A reason-based model of rational choice

Franz Dietrich
(LSE & Maastricht Univ.)

and

Christian List
(LSE)

Preference Change Workshop

May 2009
London School of Economics

Plan for the talk

1. **Classic rational choice theory vs. a reason-based model**
2. A reason-based model
3. A representation theorem for reason-based preferences
4. A closer look at three interpretations of motivating reasons
5. (perhaps) Proofs and various counterexamples showing why the theorem's assumptions and conclusions are as they are

The classic rational-choice paradigm

Preferences are fully determined by

- **fundamental desires**, captured by the utilities of fully described outcomes
- **beliefs**, themselves reduced to prior beliefs and information I

Alternative x is preferred to y (given current information I) iff x has a higher expected utility than y (conditional on current information I)

Fundamental desires and prior beliefs never change, only information changes

The classic rational-choice paradigm (cont.)

If, during the financial crisis, a banker becomes a social worker, was this just information?
–> intuitively, his fundamental reasons or motivations have changed

Reasons and motivations

- play no explicit role in the classic model
- are even implicitly assumed as fixed (never-changing), as one can argue

Critiques of rational choice theory often suffer from offering no formal alternative (except models of bounded rationality).

We proposed a reason-based model of rational choice

Our notion of a '(motivating) reason'

... is technical, thus open to different interpretations

Notably, the motivating reasons could be:

1. the (abstractly) **conceptualised** propositions
   –> players in a game usually don't conceptualise all moves 13 periods ahead
2. the **(qualitatively) understood** propositions
   –> 'worldwide inequality', 'I'll listen to Brahms music', 'we'll be friends', ... motivate after being qualitatively understood
3. the propositions taken into account in **preference formation**
   –> preference formation is a difficult and costly process
   –> many aspects has to be neglected ('heuristics')

Changes in (motivating) reasons...

- are a very real phenomenon

  –> various causes: experiences in life, physiological changes, alcohol, ...

  –> (recalling the 3 interpretations...) we re-conceptualise the world, gain or loose understanding, or re-form preference incorporating new aspects
- but arguably aren't reducible to information-learning in a classic rational choice

  –> obvious under some interpretations, non-obvious under others

Reason- vs. information-based rational choice

Classic information-based rational choice

The indiv. can be in different *states of information* I

I is the set of currently possible worlds

I gives rise to a preference order $\succeq_I$ (over relevant alternatives)

Change in information I implies change in preference $\succeq_I$

This reduces preference dynamics to belief dynamics

Reason-based rational choice

The indiv. can be in different *states of motivation* M

M is the set of currently motivating reasons

M generates a preference order $\succeq_M$

Change in motivating reasons M implies change in preference $\succeq_M$

This reduces preference dynamics to motivation dynamics

Disclaimer

Our reason-based model of preferences

- isn't meant to deny that information also plays a role
- should ultimately be merged with the classic model into a comprehensive model explaining preference by reasons and information

  –> future research!

Plan for the talk

1. Classic rational choice theory vs. a reason-based model
2. **A reason-based model**
3. A representation theorem for reason-based preferences
4. A closer look at three interpretations of motivating reasons
5. (perhaps) Proofs and various counterexamples showing why the theorem's assumptions and conclusions are as they are

Reason-based model: the alternatives

X: set of (mutually exclusive) **alternatives**, the objects of preference, e.g.:

- the indiv.'s choice alternatives (or decision paths)
- states of nature (not influenced by the indiv.)
- outcomes influence by many individuals (and perhaps nature)

Subsets of X: **propositions** or **events**

Reason-based model: reasons

Some propositions $R \subseteq X$ count as *reasons*.

- $\mathbf{R}$: fixed set of all reasons (formally, $\mathbf{R} \subseteq \mathcal{P}(X)$)
- e.g. $\mathbf{R} = \{$'war', 'am thrilled', 'am freezing', 'am rich', 'you're clinically depressed', 'No one likes us'$\}$
- *Which* propositions should count as reasons? (I.e., how should $\mathbf{R}$ be specified?)

  –> Delicate question, presumably context-dependent

  –> tentative def.: reasons are propositions that could motivate (i.e. motivate *irreducibly*, to avoid conjunctions 'R and P' of reasons R and irrel. prop.'s P to count as reasons)

  –> In some applications, reasons represent (relevant) *atomic* proposition.[1]

[1]The composition of reasons, like 'there is peace *and* it is freezing', needn't be included in $\mathbf{R}$ because motivation by them is reducible to motivation by their elemenary components.

Reason-based model: motivating reasons

A reason may or may not motivate.

A **motivating reason set**, or **motivating set**, is simply a subset M of the set of reasons $\mathbf{R}$.

Think of M as the individual's current **state**.

- a soldier in a war might have $M = \{$'war', 'am freezing'$\}$
- a teenager might have $M = \{$'No one likes me', 'I like her'$\}$

Reason-based model: motivating reasons (cont.)

