

**These are elaborate slides which are based on but go  
beyond the paper**

**"A reason-based theory of rational choice" by F. Dietrich  
and C. List**

# Plan

- 1. Reasons can rationalise 'paradoxical' choice**
2. The need to incorporate reasons into rational choice theory
3. A reason-based model
4. A representation theorem for reason-based preferences
5. Two sources of 'paradoxical' choice
6. Atomism or holism?
7. Some future avenues

## Sen's paradoxical choice example

- In an invitation, the agent has to choose from a buffet with several pieces of cake.
- Consider 'small', 'large', and 'huge' pieces.
- The agent is hungry but (wants to look) polite.
- He therefore always chooses the *second largest* available piece:

$$C(\{\text{'small'}, \text{'large'}, \text{'huge'}\}) = \{\text{'large'}\}$$

$$C(\{\text{'small'}, \text{'large'}\}) = \{\text{'small'}\}$$

etc.

where  $C$  is the agent's *choice function*

# Such behaviour is

## **not classically rationalisable:**

- There is no binary (preference) relation  $\succsim$  over all pieces of cake s.t. the agent always picks the  $\succsim$ -maximal available piece.
- (Technically,  $C$  violates *contraction-consistency*.)

## **... but reason-based rationalisable:**

- Explicable by perfectly stable motivating reasons
- ... and a stable way to weigh reasons.
- (Details later.)

# Plan

1. Reasons can rationalise 'paradoxical' choice
- 2. The need to incorporate reasons into rational choice theory**
3. A reason-based model
4. A representation theorem for reason-based preferences
5. Two sources of 'paradoxical' choice
6. Atomism or holism?
7. Some future avenues

# Two contrasting paradigms

a paradigm in **philosophy** (but  
also on the streets!)

“an agent acts on the basis  
of reasons.”

**reasons**



**action**

leading paradigm in the **social sciences**  
(the ‘rational choice’ paradigm)

“an agent acts on the basis  
of desires and beliefs.”

**desires**

**beliefs**



**action**

The contrast emerges in two areas

**Positive (explanation)**

What an agent *actually* does depends on his/her...

*motivating* reasons. | *actual* desires and beliefs.

**Normative (justification)**

What an agent *ought* to do depends on his/her...

*normative* reasons. | *idealised* desires and beliefs.

# Example

## **Positive (explanation)**

The agent killed Sam because...

Ann loves Sam,  
Sam is rich, ...

the agent has such-and-such  
utility function & probability function

## **Normative (justification)**

The agent ought not to kill Sam because...

Sam is a human being,  
Sam is no threat to anyone, ...

the agent ought to be guided by such-and-  
-such utility function & probability function

# We'll cover positive and normative dimensions

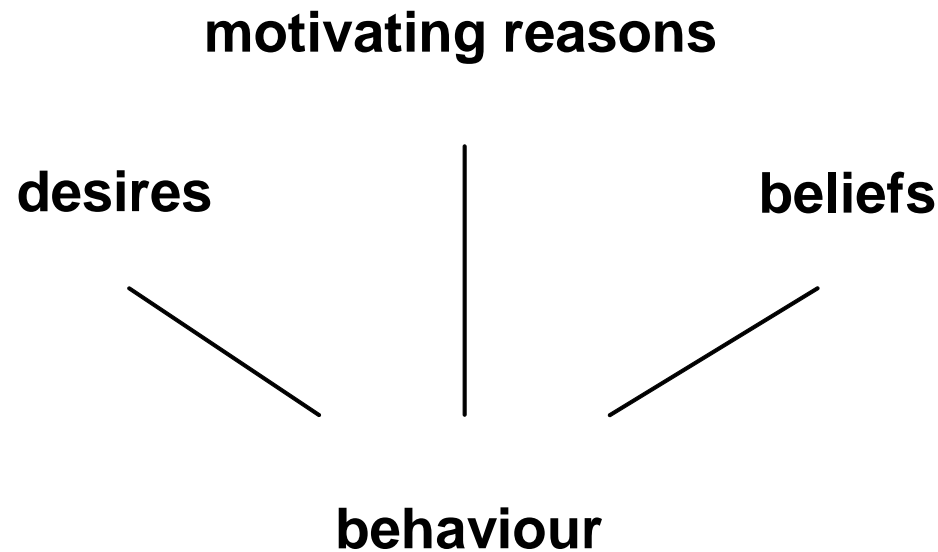
- Our formal model will allow for either reading:
  - positive or normative
  - where 'normative' could mean 'moral' or '(substantively) rational'.
- Depending on the reading, we thus get a reason-based model of *actual* choice, or *moral* choice, or (substantively) *rational* choice.
- Don't mix interpretations! One at a time!

