorised as impermissible (resp. as impossible); condition (iii) requires that, if all but one are categorised as impermissible (resp. as impossible), the remaining strategy is categorised as permissible (resp. as possible).

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 set $S_i \backslash (S_i^+ \cup S_i^-)$ containing those strategies that are not classified as elements of either component. If this third set is empty, the categorisation is *exhaustive*. However, our framework permits non-exhaustive categorisations, corresponding to statements which do not refer to all strategies. The *null categorisation* $<\varnothing, \varnothing>$ corresponds to a statement with no substantive content – that is, which says nothing at all about the permissibility or impermissibility (resp. possibility or impossibility) of strategies. This concept will be useful in representing the starting point of an iterative procedure that tracks reasoning.

We now introduce a notation which allows us to combine categorisations of the strategy sets of two or more players, while maintaining the distinction between positive and negative components. Consider any non-empty set $N' \subseteq N$ of players. For each $i \in N'$, let $<S_i^+, S_i^->$ be any categorisation of $S_i$. We define 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 the set $\cup_{i\in N'} S_i$. The set of all categorisations of the latter set is denoted $\Phi(\cup_{i\in N'} S_i)$.[6] Two kinds of combination of categorisation are particularly significant. Members of the first kind arise when $N' = N$; they combine categorisations of all players' strategy sets to produce categorisations of $\cup_{i\in N} S_i$. Members of the second kind arise when $N' = N\backslash\{i\}$; they combine categorisations of the strategy sets of all players except some player $i$ to produce categorisations of $\cup_{i\in N\backslash\{i\}} S_i$.

We use $\mathbb{S}$ as a shorthand notation for $\cup_{i\in N} S_i$; the positive and negative components of categorisations in $\Phi(\mathbb{S})$ will typically be denoted $\mathbb{S}^+$ and $\mathbb{S}^-$. Similarly, we use $\mathbb{S}_{-i}$ as a shorthand notation for $\cup_{\mathrm{i}\in\mathrm{N}\backslash\{\mathrm{i}\}} S_i$; the positive and negative components of categorisations in $\Phi(\mathbb{S}_{-i})$ will typically be denoted $\mathbb{S}_{-i}^+$ and $\mathbb{S}_{-i}^-$. For compactness, we will sometimes also use an even briefer notation whereby $C$, $C'$, and so on, denote particular categorisations of $\mathbb{S}$; $C_{-i}$, $C_{-i}'$, and so on, particular categorisations of $\mathbb{S}_{-i}$; and $C_i$, $C_i'$, and so on, particular categorisations of $S_i$.

Consider any categorisations $C_i' = <S_i^{+\prime}, S_i^{-\prime}>$, $C_i'' = <S_i^{+\prime\prime}, S_i^{-\prime\prime}>$ in $\Phi(S_i)$, for some player $i$. We define a binary relation $\sqsupseteq^*$ (read as *has weakly more content than*) between such categorisations such that $C_i'' \sqsupseteq^* C_i'$ if and only if $S_i^{+\prime\prime} \supseteq S_i^{+\prime}$ and $S_i^{-\prime\prime} \supseteq S_i^{-\prime}$. If, in addition, either $S_i^{+\prime\prime} \supset S_i^{+\prime}$ or $S_i^{-\prime\prime} \supset S_i^{-\prime}$ holds, we will say that $C_i''$ *has strictly more content than* $C_i'$, denoted $C_i'' \sqsupset^* C_i'$. On the reading of categorisations as statements about permissibility (resp. possibility), $C_i'' \sqsupset^* C_i'$ has a natural interpretation: it indicates that the statement represented by $C_i''$ asserts everything that is asserted by the statement represented by $C_i'$ and more besides. This notation is extended in an obvious way to categorisations of $\Phi(\mathbb{S})$ and $\Phi(\mathbb{S}_{-i})$. For example, consider categorisations $C' = <\mathbb{S}^{+\prime}, \mathbb{S}^{-\prime}>$, $C'' = <\mathbb{S}^{+\prime\prime}, \mathbb{S}^{-\prime\prime}>$ in $\Phi(\mathbb{S})$. In this case, $C'' \sqsupseteq^* C'$ if $\mathbb{S}^{+\prime\prime} \supseteq \mathbb{S}^{-\prime}$ and $\mathbb{S}^{+\prime\prime} \supseteq \mathbb{S}^{-\prime}$.

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

A categorisation function may be interpreted as encoding reasoning which produces categorisations of $S_i$, conditional on categorisations of $\mathbb{S}_{-i}$. On this reading, the categorisations of $S_i$ attribute permissibility and impermissibility to strategies available to $i$; and the categorisations of $\mathbb{S}_{-i}$ on which they are conditioned attribute possibility and impossibility to strategies available to players other than $i$. Thus, a particular categorisation function $f_i$ corresponds to a form of reasoning that generates statements about what is rationally permissible for player $i$, conditional on statements about what other players might do. In this sense, $f_i$ encodes a conception of *practical rationality* for player $i$. Monotonicity corresponds to the requirement on rational reasoning that any conclusions that can be obtained from a given set of premises can also be obtained from any strictly stronger set of premises.[7] As we will show, imposition of this requirement is a crucial feature of our approach.

We will work below with profiles of categorisation functions. It will simplify our subsequent definitions to express the content of a given profile $f = (f_1, \ldots, f_n)$ of categorisation functions as a single function $\zeta\colon \Phi(\mathbb{S}) \to \Phi(\mathbb{S})$, constructed as follows. Let $C = <\mathbb{S}^+, \mathbb{S}^->$ be any categorisation of $\mathbb{S}$. For each player $i$, define $C_{-i} = <\mathbb{S}^+ \backslash \mathrm{S}_i, \mathbb{S}^- \backslash \mathrm{S}_i>$. 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 $\zeta$ *summarises f*. A function $\zeta\colon \Phi(\mathbb{S}) \to \Phi(\mathbb{S})$ that summarises some profile $f$ of categorisation functions is an *aggregate categorisation function*. For a given profile $f$, there is exactly one function $\zeta$ which summarises it.