We don't assume that *all* sets $M \subseteq \mathbf{R}$ define possible motivating sets (possible individual states).

Some sets $M \subseteq \mathbf{R}$ might be impossible; examples:

- some reason, such as 'am hungry', *always* has motivating power, say genetically
- whenever 'justice' motivates, 'freedom' motivates
- whenever 'justice' motivates, 'am-rich' doesn't motivate

Assumption: the possible motivating sets form a lattice, i.e., whenever M, M' are possible, so are $M \cap M'$ and $M \cup M'$.[2]

[2]This assumption can be weakened, namely in Th. 1 to 'if M, M' are possible, so is $M \cap M''$' and in Th. 2 to 'if M, M' are possible, so is some superset $M'' \supseteq M \cup M'$. Our proofs rest on these weaker assumptions.

Reason-based model: motivating reasons (cont.)

Examples of specifications of $\mathcal{M}$, the set of possible motivating sets $M \subseteq \mathbf{R}$.

1. $\mathcal{M} = \mathcal{P}(\mathbf{R})$, i.e. *all* sets $M \subseteq \mathbf{R}$ are possible (base-line case)
2. $\mathcal{M} = \{M \subseteq \mathbf{R} : M^* \subseteq M \subseteq \mathbf{R} \backslash M^{**}\}$, where M^* is some set of *always motivating* reasons and M^{**} some set of *never motivating* reasons. (M^* or M^{**} can be empty.)
3. ('vector model', Christian's talk)
   - $X = X_1 \times ... \times X_k$ (alternatives are k-dimensional vectors)
   - $\mathbf{R} = \{R_{j:t} : j \in \{1, ..., k\}$ and $t \in X_j\}$, where the reason $R_{j:t}$ is defined as $\{x : x_j = t\}$ ('the j^{th} coordinate is t')
   - $\mathcal{M}$ consist of all sets of type $\{R_{j:t} : j \in J$ and $t \in X_j\}$ where $J \subseteq \{1, ..., k\}$ is any set of 'salient' dimensions.

Reason-based model: preferences

For each possible motivating set M, let $\succeq_M$ be a preference order on X.[3]

–> $\succeq_M$ represents the individual's preferences in state M.

An example of preferences

Let each reason $R \in \mathbf{R}$ have a 'value' $v(R) \in \mathbb{R}$.

- 'others are happy' might have a high value, 'war' a low (negative) value, ...

Let preference $\succeq_M$ under motivating set M be representable by a 'utility' function $u_M : X \to \mathbb{R}$ defined by

$$u_M(x) = \sum_{R \in M : x \in R} v(R)$$

(the utility of x is sum-total value of all motivating reasons holding of x)

[3]Throughout, an *order* (on a set) is a transitive complete binary relation (on the set). Derived relations: $\succ_M$ (strict preference) and $\sim_M$ (indifference).

Plan for the talk

1. Classic rational choice theory vs. a reason-based model
2. A reason-based model
3. **A representation theorem for reason-based preferences**
4. A closer look at three interpretations of motivating reasons
5. (perhaps) Proofs and various counterexamples showing why the theorem's assumptions and conclusions are as they are

We need conditions on preferences...

because

- not all patterns of preferences across individual states are plausible
  –> luckily!
- we aim at parsimony
  –> the reason-based model should be no less elegant than the classic model

The parsimony goal

To compare:

- The Bayesian model of beliefs is simple in that the beliefs $\mathsf{Pr}_I(.)$ across info states I are generated by a single (prior) belief $\mathsf{Pr}(.)$: for each I, $\mathsf{Pr}_I(.) = \mathsf{Pr}(.|I)$
- The classic rational choice model of preferences is simple in that the preferences $\succeq_I$ across info states I are generated by just 2 objects, a prior belief $\mathsf{Pr}(.)$ and a utility fn. U: for all I, $\succeq_I$ is given by the expectation of U w.r.t. $\mathsf{Pr}(.|I)$.

Similarly (to anticipate our findings)

- our reason-based model will, via two postulates, become simple in that the preferences $\succeq_M$ across possible motivating sets M are generated by a single *generating relation/order* $\succeq$
- $\succeq$ will rank the (consistent) sets of reasons.
- E.g., let $\mathbf{R} = \{$'I'm healthy','you're hungry'$\}$, let these two reasons be mutually consistent, and let

$$\{\text{'I'm healthy'}\} \succ \{\text{'I'm healthy', 'you're hungry'}\} \sim \emptyset \succ \{\text{'you're hungry'}\}.$$

First postulate

Postulate 1: 'Only motivating reasons matter'. The indiv. is indifferent between alternatives satisfying the same motivating reasons. Formally: for all possible motivating reason sets M and alternatives $x, y \in X$, if $\{R \in M : x \in R\} = \{R \in M : y \in R\}$ then $x \sim_M y$.