# Convention

- I now talk primarily in positive terms: actual behaviour, motivating reasons.
- (In the "applications" I come back to the normative reading.)

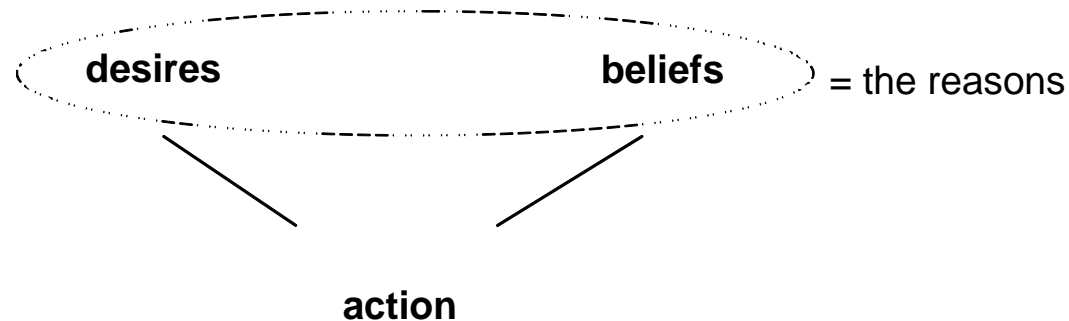
First attempt (to reconcile reasons with rational choice theory): naive

Add motivating reasons as a third determinant of actions:



## Second attempt: deflationary

- Identify the motivating reasons with the desires and beliefs!  
(Davidson, Michael Smith)