We are now in a position to define the central concept of this section. For any aggregate categorisation function $\zeta$, the *categorisation procedure* is defined 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:

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

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

This formal definition can be expressed more loosely as follows: the first stage of the procedure applies the function $\zeta$ to the null categorisation $<\varnothing, \varnothing>$; and then, each subsequent stage applies the function $\zeta$ to the output of the previous stage. Obviously, if there exists $k' \in \{1, 2, \ldots..\}$ such that $C(k') = C(k'-1)$ then, for all $k'' > k'$, $C(k'') = C(k')$. Since this renders further application of the continuation rule redundant, we will say that the procedure *halts* at the lowest value of $k'$ for which $C(k') = C(k'-1)$; this value of $k'$ will be denoted by $k^*$. Then, $C(k^*)$ is the *categorisation solution* of the game, relative to $\zeta$.

Though formally not essential, in our applications of the concept of a categorisation function, we will in fact always attribute the same conception of practical rationality to all players. Since this conception is embedded in a particular $\zeta$, the categorisation procedure for that $\zeta$ can be interpreted as tracking a sequence of phases of reasoning based on this conception. The first phase starts with no substantive premises (the absence of such premises being represented by the null categorisation) and reaches conclusions about the permissibility or impermissibility of strategies, represented by the categorisation $C(1)$.

These conclusions are unconditional implications of the underlying conception of practical rationality. For the second phase, the strategies whose permissibility (resp. impermissibility) was established in the first phase are taken as possible (resp. as impossible). These transitions from (im)permissibility to (im)possibility can be interpreted as tracking inferences that players can make, given that they attribute rationality to one another. Further conclusions about permissibility and impermissibility, captured by the categorisation $C(2)$, can now be drawn. And so on.

In this way, the categorisation procedure tracks the reasoning of all players. This idea corresponds to the informal notion that players have mutual awareness of each other's rationality. More formally, it can be seen as reflecting a Lewisian conception of common knowledge of rationality whereby there is some mode of actual reasoning that is accessible to all players (Lewis, 1969); our interpretation is that this is the mode of reasoning that is tracked by the categorisation procedure.

Some important properties of a categorisation procedure can be seen to flow from the definition of the concept, even without specifying a particular function ζ. These properties are encapsulated in the following result (proved in the appendix).

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

Part (b) of Proposition 1 guarantees the existence of a categorisation solution for any ζ. Part (a) is used to prove part (b), but is also significant in its own right, in terms of the interpretation of a categorisation procedure as tracking a reasoning process. On that reading, it shows that each phase of reasoning reaffirms the classifications made by previous phases. We call this the *reaffirmation* property.

The intuition for Proposition 1 is straightforward but important. By the Initiation rule, the sequence of categorisations generated by the categorisation procedure begins with $C(0) = <\varnothing, \varnothing>$. Thus, trivially, $C(1) \supseteq^* C(0)$. Since $C(2) = \zeta[C(1)]$ and $C(1) = \zeta[C(0)]$, the fact that ζ is an aggregate categorisation function (and so summarises a profile of categorisation functions each of which satisfies Monotonicity), implies that $C(2) \supseteq^* C(1)$. And so on. Each time the procedure generates a categorisation with weakly more content than the previous one, a corresponding use of the definition of ζ forces the next categorisation generated by the procedure to have at least weakly more content again. Eventually the procedure must halt, if only because the game is finite; but, up to that stage, the procedure generates categorisations with strictly more content at each successive stage.

The categorisation procedure for a given ζ can be interpreted as a process which constructs the categorisation solution by successive "accumulations" (additions to the set of strategies that have been found to be permissible) and "deletions" (additions to the set of strategies that have been found to be impermissible). For each $k > 0$, we will say that strategies in $\mathbb{S}^+(k)\backslash\mathbb{S}^+(k–1)$ are *accumulated at stage k*, and that strategies in $\mathbb{S}^-(k)\backslash\mathbb{S}^-(k–1)$ are *deleted at stage k*.

The properties of categorisation procedures presented in this section have special significance in relation to the undercutting and order-sensitivity problems that can arise for some iterative procedures, notably IDWDS. There is undercutting when an operation (in IDWDS, the deletion of a strategy) is made at one stage of an iterative procedure, but the justification for that operation is invalidated at a later stage. Undercutting can give rise to order-sensitivity: if two operations are both valid at a given stage, carrying out only one of them may invalidate the justification for the other. Conversely, if there is no undercutting, an operation that becomes valid at some stage remains valid until it is carried out, irrespective of which other valid operations are carried out in the interim.

Because of the reaffirmation property, categorisation procedures are immune to undercutting. The profile of categorisation functions summarised by ζ captures the rationale for the deletions (and accumulations) made for each player at each stage. The requirement that each player i's categorisation function satisfy Monotonicity imposes the discipline that whatever statements about permissibility and impermissibility of player *i*'s strategies are warranted, given the categorisation generated by the procedure at stage *k*, are still warranted, given the output of subsequent stages.

To analyse order-sensitivity, we introduce the concept of a "potentially negligent" variant of a categorisation procedure. The idea is that, where the categorisation procedure specifies a set of deletions and accumulations at a given stage *k*, a potentially negligent variant might make only some of them at that stage. This is analogous to what is often regarded as legitimate variation in the specification of IDSDS and IDWDS. Suppose that, at some stage in a procedure of iterative deletion of dominated strategies, two or more strategies are dominated. Must all of these strategies be deleted simultaneously? Or should each deletion of a single strategy count as a separate "stage" – and if so, which strategy should be deleted first? The order-sensitivity problem of IDWDS is that which strategies ultimately survive the procedure can depend on how these specification questions are answered. Thus, in our context, order-*in*sensitivity can be represented as the requirement that all potentially negligent variants of a categorisation procedure reach the same final output as the categorisation procedure itself.