- Satisfied nearly by definition under two of our interpretations of 'being motivating': 'being conceptualised' and 'being considered in a heuristic'.
- A non-obvious psychological hypothesis under the third interpretation: 'being qualitatively understood'.

Second postulate

While Postulate 1 addresses preference in a *fixed* individual state M, Postulate 2 addresses preference *change*.

Postulate 2. For every alternatives $x, y \in X$, possible motivating set M, and possible larger motivating set $M' \supseteq M$, if all additional motivating reasons $R \in M' \backslash M$ hold neither of x nor of y then $x \succeq_{M'} y \Leftrightarrow x \succeq_{M} y$.

Example: If 'red wine' becomes motivating, preference between two dishes without wine doesn't change (but preference between two dishes of which at least one includes wine may change).

Seems plausible, under all three interpretations of 'being motivating'.

Theorem 1: a single generating relation

A *relation between consistent reason sets* is an arbitrary binary relation $\succeq$ on the set $\{C \subseteq \mathbf{R} : \cap_{R \in C} \neq \emptyset\}$ of consistent sets of reasons.

Theorem 1. Postulates 1 and 2 hold if and only if there exists a relation $\succeq$ between consistent reason sets that generates each preference order $\succeq_M$ in the sense that

$x \succeq_M y \Leftrightarrow \{R \in M : x \in R\} \succeq \{R \in M : y \in R\}$ for all $x, y \in X$.

So, under the postulates, a single relation $\succeq$ – called *generating relation* – represents preferences $\succeq_M$ whatever the current motivating set M: x is (currently) preferred to y if and only if $\succeq$ ranks x's (currently) motivating reasons above y's.

Theorem 2: a single generating order

When is the generating relation $\succeq$ an *order* of consistent reason sets (i.e. a complete and transitive relation)?
While $\succeq$ can always be chosen as complete, there exist genuinely intransitive cases (see example slide).
But transitivity is guaranteed under a richness condition.

Richness. For any set of reasons $\mathbf{R}^* \subseteq \mathbf{R}$, if some alternative in X satisfies all reasons in $\mathbf{R}^*$ then some alternative in X satisfies all these reasons *and no others*.

Theorem 2. Assume Richness. Postulates 1 and 2 hold if and only if there exists an order of consistent reason sets $\succeq$ that generates each order $\succeq_M$ in the sense that

$$x \succeq_M y \Leftrightarrow \{R \in M : x \in R\} \succeq \{R \in M : y \in R\} \text{ for all } x, y \in X.$$

The generating order $\succeq$ is moreover **unique** if (i) Richness is strengthened to *full independence* of the reasons[4] and (ii) the *full* motivation set $M = \mathbf{R}$ is possible.

[4]Full independence: for every set $\mathbf{R}^* \subseteq \mathbf{R}$, some alternative $x \in X$ satisfies all reasons in $\mathbf{R}^*$ and no others.

Different types of generating relations

Some types of generating orders $\succeq$, in increasing order of generality:

1. *Additive order*: $C \succeq D \Leftrightarrow \sum_{R \in C} v(R) \geq \sum_{R \in D} v(R)$ ($v(R)$ the 'value' of reason R)

–> as in the Theorem in the 'vector model' (Christian's talk)

2. *Separable order*: $C \succeq D \Leftrightarrow C' \succeq D'$ whenever C', D' arise from C, D by adding the *same* reasons (e.g., {'peace', 'justice', ...} is better than {'fighting', ...} regardless of what '...' is provided it is the *same*)

–> the most natural example of separable orders are additive orders

3. *General order* (possibly non-separable; so, whether a reason is a 'for' or 'against' reason depends on which other reasons hold)

–> as in Th, 2

4. *General relation* (possibly with cycles)

–> as in Th. 1

Interpreting the generating relation/order

'$C \succeq D$' is interpretable as 'set C is (weakly) preferred to set D'

... but from what perspective? (Recall the agent's perspective changes!)

–> for simplicity, let *all* reason sets $M \subseteq \mathbf{R}$ be possible (otherwise the answer is more complicated to state)

A possible answer (see proof of Th. 1): From the perspective $M = C \cup D$ in which only the reasons in C and those in D motivate

–> so the perspective changes as the compared pair C, D changes

–> this explains why $\succeq$ can be intransitive!

An under Richness possible answer (see proof of Th. 2): From the full-motivation perspective $M = \mathbf{R}$

–> the perspective doesn't depend on what's being compared (C, D)

–> this explains why $\succeq$ is transitive

Plan for the talk

1. Classic rational choice theory vs. a reason-based model
2. A reason-based model
3. A representation theorem for reason-based preferences
4. **A closer look at three interpretations of motivating reasons**
5. (perhaps) Proofs and various counterexamples showing why the theorem's assumptions and conclusions are as they are

Being motivating as being conceptualised

M contains the currently **conceptualised** reasons.