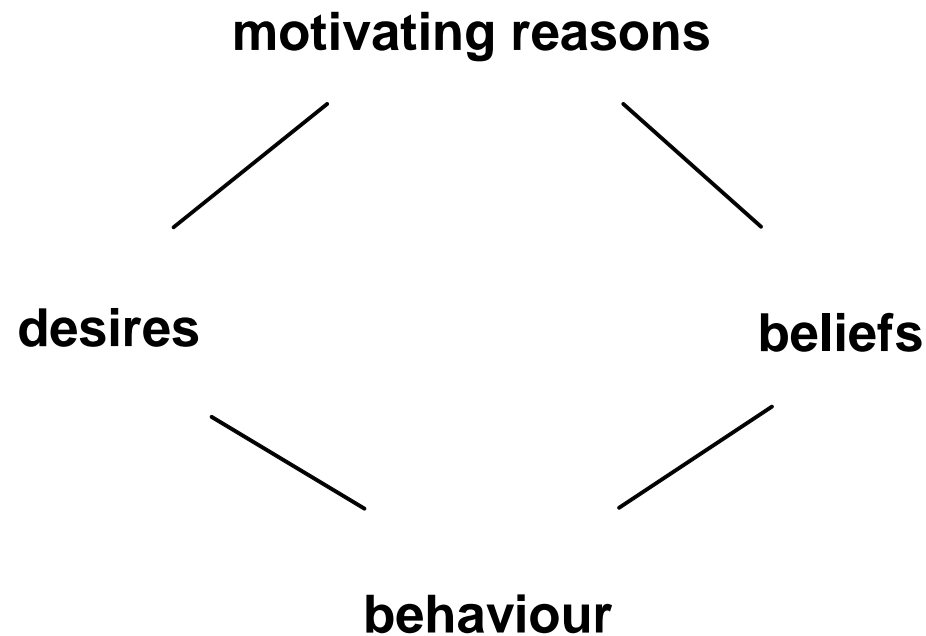


### Problems:

- Not faithful to the daily-life notion of reasons as propositions
- We can't ask 'For what reasons does the agent hold his desires/beliefs?'

## Third attempt: the right way

- Reasons are more fundamental than desires and beliefs.
- Reasons *explain* the desires and beliefs.



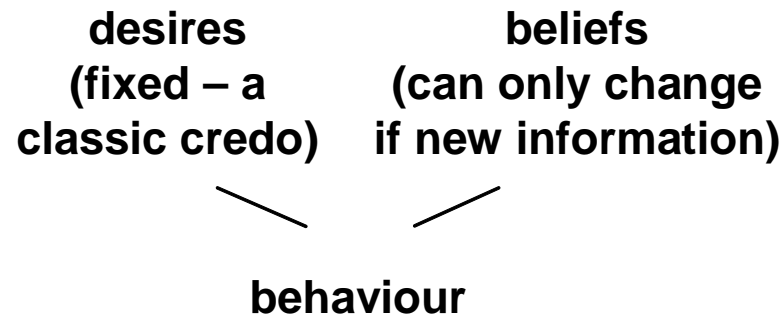
# Reasons for reasons

- '*More explanation*': Any approach that doesn't address *why* agents hold their desires and (prior) beliefs is question-begging.
- '*More realism*': Overcome the unrealistic nature of standard rational choice modelling.
- '*Understanding change*': overcome the narrow classic account of human change (next slide).