Consider any game in $G$; and let $\zeta$ be any aggregate categorisation function for the game and CP($\zeta$) be the categorisation procedure for $\zeta$. An iterative procedure IP($\zeta$) is a *potentially negligent variant* of CP($\zeta$) if it generates a sequence of categorisations $C'(k)$ of $\mathbb{S}$ for successive stages $k \in \{0, 1, 2, \ldots.\}$ that satisfy (i) $C'(0) = <\varnothing, \varnothing>$; and, for all $k > 0$, (ii) $\zeta[C'(k-1)] \supseteq^* C'(k)$; (iii) if $\zeta[C'(k-1)] \supset^* C'(k-1)$, then $C'(k) \supset^* C'(k-1)$; and (iv) if $\zeta[C'(k-1)] = C'(k-1)$, then $C'(k) = C'(k-1)$. IP($\zeta$) *halts* at the lowest value of $k'$ for which $C'(k') = C'(k'-1)$; this value of $k'$ will be denoted $k^{**}$.

For intuition, consider any stage $k$. Think of application of $\zeta$ to the previous categorisation $C(k-1)$ as defining a maximal set of "instructions" for the deletion and accumulation of strategies. In CP($\zeta$), these instructions are fully carried out at stage $k$; but in IP($\zeta$), some instructions may be neglected. Condition (i) requires IP($\zeta$) to begin with the null categorisation, as CP($\zeta$) does. Condition (ii) requires that, though IP($\zeta$) may be negligent, it never deletes or accumulates a strategy at any stage $k$ unless it has been instructed to do so at that stage. Condition (iii) requires that, if the instructions at stage $k$ are to make some new deletions and/or accumulations (and not to undo any existing ones), then IP($\zeta$) makes at least some of these at stage $k$. Condition (iv) requires that, if the instructions at stage $k$ are just to repeat the previous output, then IP($\zeta$) does so, thereby halting.

The following result is proved in the appendix:

> *Proposition 2*: Consider any game in $G$ and any aggregate categorisation function $\zeta$ for the game. Let CP($\zeta$) be the categorisation procedure for $\zeta$ and $C(k^*)$ be the corresponding categorisation solution. Let IP($\zeta$) be any potentially negligent variant of CP($\zeta$) and $C'(0)$, $C'(1)$, …., be the sequence of categorisations generated by IP($\zeta$). Then:
>
> (a) For all $k \in \{1, 2, \ldots.\}$, $C'(k') \supseteq^* C'(k'-1)$.
>
> (b) IP($\zeta$) halts at some $k^{**} \in \{1, 2, \ldots.\}$.
>
> (c) $C'(k^{**}) = C(k^*)$.

This proposition shows that for every potentially negligent variant of CP($\zeta$), the following is true: it has the reaffirmation property; it halts; and when it halts, the categorisation it has generated coincides with the categorisation solution generated by CP($\zeta$) itself. Put more loosely, it makes no difference to the final output of a categorisation procedure if some deletions or accumulations are omitted at some stage(s), as long as *some* such operations are carried out whenever *any* are warranted.

## 3. The reasoning-based EU categorisation procedure

We are now able to define our RBEU procedure. This is an instance of the more general concept of a categorisation procedure, introduced in Section 2.

In order to define a particular categorisation procedure, we have only to define a profile of categorisation functions for the players (which will, in turn, define the function $\zeta$). We require, for each player $i$, a function that maps $\Phi(\mathbb{S}_{-i})$ to $\Phi(S_i)$; and, crucially, which satisfies Monotonicity.

On our interpretation, specifying this function corresponds to specifying a conception of practical rationality, for each player $i$. To do this, we need criteria of rational permissibility and impermissibility of $i$'s strategies, conditional on any categorisation $C_{-i}$ of $S_{-i}$. The approach we adopt is orthodox, in the sense of being based on expected utility maximisation. We proceed in two steps. Intuitively, the first step defines a rule to indicate which probability distributions over $S_{-i}$ are "allowable", given a categorisation $C_{-i}$ of $\mathbb{S}_{-i}$. The second step defines a rule for assigning strategies to the positive and negative components of a categorisation of $S_i$, given a set of allowable probability distributions over $S_{-i}$. To formalise these concepts, we use $\Delta(S_{-i})$ to denote the set of probability distributions over $S_{-i}$.

An *allowability rule* for player $i$ associates a non-empty subset $A(C_{-i})$ of $\Delta(S_{-i})$ with each categorisation $C_{-i}$ in $\Phi(\mathbb{S}_{-i})$. The *reasoning-based* allowability rule is defined by the requirement that a probability distribution is allowable if and only if it satisfies the following two conditions:[8]

*Positive sub-rule*: Each strategy in the positive component of $C_{-i}$ has strictly positive marginal probability.

*Negative sub-rule*: Each strategy in the negative component of $C_{-i}$ has zero marginal probability.

The *EU assignment rule* for player $i$ comprises the following pair of sub-rules for specifying $S_i^+(A) \subseteq S_i$ and $S_i^-(A) \subseteq S_i$, conditional on any non-empty set $A \subseteq \Delta(S_{-i})$ of allowable probability distributions:

> *Positive sub-rule*: $S_i^+(A) = \{s_i \in S_i \mid s_i$ is expected utility maximising for $i$, for every probability distribution in $A\}$;
>
> *Negative sub-rule*: $S_i^-(A) = \{s_i \in S_i \mid s_i$ is not expected utility maximising for $i$, for any probability distribution in $A\}$.