The indiv. can distinguish alternatives $x, y \in X$ only on the basis of which conceptualised reasons are satisfied

So, the indiv.'s subjective representation of alternatives is coarser than the modeller's set of alternatives X.

- X_M : a coarsening of X, containing the 'subjective alternatives' given state M.[5]
- $\succeq^*_M$: the indiv.'s subj. preference order on X_M given state M
- $\succeq_M$: the modeller's representation of $\succeq^*_M$
  –> the modeller 'lifts' subj. preferences $\succeq^*_M$ to 'his' space X
  –> $x \succeq_M y$ stands for $x^* \succeq^* y^*$, where x^* (y^*) is the subjective alternative that is a coarsening of x (y).

N.B.: Postulate 1 holds by construction.

[5]Formally, one may view X_M as a partition of X into sets of subjectively non-distinghished 'modeller alternatives' $x \in X$. These sets are the equivalences classes w.r.t. the equivalence relation $\equiv_M$ on X defined by $x \equiv_M y \Leftrightarrow \{R \in M : x \in R\} = \{R \in M : y \in R\}$. Subjective alternative $x^* \in X_M$ is a coarsening of modeller's alternative $x \in X$ iff $x^* \ni x$.

Being motivating as being qualitatively understood

What exactly means (qualitatively) understanding R?
Tricky...

An informal definition: Someone **qualitatively understands** a proposition R if he can *imagine* its qualitative implications.

N.B.: lacking understanding of R doesn't mean not conceptualising R

- I can perfectly conceptualise 'Brahms music is on the programme' without any qualitative understanding
- so the 2 interpretations of motivating reasons as conceptualised resp. understood reasons are genuinely distinct

Qualitatively understanding R as conceptualising certain propositions *related* to R

We propose to think of (non-)understanding R as (non-)conceptualising certain *related* propositions.

Understanding R : 'Brahms music is on the programme'

$\Leftrightarrow$ Imagining the character or experience of Brahms music (call it B)

$\Leftrightarrow$ Conceptualising P_R : 'Brahms music is like B'.[6]

Proposition P_R is usually not expressible as a subset of X

[A proposition experessible in X if each $x \in X$ entails either it or its negation

–> 'am hungry' isn't expressible in $X = \{$'it's sunny', 'it rains'$\}$]

[6] Instead of one related proposition (P_R) there might be many ($P_R^1, P_R^2, ...$), each of which may or may not be conceptualised. This suggests generalising our model by allowing more than one way in which a reason can be understoood resp. be motivating. Feasible! Not for today!

Can't we reduce (non-)understanding of R to classic rational choice theory by...

- refining X so that the proposition P_R becomes expressible
- representing (non-)understanding of R as (non-)*information* of P_R?

This would (re-)model qualitative understanding of 'Brahms is on the programme' as being *informed* of 'Brahms is like B'.

A mistake:

- This confuses conceptualising P_R with knowing that P_R
- Conceptualising P_R needn't imply knowing (for sure) that P_R
  –> one can conceptualise P_R 'Brahms is like B' but be uncertain of whether P_R (is Brahms really as depressive as B?)
- Not conceptualising P_R is distinct from being uncertain of whether P_R is true
  –> If not understanding P were only a matter of *uncertainty* of whether P_R holds, then Postulate 1 would be implausible because already a mere *probability* of P_R (i.e. of Brahms being like B) can break the indifference.

Being motivating as being considered by a heuristic

No doubt, we often form our preferences based on a limited set of criteria

- like Gerd Gigerenzer, we consider heuristics based on a limited set of criteria M
- unlike Gigerenzer, we focus on how preference depends on and changes with the set of criteria M
- $\succeq_M$: preferences if the set of criteria is M
- Theorem 1: given Postulates 1-2, the preferences (heuristics) across all possible sets of criteria M are generated by a single relation $\succeq$ over consistent sets of criteria.

Plan for the talk

1. Classic rational choice theory vs. a reason-based model
2. A reason-based model
3. A representation theorem for reason-based preferences
4. A closer look at three interpretations of motivating reasons
5. **(perhaps) Proofs and various counterexamples showing why the theorem's assumptions and conclusions are as they are**

A preliminary to the proofs

The following simple lemma is used to prove both theorems and true independently of any assumptions on the set $\mathcal{M}$ of possible motivating sets M.

Lemma 1. Suppose Postulate 1. For all $x, y, x', y' \in X$ and all $M \in \mathcal{M}$, if $\{R \in M : x \in R\} = \{R \in M : x' \in R\}$ and $\{R \in M : x' \in R\} = \{R \in M : y' \in R\}$ then $x \succeq_M y \Leftrightarrow x' \succeq_M y'$.