# Reasons for reasons

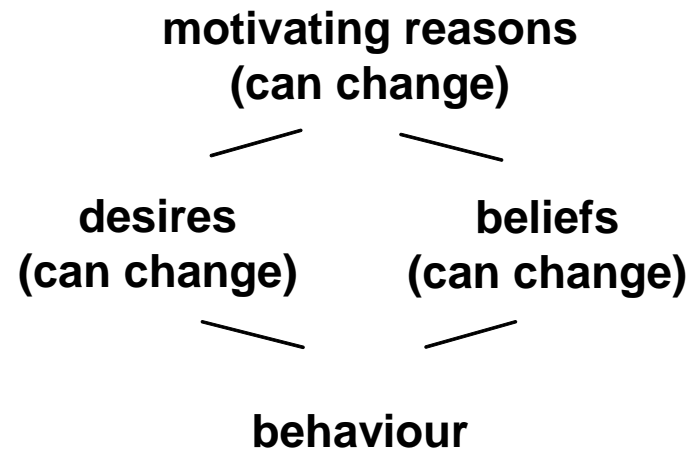
## ***Classical approach:***

**“all behavioural change comes from new information”**



## ***New approach:***

**“behavioural change can come from new motivating reasons”**

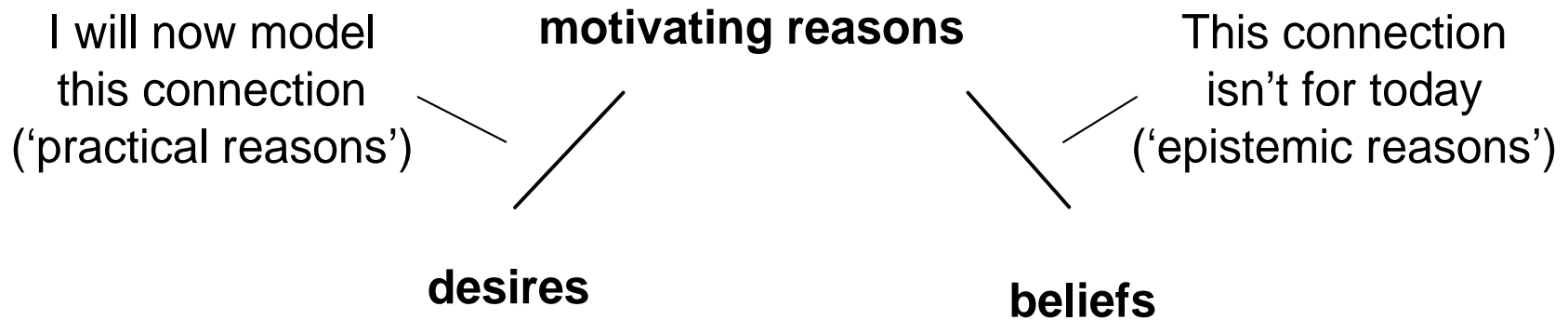


# Plan

1. Reasons can rationalise 'paradoxical' choice
2. The need to incorporate reasons into rational choice theory
- 3. A reason-based model**
4. A representation theorem for reason-based preferences
5. Two sources of 'paradoxical' choice
6. Atomism or holism?
7. Some future avenues

# Model

Critiques of rational choice theory often suffer from offering no formal alternative.



## Model: the alternatives

- Set  $X$  of (mutually exclusive) **alternatives**  $x, y, z, \dots$
- Alternatives could be:
  - fully specified possible worlds
  - ‘outcomes’
  - ...
- A **proposition** (‘it rains’, ‘there is global injustice’, ...) is true of certain alternatives, and false of the others.
- (Instead of ‘propositions’ one may think of ‘properties’.)

# Model: reasons

- ‘Definition’: We think of a reason as a proposition with a particular kind of relevance for the agent’s preferences towards the alternatives of which it is true.
- Depending on how we spell out ‘relevance’, we obtain different kinds of reason:
  - a *motivating reason* is a proposition that is motivationally relevant: if true of an alternative, it may affect the agent’s *actual* preference towards this alternative.
  - a *normative reason* is a proposition that is normatively relevant: if true of an alternative, it may affect the preference the agent *ought* to have towards this alternative.
- As mentioned, I’ll focus on *motivating reasons*

## Model: motivating reasons

- The individual has some set of motivating reasons  $M = \{R, R', \dots\}$ .
- Think of  $M$  as the individual's current **psychological state**.
- A soldier in a war might have  $M = \{\text{'It's war'}, \text{'I'm freezing'}\}$ .
- A teenager might have  $M = \{\text{'No one likes me'}, \text{'I like her'}\}$ .

# Model: psychological states

- The set of motivating reasons  $M$  can change.
- But only certain sets  $M$  might be *psychologically possible*.
  - Perhaps the agent cannot have an empty set  $M = \emptyset$  of motivating reasons.
- The model can handle various assumptions on psychologically (im)possibility.
  - Formally, there is a set  $\mathcal{M}$  of psychologically possible sets of motivating reasons  $M$
  - $\mathcal{M}$  must satisfy weak regularity conditions (ask me if curious!).

## Model: possible reasons

- A proposition that belongs to at least one psychologically possible set of motivating reasons  $M$  will be called a *possible reason*.

A possible reason might become a motivating reason once it is...

1. **abstractly conceptualised**

- The agent has awareness (i.e., a mental representation) of the proposition.

2. **qualitatively understood**

- Becoming motivated by 'there is injustice' requires *qualitative* understanding, not just abstract conceptualisation.

3. **attentionally salient** (borrowed from psychology)

- In real life, we pay attention only to few reasons/aspects when forming desires.

# Model: preferences

- For each psychologically possible motivating reason set  $M$ , let  $\succsim_M$  be a preference order on  $X$ .<sup>1</sup>
- $\succsim_M$  represents the individual's preferences in psychological state  $M$ .