This specification guarantees that, for any non-empty $A \subseteq \Delta(S_{-i})$, the ordered pair $<S_i^+(A)$, $\mathrm{S_i}^-(A)>$ satisfies parts (i) – (iii) of the definition of a categorisation of $S_i$. Thus, the EU assignment rule for player $i$ associates a categorisation of $S_i$ with each non-empty set of allowable probability distributions over $S_{-i}$. In general, this categorisation may or may not be exhaustive; it is non-exhaustive when there are strategies in $S_i$ that maximise player $i$'s expected utility for some, but not all, probability distributions in $A$.

The reasoning-based allowability rule is a particular function from $\Phi(\mathbb{S}_{-i})$ to the set of non-empty subsets of $\Delta(S_{-i})$; and the EU assignment rule is a particular function from the latter set to $\Phi(\mathbb{S}_{-i})$. Thus, the composition of these two functions is a particular function $f_i$: $\Phi(\mathbb{S}_{-i}) \rightarrow \Phi(S_i)$. We call this function the *reasoning-based expected utility (RBEU) categorisation function* for player $i$, anticipating the following result:

> *Proposition 3*: Consider any game in $G$ and any player $i$ of the game. The composition $f_i$ of the reasoning-based allowability rule for $i$ and the EU assignment rule for $i$ is a categorisation function for player $i$.

To prove this proposition, it is sufficient to establish that $f_i$ satisfies Monotonicity. That is, we must show that if some categorisation $C_{-i}''$ has strictly more content than another categorisation $C_{-i}'$, then $f_i(C_{-i}'')$ has weakly more content than $f_i(C_{-i}')$. To see that this is the case, note that as the content of $C_{-i}$ increases, the reasoning-based allowability rule imposes (strictly) tighter restrictions on the set $A$ of allowable probability distributions. This makes it "easier" for a strategy to be expected utility maximising for *all* allowable distributions (and so to be assigned to $S_i^+(A)$ by the EU assignment rule); and also "easier" to be expected utility maximising for *no* such distributions (and so to be assigned to $S_i^-(A)$ by the EU assignment rule).

For any game in $G$, the profile of RBEU categorisation functions is summarised by a unique aggregate categorisation function. The *RBEU procedure* (henceforth RBEU) is the categorisation procedure for this aggregate categorisation function. Since RBEU is a categorisation procedure, Proposition 1 applies to it. Thus, RBEU induces a unique categorisation solution, the *RBEU solution*.

As an initial illustration, consider the following game:

*Game 1:*

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

For this game, RBEU runs as follows: *C*(0) = <∅, ∅>; *C*(1) = <{*left*}, {*fourth*}>; *C*(2) = <{*left, right*}, {*second*, *fourth*}>; *C*(3) = *C*(2). Thus, the RBEU solution of the game is <{*left, right*}, {*second*, *fourth*}>. In words, RBEU accumulates *left* and deletes *fourth* at stage 1; then, at stage 2, accumulates *right* and deletes *second*; and then halts.

Game 1 illustrates several features of RBEU. First, RBEU can accumulate strategies as well as deleting them. Second, accumulation and deletion both feed on the conclusions of prior stages; and each can feed on the other. It is the deletion of *fourth* at stage 1 that allows the accumulation of *right* at stage 2; and it is the accumulation of *left* at stage 1 that allows the deletion of *second* at stage 2. The fact that a subsequent deletion can be driven by an earlier accumulation shows that the concept of accumulation has real bite in RBEU; it is not mere semantics. Third, the procedure can halt with some strategies neither accumulated nor deleted. Thus, in general, the RBEU solution induces a trinary partition of strategies.

This implies that RBEU may distinguish between two classes of undeleted strategy: those accumulated (*left* and *right* in the example) and those neither accumulated nor deleted (*first* and *third*). To say that a strategy has not been deleted is to say that it is optimal for *some* beliefs that have not been definitely ruled out; to say that it has been accumulated is to make the stronger statement that it is optimal for *all* such beliefs. Or, interpreting RBEU as tracking a process of reasoning: to say that a strategy has not been deleted is to say that it *has not been shown to be impermissible*; to say that is has been accumulated is to say that is *has been shown to be permissible*.

## 4. The RBEU procedure compared to others

In this section, we compare RBEU to existing iterative procedures, continuing to confine our attention to finite games.[9]

### *(i) IDSDS*

Since RBEU uses the concept of accumulation while IDSDS does not, it is obvious that the two procedures do not coincide. But we may usefully compare the *deletions* they make.

We begin with a very simple example which shows that RBEU can delete strategies that IDSDS does not:

*Game 2:*

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

In this game, IDSDS does not delete any strategies. In contrast, RBEU runs as follows: $C(0) = <\emptyset, \emptyset>$; $C(1) = <\{\textit{first, left}\}, \emptyset>$; $C(2) = <\{\textit{first, left}\}, \{\textit{second, right}\}>$; $C(3) = C(2)$. That is, *first* and *left* are optimal for all beliefs, and so are accumulated at stage 1; *second* and *right* are then deleted at stage 2, on the grounds that they are not optimal for any beliefs which assign strictly positive probability to *left* and *first*. Intuitively, the strategies that are deleted by RBEU but not IDSDS in this game seem very unattractive.