Proof. Let $x, y, x', y' \in X$ and $M \in \mathcal{M}$ such that $\{R \in M : x \in R\} = \{R \in M : x' \in R\}$ and $\{R \in M : x' \in R\} = \{R \in M : y' \in R\}$. By Postulate 1, $x \sim_M x'$ and $y \sim_M y'$. So, as $\succeq_M$ is transitive, $x \succeq_M y \Leftrightarrow x' \succeq_M y'$. ■

Proof of Theorem 1

Proof. Throughout we write $\mathcal{M}$ for the set of all possible motivating sets, $M_x := \{R \in M : x \in R\}$ for the set of reasons in $M \subseteq \mathbf{R}$ holding of $x \in X$, and $\mathbf{C} := \{C \subseteq \mathbf{R} : \cap_{R \in C} \neq \emptyset\}$ for the set of consistent reason sets.

1. First, suppose a relation $\succeq$ on $\mathbf{C}$ generates all orders $\succeq_M$, $M \in \mathcal{M}$. Postulate 2 holds obviously. As for Postulate 1, consider any $M \in \mathcal{M}$ and any $x, y \in X$ such that $M_x = M_y$; we have to show that $x \sim_M y$. As $\succeq_M$ is reflexive, $x \sim_M x$. So, as $\succeq$ generates $\succeq_M$, $M_x \sim M_x$, which by $M_x = M_y$ implies $M_x \sim M_y$. Hence, again using that $\succeq$ generates $\succeq_M$, $x \sim_M y$, as desired.

Proof of Theorem 1 (cont.)

2. Now assume Postulates 1 and 2. Recall that $\mathcal{M}$ is by assumption closed under finite intersection $\cap$ (no need to assume closedness under the other lattice operation, finite union $\cup$).

Claim 1. For all $x, y, x', y' \in X$ and all $M, M' \in \mathcal{M}$, if $M_x = M'_{x'}$ and $M_y = M'_{y'}$ then $x \succeq_M y \Leftrightarrow x' \succeq_{M'} y'$.

Let $x, y, x', y' \in X$ and $M, M' \in \mathcal{M}$ with $M_x = M'_{x'}$ and $M_y = M'_{y'}$. As $\mathcal{M}$ is closed under $\cap$, $M \cap M' \in \mathcal{M}$. We first show that

$(M \cap M')_x = (M \cap M')_{x'} = M_x = M'_{x'}$ and $(M \cap M')_y = (M \cap M')_{y'} = M_y =$

The first set of equalities holds because, firstly, $M_x = M'_{x'}$ by assumption, secondly $(M \cap M')_x = M_x$ by $(M \cap M')_x = M_x \cap M'_x = M_x$ (in the last equality using that $M'_x \supseteq (M'_{x'})_x = (M_x)_x = M_x$) and, thirdly, $(M \cap M')_{x'} = M'_{x'}$ by $(M \cap M')_{x'} = M_{x'} \cap M'_{x'} = M'_{x'}$ (in the last equality using that $M_{x'} \supseteq (M_x)_{x'} = (M'_{x'})_{x'} = M'_{x'}$). The second block of equalities holds for parallel reasons.

Proof of Theorem 1 (cont.)

By $(M \cap M')_x = M_x$ and $(M \cap M')_y = M_y$, Postulate 2 yields

(*) $x \succeq_{M \cap M'} y \Leftrightarrow x \succeq_M y$.

By $(M \cap M')_{x'} = M'_{x'}$ and $(M \cap M')_{y'} = M'_{y'}$, Postulate 2 yields

(**) $x' \succeq_{M \cap M'} y' \Leftrightarrow x' \succeq_{M'} y'$;

By $(M \cap M')_x = (M \cap M')_{x'}$ and $(M \cap M')_y = (M \cap M')_{y'}$, Lemma 1 yields

(***) $x \succeq_{M \cap M'} y \Leftrightarrow x' \succeq_{M \cap M'} y'$.

The equivalences (*)-(***) together imply that $x \succeq_M y \Leftrightarrow x' \succeq_{M'} y'$. QED.

Proof of Theorem 1 (cont.)

Claim 1 allows us to define a binary relation $\succeq$ on $\mathbf{C}$ as follows: for all $C, D \in \mathbf{C}$, $C \succeq D$ holds if and only if $x \succeq_M y$ for *some* (hence by Claim 1 *all*) $x, y \in X$ and $M \in \mathcal{M}$ such that $M_x = C$ and $M_y = D$.

Claim 2 (which completes the proof). For each $M \in \mathcal{M}$, $\succeq$ generates $\succeq_M$, i.e. $x \succeq_M y \Leftrightarrow M_x \succeq M_y$ for all $x, y \in X$.

Let $M \in \mathcal{M}$ and $x, y \in X$. First, assume $x \succeq_M y$. We show that $M_x \succeq M_y$, i.e. that $x' \succeq_{M'} y'$ for some $x', y' \in X$ and $M' \in \mathcal{M}$ with $M'_{x'} = M_x$ and $M'_{y'} = M_y$. This obviously holds: simply take $x' = x$, $y' = y$ and $M' = M$. Conversely, assume that $M_x \succeq M_y$. Then, by $\succeq$'s definition and by Claim 1, $x' \succeq_{M'} y'$ for *all* $x', y' \in X$ and $M' \in \mathcal{M}$ satisfying $M'_{x'} = M_x$ and $M'_{y'} = M_y$. In particular, $x \succeq_M y$. ■

Proof of Theorem 2

Proof. Assume Richness. The proof is written so as to maximise parallels with the proof of Theorem 1. The notation '$\mathcal{M}$', 'M_x' and 'C' is as in that proof. Already by Theorem 1, Postulates 1 and 2 hold if some order $\succeq$ of consistent reason sets generates all preference orders $\succeq_M$, $M \in \mathcal{M}$.

Now assume Postulates 1 and 2. Recall that if $\mathcal{M}$ contains M, M' then $\mathcal{M}$ contains some superset of $M \cup M'$ (no need to assume that $\mathcal{M}$ is closed under $\cup$ and $\cap$).

Proof of Theorem 2 (cont.)

Claim 1. For all $x, y, x', y' \in X$ and all $M, M' \in \mathcal{M}$, if $M_x = M'_{x'}$ and $M_y = M'_{y'}$ then $x \succeq_M y \Leftrightarrow x' \succeq_{M'} y'$.

This claim, though analogous to the first claim in Theorem 1's proof, needs a different proof. Let $x, y, x', y' \in X$ and $M, M' \in \mathcal{M}$ such that $M_x = M'_{x'}$ and $M_y = M'_{y'}$. As $\mathcal{M}$ contains M, M', it by assumption contains some $\overline{M} \supseteq M \cup M'$. By Richness there are $\bar{x}, \bar{y} \in X$ such that $\mathbf{R}_{\bar{x}} = M_x$ and $\mathbf{R}_{\bar{y}} = M_y$, whence $\overline{M}_{\bar{x}} = M_x$ and $\overline{M}_{\bar{y}} = M_y$, so that by Postulate 2

(*) $\bar{x} \succeq_{\overline{M}} \bar{y} \Leftrightarrow \bar{x} \succeq_M \bar{y}$.

Proof of Theorem 2 (cont.)

By an analogous argument (preformed on x', y', M' instead of x, y, M), there are $\bar{x}', \bar{y}' \in X$ such that $\overline{M}_{\bar{x}'} = M'_{x'}$ and $\overline{M}_{\bar{y}'} = M'_{y'}$ and

(**) $\bar{x}' \succeq_{\overline{M}} \bar{y}' \Leftrightarrow \bar{x}' \succeq_{M'} \bar{y}'$.

Using Lemma 1, the right hand side of (*) is equivalent to $x \succeq_M y$ (because $M_{\bar{x}} = M_x$ and $M_{\bar{y}} = M_y$), the right hand side of (**) is equivalent to $x' \succeq_{M'} y'$ (because $M'_{\bar{x}'} = M'_{x'}$ and $M'_{\bar{y}'} = M'_{y'}$) and the left hand sides of (*) and (**) are equivalent to each other (because $\overline{M}_{\bar{x}} = \overline{M}_{\bar{x}'}$ and $\overline{M}_{\bar{y}} = \overline{M}_{\bar{y}'}$). These three equivalences together with the equivalences (*) and (**) imply that $x \succeq_M y \Leftrightarrow x' \succeq_{M'} y'$. QED.

Proof of Theorem 2 (cont.)

Claim 1 allows us to define a binary relation $\succeq^*$ on $\mathbf{C}$ (analogous to that defined when proving Theorem 1 but this time not our ultimate relation): for any $C, D \in \mathbf{C}$, $C \succeq^* D$ holds if and only if $x \succeq_M y$ for *some* (hence by Claim 1 *all*) $x, y \in X$ and $M \in \mathcal{M}$ such that $M_x = C$ and $M_y = D$.

Claim 2. For each $M \in \mathcal{M}$, the binary relation $\succeq^*$ generates $\succeq_M$, i.e. $x \succeq_M y \Leftrightarrow M_x \succeq M_y$ for all $x, y \in X$.

The proof is analogous to that of the second claim in Theorem 1's proof. QED.

Proof of Theorem 2 (cont.)

Claim 3. $\succeq^*$ is transitive.