We now show that RBEU deletes every strategy that is deleted by IDSDS. For this purpose, it is convenient to describe (maximal) IDSDS in terms of "allowable" probability distributions. Consider any game in *G*. IDSDS deletes strategies in a series of stages $k = 1, 2, \ldots$ . At each stage $k > 1$ of IDSDS, there is for each player *i* a set $D_i(k–1) \subseteq S_i$, containing strategies for player *i* that have been deleted in previous stages. We set $D_i(0) = \emptyset$. We define $D_{–i}(k–1) \equiv \cup_{j \in N \setminus \{i\}} \mathrm{D}_j(k–1)$ and $D(k–1) \equiv \cup_{i \in N} D_i(k–1)$. For each stage $k \geq 1$, and for each player *i*, let $A[D_{–i}(k–1)] \subseteq \Delta(S_{–i})$ be the set of probability distributions over $S_{–i}$ which assign zero marginal probability to every element of $D_{–i}(k–1)$. The deletion operation of IDSDS can then be expressed as the rule that a strategy for player *i* (if not already deleted) is deleted at stage *k* if and only if it is not expected utility maximising for any probability distribution in $A[D_{–i}(k–1)]$.[10] $D(k)$ can then be defined as the union of $D(k–1)$ and the set of strategies deleted at stage *k*. The procedure halts at the first *k* at which $D(k) = D(k–1)$.

This formulation shows that, *as far as deletions are concerned*, IDSDS and RBEU differ only in that, for given previous deletions, RBEU imposes tighter restrictions on allowable beliefs at each stage *k*. (Both procedures require that previously deleted strategies have zero probability, but RBEU imposes the additional condition that previously accumulated strategies have strictly positive probability.) Thus, in general, RBEU makes it "easier" for a strategy to be expected utility maximising for no allowable beliefs. Hence, every strategy that is deleted by IDSDS is also deleted by RBEU.

It is well known that, in finite games, IDSDS does not exhibit the order-sensitivity and undercutting problems. As this attractive property of IDSDS is also a property of all categorisation procedures, it is natural to ask whether it is possible to define a categorisation procedure which makes exactly the same deletions as IDSDS. As we now illustrate, this *is* possible. Recall that, for each player, the RBEU categorisation function is the composition of the reasoning-based allowability rule and the EU assignment rule. Consider the categorisation procedure which differs from RBEU in only one respect, namely the removal, for each player *i*, of the positive sub-rule of the reasoning-based allowability rule. This amendment has the effect of making deletions at each stage dependent only on previous deletions; although strategies can be accumulated, accumulations have no implications for subsequent operations of deletion. It is easy to show that, at each stage, this variant procedure makes exactly the same deletions as IDSDS.[11]

*(ii) IDWDS*

We now compare RBEU to IDWDS. We use the term "IDWDS" to refer to the family of iterative procedures in which weakly dominated strategies are successively deleted. Because of the order-sensitivity problem, an IDWDS procedure is not fully specified unless the order in which deletions are made is defined. The most common such specification is *maximal* IDWDS – that is, at each stage, *all* strategies that are weakly dominated at that stage are deleted together. However, most of the conclusions that we will derive in this sub-section apply to all forms of IDWDS. For the same reason as in discussion of IDSDS, we focus on comparison of RBEU and IDWDS in terms of deletions.

It is apparent from Game 2 that there are some games in which IDWDS and RBEU delete precisely the same strategies. However, even when this is so, they do not always delete them for the same reasons. In Game 2, irrespective of the order of deletion, IDWDS deletes both *second* and *right* on grounds of weak dominance, because both are weakly dominated in the initial game and each remains dominated if the other is deleted; *first* and *left* remain as an undeleted residual. In contrast, in RBEU, *second* and *right* are deleted only after *first* and *left* have been accumulated. Interpreting RBEU as tracking a reasoning process, we can say, for example, that it is only after player 1 has established that *left* is permissible for player 2 that she can conclude that she must assign strictly positive probability to that strategy, and hence that *second* is impermissible for her.

Although RBEU and IDWDS delete the same strategies in Game 2, there are many games where this is not so. We first show that, even if its order of deletion is uniquely determined, IDWDS may delete strategies that RBEU does not delete. Our first example of this is a game discussed by Samuelson (1992, esp. pp. 304-5), which provides perhaps the simplest possible illustration of the undercutting problem.[12]

*Game 3:*

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

For this game, all IDWDS procedures delete *second* (and nothing else) at the first stage, followed by *right* at the second stage, leaving *first* and *left* undeleted. But the deletion of *right* undercuts the reason for the earlier deletion of *second*. That is, if *right* will not be played, there seems no reason to discard *second*, which is a best reply to *left*. In contrast, when RBEU is applied to this game, it accumulates *first* at the first stage, but does not delete any strategy. The reason for accumulating *first* is that it is optimal for all beliefs; the reason for *not* deleting *second* is that such deletion is not required unless *right* is accumulated first. But *right* is never accumulated, as it is not a best reply to *first,* which is accumulated at the first stage.

It is interesting to note what happens if player 2's best replies are transposed, as in:

*Game 4:*

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

In this game, all IDWDS procedures delete *second* (and nothing else) at the first stage, and then *left* at the second stage, leaving *first* and *right* undeleted. In this case, IDWDS is not subject to an undercutting problem: to the contrary, the deletion of *left* "confirms" the deletion of *second* by making the latter strictly dominated. But this kind of ex post "confirmation" is incompatible with a reasoning-based interpretation of iterative procedures, since the latter requires deletions (and accumulations) to be justified at the stage at which they are made. RBEU cannot delete *second* unless *right* has previously been accumulated; and it never is. RBEU just accumulates *first* and then halts.

Can RBEU delete strategies that are not deleted by IDWDS? The answer depends on how the order of deletion under IDWDS is specified, as we will explain.

A general result is that, in any given game, all the deletions made by RBEU would also be made by IDWDS *under some order of deletions*. Specifically, deletions made in the same order as they are made by RBEU are always consistent with IDWDS. To demonstrate this, we describe IDWDS in terms of "allowable" probability distributions.