Consider $C, D, E \in \mathbf{C}$ such that $C \succeq^* D$ and $D \succeq^* E$; we have to show that $C \succeq^* E$. By $C \succeq^* D$ there exist $x, y \in X$ and $M \in \mathcal{M}$ such that $M_x = C$, $M_y = D$ and $x \succeq_M y$. By $D \succeq^* E$ there exist $y', z \in X$ and $M' \in \mathcal{M}$ such that $M'_{y'} = D$, $M'_z = E$ and $y' \succeq_{M'} z$. By $M, M' \in \mathcal{M}$ and by assumption on $\mathcal{M}$, $\mathcal{M}$ contains some $\overline{M} \supseteq M \cup M'$. By Richness there are $\bar{x}, \bar{y}, \bar{z} \in X$ such that $\mathbf{R}_{\bar{x}} = C$, $\mathbf{R}_{\bar{y}} = D$ and $\mathbf{R}_{\bar{z}} = E$, whence $\overline{M}_{\bar{x}} = C$, $\overline{M}_{\bar{y}} = D$ and $\overline{M}_{\bar{z}} = E$. By $x \succeq_M y$, $M_x = \overline{M}_{\bar{x}}$ $(= C)$, $M_y = \overline{M}_{\bar{y}}$ $(= D)$ and Claim 1, we have $\bar{x} \succeq_{\overline{M}} \bar{y}$. Similarly, by $y' \succeq_{M'} z$, $M'_{y'} = \overline{M}_{\bar{y}}$ $(= D)$, $M'_z = \overline{M}_{\bar{z}}$ $(= E)$ and Claim 1, we have $\bar{y} \succeq_{\overline{M}} \bar{z}$. By $\bar{x} \succeq_{\overline{M}} \bar{y}$, $\bar{y} \succeq_{\overline{M}} \bar{z}$ and transitivity of $\succeq_{\overline{M}}$, we have $\bar{x} \succeq_{\overline{M}} \bar{z}$. So, by definition of $\succeq^*$ (and using that $\overline{M}_{\bar{x}} = C$ and $\overline{M}_{\bar{z}} = E$) we have $C \succeq^* E$. QED.

Proof of Theorem 2 (cont.)

Claim 4. There exists an order $\succeq$ on $\mathbf{C}$ that *extends* $\succeq^*$, in the usual sense that $C \succ^* D \Rightarrow C \succ D$ and $C \sim^* D \Rightarrow C \sim D$ for all $C, D \in \mathbf{C}$ (equivalently, that $C \succeq D \Leftrightarrow C \succeq^* D$ for all $C, D \in \mathbf{C}$ that are ranked relative to each other by $\succeq^*$).

This follows from Claim 3 via a classic extension theorem for binary relations (proven in its most general form by Suzumura, K. ,1976, Remarks on the theory of collective choice, *Economica 43*: 381–90). QED.

Claim 5 (which completes the proof). For each $M \in \mathcal{M}$, the order $\succeq$ defined in Claim 4 generates $\succeq_M$.

Let $M \in \mathcal{M}$ and $x, y \in X$. First, if $x \succeq_M y$, then $M_x \succeq^* M_y$ as $\succeq^*$ generates $\succeq_M$ (by Claim 2), whence $M_x \succeq M_y$ as $\succeq$ extends $\succeq^*$. Conversely, if $x \not\succeq_M y$, then $y \succ_M x$ as $\succeq_M$ is connected, so that $M_y \succ^* M_x$ as $\succeq^*$ generates $\succeq_M$ (by Claim 2); which implies that $M_y \succ M_x$ as $\succeq$ extends $\succeq^*$, hence that $M_x \not\succeq M_y$. ■

Proof that the generating order is unique under full independence of the reasons and if $\mathbf{R} \in \mathcal{M}$

Proof. Assume full independence and $\mathbf{R} \in \mathcal{M}$ (the latter strengthens Theorem 2's condition on $\mathcal{M}$); the postulates needn't be assumed. Any two generating relations $\succeq'$ and $\succeq''$ (on the set $\mathbf{C}$ of consistent reason sets) are identical because, for all $C, D \in \mathbf{C}$, letting $x, y \in X$ be two (by full independence existing) alternatives such that $\mathbf{R}_x = C$ and $\mathbf{R}_y = D$, $C \succeq' D$ is equivalent to $x \succeq_{\mathbf{R}} y$ (as $\succeq'$ generates $\succeq_{\mathbf{R}}$), which is equivalent to $C \succeq'' D$ (as $\succeq''$ generates $\succeq_{\mathbf{R}}$). ■

Example of a cyclical generating relation

- $X = \{110, 101, 011\}$ (each altern. describes which of three facts/reasons hold; e.g., 110 means that the first two hold).
- $\mathbf{R} = \{R_1, R_2, R_3\}$, where $R_i = \{t_1t_2t_3 : t_i = 1\}$ ('i^{th} fact holds')

  –> Richness is violated!
- Consider orders $\succeq_M$, $M \subseteq \mathbf{R}$, given as follows:

(i) $\succeq_{\{R_1,R_2\}}$ is the order given by $101 \succ_{\{R_1,R_2\}} 011 \succ_{\{R_1,R_2\}} 110$;

(ii) $\succeq_{\{R_2,R_3\}}$ is the order given by $110 \succ_{\{R_2,R_3\}} 101 \succ_{\{R_2,R_3\}} 011$;

(iii) $\succeq_{\{R_1,R_3\}}$ is the order given by $011 \succ_{\{R_1,R_3\}} 110 \succ_{\{R_1,R_3\}} 101$;

(iv) $\succeq_M$ is the full-indifference order whenever $|M| \in \{0, 1, 3\}$.