Consider a sequence of stages $k = 1, 2, ...$ at which deletions are made in accordance with IDWDS. For each stage $k$, let $D_i(k)$ and $D_{-i}(k)$ be defined as in our reformulation of IDSDS. For each $k \geq 1$, the set of allowable beliefs for each player $i$, under IDWDS, is the set of probability distributions that assign zero probability to every element of $D_{-i}(k-1)$ *and strictly positive probability to every other strategy*. (The presence of the italicised clause distinguishes IDWDS from IDSDS.) It is common to all IDWDS procedures that a strategy for player $i$ may be deleted at stage $k$ *only if* it is not expected utility maximising for any allowable probability distribution. If this criterion permits any deletions at a given stage, at least one permitted deletion is made. Notice that, at $k = 1$, IDWDS imposes tighter restrictions on allowable beliefs than RBEU does. Thus, any deletions made by RBEU at this stage are also permitted by IDWDS. Now suppose that the deletions made at $k = 1$ are exactly those required by RBEU. The argument can then be repeated: at $k = 2$, any deletions made by RBEU are also permitted by IDWDS; and so on. Thus, the sequence of deletions made by RBEU coincides with *one possible* sequence of IDWDS *up to the stage at which the RBEU procedure halts.* But there may be no sequence of IDWDS that stops deleting when RBEU does. (In Games 3 and 4, for example, RBEU halts without deleting anything, but all IDWDS procedures delete some strategies.)

However, as is shown by Game 5 below, *specific* IDWDS procedures may fail to delete strategies which RBEU does delete. (The argument of the previous paragraph implies that this eventuality can only arise in a game with an order-sensitivity problem for IDWDS. It is easy to see that Game 5 has this feature.)

*Game 5:*

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

We compare RBEU with maximal IDWDS. Maximal IDWDS deletes *centre* and *right*, then *third*, leaving *first*, *second*, and *left* undeleted. RBEU accumulates *left* and deletes *right*;

then accumulates *first* and deletes *third*; then accumulates *centre*;[13] and finally deletes *second*. Thus, *second* is deleted by RBEU but not by maximal IDWDS. RBEU is eventually able to delete *second* because it has previously accumulated *centre*. It was able to accumulate *centre* because it had previously deleted *third*. However, before any strategies are deleted, *centre* is weakly dominated, and so is deleted immediately by maximal IDWDS. This is another example of undercutting: the initial justification for deleting *centre* is undercut by the later deletion of *third*.

Unlike IDWDS, RBEU is not vulnerable to order-sensitivity and undercutting problems. Formally, this is an implication of the fact that RBEU is a categorisation procedure (see Section 2). More intuitively, it is a product of the distinction between accumulation and non-deletion. At each stage of RBEU, players are required to assign zero probability to previously deleted strategies and strictly positive probability to *previously accumulated* strategies. In contrast, at a given stage of IDWDS, players are implicitly required to assign zero probability to previously deleted strategies and strictly positive probability to strategies *that have not yet been deleted*. This difference is crucial. Strategies which have been accumulated in RBEU could never be deleted later, and so the rationale for requiring strictly positive probability on them is secure. In contrast, strategies which have not been deleted at a given stage of IDWDS may still be deleted at a later stage, so the case for requiring strictly positive probability on them can be undercut.

### *(iii) The Dekel–Fudenberg procedure*

Dekel and Fudenberg (1990) propose an iterative procedure which combines elements of IDSDS and maximal IDWDS. In this procedure (which we denote DF), there is one stage of maximal deletion of weakly dominated strategies, followed by IDSDS on the game that remains. In some games, DF deletes strategies that RBEU fails to delete (or even accumulates); in others, RBEU deletes strategies that DF does not. Game 5 illustrates both these possibilities. In this game, DF coincides exactly with maximal IDWDS; DF deletes *centre* but not *second*, while RBEU deletes *second* and accumulates *centre*. The deletion of *centre* (and more generally, the deletion of every strategy that would be weakly dominated in the absence of any deletions) might be justified as a principle of "caution", if caution is understood as requiring some non-zero degree of belief, even if only at some level in a lexicographic probability system, to be assigned to *every* possible combination of one's opponents' strategies.[14] However, RBEU rests on a stronger interpretation of the players' mutual understanding of rationality, according to which strategies that can be shown to be impermissible are assigned zero probability. Thus, having deleted *third*, RBEU can accumulate *centre*, even though the latter strategy is weakly dominated in the game as a whole.

(iv) *The Asheim-Dufwenberg procedure*

Asheim and Dufwenberg (2003) present a procedure of *iterative elimination of choice sets*. This begins, for each player $i$, with the collection of non-empty subsets of $S_i$ ("choice sets") and then iteratively deletes elements from this collection. Thus, at each stage, the Asheim-Dufwenberg procedure (henceforth AD) generates, for each player $i$, a collection of so-far surviving choice sets for $i$. The final output of the procedure is a binary partition of the collection of choice sets: each such set is either eliminated or not.

Indirectly, however, AD induces a trinary partition of each strategy set $S_i$. One element of this partition, which we may denote $S_i^+$, contains those strategies that are members of *all* surviving choice sets for player $i$; the second element $S_i^-$ contains those strategies that are members of *no* such sets; the third element is the residual. There is an obvious sense in which the elements of $S_i^+$ have been categorised as "permissible" and the elements of $S_i^-$ as "impermissible".[15] One might ask whether, given this interpretation, AD coincides with RBEU.