–> Postulate 1 holds (easy to check).

–> Postulate 2 holds because if $x, y \in X$, $M \subseteq \mathbf{R}$ and $R \in \mathbf{R}\backslash M$ s.t. $x, y \notin R$, then $x = y$, whence $x \sim_M y$ and $x \sim_{M\cup\{R\}} y$.

Example of a cyclical generating relation (cont.)

The preferences $\succeq_M$, $M \subseteq \mathbf{R}$, are generated by the (cyclic!) relation $\succeq$ that

(i) ranks $\{R_1\}$ over $\{R_2\}$, and both over $\{R_1, R_2\}$,

(ii) ranks $\{R_2\}$ over $\{R_3\}$, and both over $\{R_2, R_3\}$,

(iii) ranks $\{R_3\}$ over $\{R_1\}$, and both over $\{R_1, R_3\}$,

(iv) ranks as indifferent any pair of sets C, C' not yet compared in (i)-(iii).

Cyclicity is unavoidable: *every* specification of the generating relation must obey (i)-(iii).

Th. 1 wouldn't hold w/o assuming that $\mathcal{M}$ is closed under $\cap$: counterexample

- $X = \{110, 101, 000\}$ (alternatives describe which of three facts/reasons hold; e.g., 110 means that the first two hold)
- $\mathbf{R} = \{R_1, R_2, R_3\}$, where $R_1 = \{110, 101\}$ ('first fact holds'), $R_3 = \{110\}$ ('second fact holds') and $R_3 = \{101\}$ ('third fact holds')
- $\mathcal{M} = \{\{R_1, R_2\}, \{R_1, R_3\}, \{R_1, R_2, R_3\}\}$ (not closed under $\cap$, just under $\cup$)
- $\succeq_{\{R_1,R_2,R_3\}}$ is the full-indifference order
- $\succeq_{\{R_1,R_2\}}$ is the order given by $101 \succ_{\{R_1,R_2\}} 000 \sim_{\{R_1,R_2\}} 110$
- $\succeq_{\{R_1,R_3\}}$ is the order given by $101 \sim_{\{R_1,R_3\}} 000 \succ_{\{R_1,R_3\}} 110$

–> Although Postulates 1 and 2 hold (!), no generating relation $\succeq$ exists as $\succeq$ would have to satisfy $\{R_1\} \succ \emptyset$ (by $101 \succ_{\{R_1,R_2\}} 000$) but also $\emptyset \succ \{R_1\}$ (by $000 \succ_{\{R_1,R_3\}} 110$).

Th. 2 wouldn't hold w/o assuming that if $\mathcal{M}$ contains M, M' then it contains some $M'' \supseteq M \cup M'$: counterexample

- $X = \{0,1\}^3 = \{(0,0,0),(1,0,0),...\}$
- $\mathbf{R} = \{R_1, R_2, R_3\}$ with $R_i = \{(t_1,t_2,t_3) \in X : t_i = 1\}$
- $\mathcal{M} = \{M \subseteq \mathbf{R} : |M| \leq 2\}$
  –> violates our condition (but closed under $\cap$)
- $\succeq_\emptyset$ is the full-indifference order
- $\succeq_{\{R_i\}}$ is generated by an order $\succeq$ s.t. $\{R_i\} \succ \emptyset$
- $\succeq_{\{R_1,R_2\}}$ is generated by an order $\succeq$ s.t. $\{R_1\} \succ \{R_2\} \succ \emptyset \sim \{R_1, R_2\}$
- $\succeq_{\{R_2,R_3\}}$ is generated by an order $\succeq$ s.t. $\{R_2\} \succ \{R_3\} \succ \emptyset \sim \{R_2, R_3\}$
- $\succeq_{\{R_1,R_3\}}$ is generated by an order $\succeq$ s.t. $\{R_3\} \succ \{R_1\} \succ \emptyset \sim \{R_1, R_3\}$

–> Postulates 1-2 hold (Postu. 1 holds as each $\succeq_M$ is 'generated')

–> But no order $\succeq$ generates all $\succeq_M$, $M \in \mathcal{M}$, as $\succeq$ would have to satisfy $\{R_1\} \succ \{R_2\}$, $\{R_2\} \succ \{R_3\}$, $\{R_3\} \succ \{R_1\}$, cycle!