The answer is that it does not. The specification of AD is such that if a strategy $s_i$ is weakly dominated, it cannot be an element of any surviving choice set – that is, in the terms used in the previous paragraph, it is assigned to $S_i^-$. The corresponding property does not hold in general for RBEU. Consider the following game (which is $G_3$ of Asheim and Dufwenberg (2003)):

*Game 6:*

| | | *Player 2* | |
|---|---|---|---|
| | | *left* | *right* |
| *Player 1* | *first* | 1, 1 | 1, 1 |
| | *second* | 1, 1 | 1, 0 |
| | *third* | 1, 0 | 0, 1 |

Asheim and Dufwenberg (2003, pp. 211, 214) show that their procedure first deletes all choice sets for player 1 except {*first*, *second*}. Then, it deletes all choice sets for player 2 except {*left*}. No more deletions are possible. Thus, $S_1^+$ = {*first*, *second*}, $S_1^-$ = {*third*}, $S_2^+$ = {*left*}, $S_2^-$ = {*right*}. Notice that *third*, which is weakly dominated, has been categorised as impermissible. RBEU does not have this implication: it accumulates *first* and *second* and then halts.

## 5. Conclusion

We have argued that RBEU has a novel and attractive combination of properties, including ability to delete more strategies than IDSDS, order-insensitivity, and the absence of an undercutting problem. Its possession of these properties is intimately connected with its capacity to be interpreted as tracking successive steps of reasoning that can be carried out by the players.

This capacity is induced by a fundamental feature of RBEU, namely that it is a "categorisation procedure". RBEU is of particular interest because of the analogies and disanalogies between it and IDWDS. However, other categorisation procedures may be of interest too. To define such a procedure, what is required is to specify an aggregate categorisation function. Our analysis in Section 3 illustrates a recipe for achieving this, using the concept of a set of probability distributions that are allowable for player *i*, given a categorisation of other players' strategies. In this set-up, two ingredients provide the key (by being jointly sufficient for the crucial Monotonicity property of a categorisation function for player *i*). The first is that, as the categorisation relative to which they are defined acquires strictly more content, the rules defining the allowable probability distributions tighten. The second is that, when this happens, the rules assigning strategies to the positive component of the resulting categorisation of *i*'s strategies become easier to satisfy, and likewise for the negative component. Each of these ingredients is consistent with a variety of modifications of the sub-rules that define RBEU. For example, we have described a categorisation procedure which makes the same deletions as IDSDS but which also has an accumulation operation. Another possibility is to impose on RBEU the additional restriction that, in allowable probability distributions, the probabilities assigned to the strategies of different players should be independent.[16] Finally, a more radical possibility would be to substitute some other theory of choice under uncertainty for expected utility theory. This could be done by replacing "is expected-utility maximising" in the assignment sub-rules with some other predicate defined relative to probability distributions. For example, rank-dependent expected utility theory, in which probabilities are transformed non-linearly into "decision weights" (Quiggin, 1982) could be used in place of conventional expected utility theory as the underlying conception of “rational” choice.  We suggest that the concept of a categorisation procedure provides a general theoretical framework for the development and investigation of reasoning-based iterative procedures.

## Appendix : Proofs of Propositions 1 and 2

It is convenient to begin with the following lemma:

> *Lemma*: Consider any game in $G$ and any profile $f = (f_1, \ldots, f_n)$ of categorisation functions for its players. Let $\zeta$ be the aggregate categorisation function that summarises $f$. $\zeta$ has the following property: for all $C'$, $C'' \in \Phi(\mathbb{S})$, if $C'' \sqsupseteq^* C'$ then $\zeta(C'') \sqsupseteq^* \zeta(C')$.

*Proof of Lemma*: Fix any game in $G$ and any profile $f = (f_1, \ldots, f_n)$ of categorisation functions for its players. Let $\zeta$ be the aggregate categorisation function that summarises $f$. Consider any $C'$, $C'' \in \Phi(\mathbb{S})$; and let $\mathbb{S}^{+\prime\prime}$ and $\mathbb{S}^{-\prime\prime}$ be, respectively, the positive and negative components of $C''$; and $\mathbb{S}^{+\prime}$ and $\mathbb{S}^{-\prime}$ be, respectively, the positive and negative components of $C'$. First, suppose that $C'' = C'$. Then, it is immediate, from the construction of $\zeta$, that $\zeta(C'') = \zeta(C')$. Now, suppose that $C'' \sqsupset^* C'$. There must exist a unique and non-empty subset $N'$ of $N$ such that (i) for all $i \in N'$, $\mathbb{S}^{+\prime\prime} \backslash S_i \sqsupseteq \mathbb{S}^{+\prime} \backslash S_i$ and $\mathbb{S}^{-\prime\prime} \backslash S_i \sqsupseteq \mathbb{S}^{-\prime} \backslash S_i$, with at least one of those two superset relationships strict; and (ii) for any $i \in N \backslash N'$, $\mathbb{S}^{+\prime\prime} \backslash S_i = \mathbb{S}^{+\prime} \backslash S_i$ and $\mathbb{S}^{-\prime\prime} \backslash S_i = \mathbb{S}^{-\prime} \backslash S_i$. Thus, from the construction of $\zeta$ and the fact that, for all $i \in N'$, $f_i$ satisfies Monotonicity, $\zeta(C'') \sqsupseteq^* \zeta(C')$. □

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

*Proof of Proposition 2*:

In this proof, reference to rules (i), (ii), (iii) and (iv) are to the rules defining a potentially-negligent variant of a categorisation procedure.

*Proof of (a)*: Since $C'(1) \sqsupseteq^* C'(0)$ follows trivially from rule (i), it is sufficient to prove that, for all $k \in \{0, 1, 2, \ldots.\}$, $C'(k+1) \sqsupseteq^* C'(k)$ implies $C'(k+2) \sqsupseteq^* C'(k+1)$. Suppose $C'(k+1) \sqsupseteq^* C'(k)$. By the Lemma, $\zeta[C'(k+1)] \sqsupseteq^* \zeta[C'(k)]$. By rule (ii), $\zeta[C'(k)] \sqsupseteq^* C'(k+1)$. Thus $\zeta[C'(k+1)] \sqsupseteq^* C'(k+1)$. Then, by rules (iii) and (iv), $C'(k+2) \sqsupseteq^* C'(k+1)$.

*Proof of (b)*: Given (a), (b) follows from the finiteness of the game (compare the proof of part (b) of Proposition 1).

*Proof of (c)*: We first show that $C(k^*) \sqsupseteq^* C'(k^{**})$. Since $C(k^*) \sqsupseteq^* C'(0)$ is trivially true, by rule (i), it is sufficient to prove that, for all $k \in \{0, 1, 2, \ldots.\}$, $C(k^*) \sqsupseteq^* C'(k)$ implies $C(k^*) \sqsupseteq^* C'(k+1)$. Suppose $C(k^*) \sqsupseteq^* C'(k)$. By the Lemma, $\zeta[C(k^*)] \sqsupseteq^* \zeta[C'(k)]$. But $\zeta[C(k^*)] = C(k^*)$ by the definition of $k^*$, and $\zeta[C'(k)] \sqsupseteq^* C'(k+1)$ by rule (ii). Thus $C(k^*) \sqsupseteq^* C'(k+1)$. We complete the proof by showing that $C'(k^{**}) \sqsupseteq^* C(k^*)$. Since $C'(k^{**}) \sqsupseteq^* C(0)$ follows trivially from the initiation rule of the categorisation procedure, it is sufficient to prove that, for all $k \in \{0, 1, 2, \ldots.\}$, $C'(k^{**}) \sqsupseteq^* C(k)$ implies $C'(k^{**}) \sqsupseteq^* C(k+1)$. Suppose $C'(k^{**}) \sqsupseteq^* C(k)$. By the Lemma, $\zeta[C'(k^{**})] \sqsupseteq^* \zeta[C(k)]$. But $\zeta[C'(k^{**})] = C'(k^{**})$ by the definition of $k^{**}$ and rule (iii), and $\zeta[C(k)] = C(k+1)$ by the continuation rule of the categorisation procedure. Thus $C'(k^{**}) \sqsupseteq^* C(k+1)$.

□

## References

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

Asheim, Geir B. and Andrés Perea (2009). Algorithms for cautious reasoning in games. Mimeo, Universities of Oslo and Maastricht.

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

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

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

Cubitt, Robin P. and Robert Sugden (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 (2008). Common reasoning in games. Centre for Decision Research and Experimental Economics (CeDEx) working paper 2008-01, University of Nottingham.

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

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

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

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

Quiggin, John (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3, 323-43.

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

Stahl, Dale (1995). Lexicographic rationality, common knowledge, and iterated admissibility. *Economics Letters* 47, 155-9.

# Notes

[1] More recently, Asheim and Dufwenberg (2003) show how an iterative procedure (which we discuss in Section 4) can be used to identify what they call "fully permissible sets". Such sets provide an equilibrium concept in the sense that a game may have more than one profile of fully permissible sets. Asheim and Dufwenberg's procedure identifies exactly all the fully permissible sets. Another way in which an iterative procedure might assist an equilibrium search is illustrated by the results of Brandenburger, Friedenberg and Kiesler (2008). They show that the strategies which survive (maximal) IDWDS must comprise a "self-admissible set". In this case, the iterative procedure always identifies exactly one self-admissible set, but the game may still have others.

[2] Cubitt and Sugden (2003) describe how Lewis's conception of common knowledge, which has the commonality of certain modes of reasoning as its central ingredient, differs from those which are now more familiar in the game theory literature. Cubitt and Sugden (2008) sets out formal foundations for a Lewisian approach but here we focus, not on foundations, but on solution concepts.

[3] 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 $S_i$.

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

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

[6] This notation uses unions of sets of strategies, some of which "belong" to one player and some to another. Recall that our labelling convention enables us, as analysts, to keep track of which strategies belong to whom.

[7] Monotonicity is not the *only* formal restriction on categorisation functions that might be justified by appeal to principles of reasoning; but it is sufficient for our purposes.

[8] For any categorisation $C_{-i}$, the set of probability distributions satisfying these conditions is well-defined and non-empty.

[9] For discussion of IDSDS in infinite games, see Dufwenberg and Stegeman (2002) and Chen *et al* (2007).

[10] This formulation of IDSDS, and later analogous formulations of IDWDS, rely on Theorems 1.6 and 1.7 of Myerson (1991). These theorems establish equivalences between propositions about dominance and propositions about optimality, conditional on allowable probability distributions. By using and "if and only if" here, we are defining maximal IDSDS, i.e. the form of IDSDS in which all strategies which are strictly dominated at a given stage are deleted at that stage. It is well-known that, in finite games, it makes no difference to the strategies which survive IDSDS whether this requirement is imposed or some non-maximal variant used. The maximal variant is, however, simpler to define.

[11] The procedure we have defined differs from IDSDS by distinguishing between two ways in which strategies may survive the deletion process: a strategy may be optimal for *all* beliefs that attach zero marginal probability to deleted strategies (and therefore be accumulated), or it may merely be optimal for *some* but not all such beliefs (and therefore be neither accumulated nor deleted).

[12] Samuelson uses the game in support of an argument that common knowledge of admissibility is not a consistent concept.

[13] Notice that, by accumulating *centre*, RBEU actually accumulates (and not merely fails to delete) a strategy that IDWDS deletes.

[14] This conception of caution is discussed by Asheim and Dufwenberg (2003). See Asheim and Perea (2009) for further exploration of iterative procedures that incorporate this concept.

[15] We use the term "permissible" here in the same sense as in our informal interpretation of categorisations in Section 2. Asheim and Dufwenberg (2003) have a different, and formal, concept of permissible sets.

[16] This would result in the "ICEU"procedure proposed by Cubitt and Sugden (2008).