<sup>1</sup>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

1. Reasons can rationalise 'paradoxical' choice
2. The need to incorporate reasons into rational choice theory
3. A reason-based model
4. **A representation theorem for reason-based preferences**
5. Two sources of 'paradoxical' choice
6. Atomism or holism?
7. Some future avenues

# An axiom of internal consistency in a fixed psychological state

## **Axiom 1 ('Principle of insufficient motivating reason').**

- Informally: If (for instance) the agent is motivated only by whether fish and whether wine is served for dinner, then he/she is indifferent between any two dinner plans that include the *same* main course and drink.
- Formally: for any psychologically possible motivating reason set  $M$  and any alternatives  $x, y$  in  $X$ , if  $\{R \in M : R \text{ is true of } x\} = \{R \in M : R \text{ is true of } y\}$  then  $x \sim_M y$

Plausible! Only what's motivating matters!

# An axiom about change in psychological state

## **Axiom 2 ('Consistent updating').**

- Informally: if (for instance) 'wine is served for dinner' becomes additionally motivating, then the preference between two dinner plans *that don't include wine* remains the same.
- Formally: for any alternatives  $x, y$  in  $X$ , any psychologically possible motivating reason set  $M$ , and any increased one  $M^+ \supseteq M$ , if all additional motivating reasons  $R \in M^+ \setminus M$  are false of  $x$  and of  $y$ , then  $x \succsim_{M^+} y \Leftrightarrow x \succsim_M y$ .

Plausible! Only what's motivating matters!

# Weighing reasons

- A *reason combination* is simply a set of possible reasons.
- A *weighing relation* is a binary relation  $\geq$  over reason combinations.
- ' $C \geq D$ ' means ' $C$  weighs at least as much as  $D$ '
- The weighing relation might say this (among others):  
...  $>$  {'I'm healthy'}  $>$  {'I'm healthy', 'you're tired'}  $>$  {'you're tired'}  
(where ' $>$ ' denotes the strict relation induced by  $\geq$ ).

# A representation theorem

**Theorem 1.** The agent's preference orders  $\succsim_M$  across psychologically possible motivating reason sets  $M$  in  $\mathcal{M}$  satisfy Axioms 1 and 2 if and only if these preferences are given by

$$x \succsim_M y \Leftrightarrow \{R \in M : R \text{ is true of } x\} \geq \{R \in M : R \text{ is true of } y\}$$

for some fixed weighing relation  $\geq$ .

- *In short:* The axioms hold if and only if preferences go by weighing reasons: an alternative  $x$  is preferred to another  $y$  just in case the motivating reasons true of  $x$  weigh more than those true of  $y$ .

# A simple example

- $X$  contains holiday destinations
- Only three possible reasons: it's exotic ( $E$ ), it's safe ( $S$ ), it's french-speaking ( $F$ ). (B.t.w., the agent is a Frenchman.)
- The various reason combinations are weighed like this:

$$\{E, S, F\} > \{E, S\} > \{S, F\} > \{E, F\} > \{F\} > \{E\} > \{S\} > \emptyset$$

- Then the preference between travelling to Provence (safe, french-speaking) or to Algeria (exotic, french-speaking) are:

$$M = \{E, S, F\} \Rightarrow \text{'Provence'} \succ_M \text{'Algeria'}$$

$$M = \{E, S\} \Rightarrow \text{'Algeria'} \succ_M \text{'Provence'}$$

etc.

- Note: preference changes if motivating reasons change.  
→ impossible in orthodox decision theory

# The weighing relation

A philosophically interesting object:

- It captures the (hard) exercise of weighing between reason combinations.
- Its meaning depends on whether we model actual, moral, or (substantively) rational choice.

# The weighing relation

A mathematically interesting object:

- possibly intransitive (weighing cycles)
  - the paper contains a theorem without intransitivity
- replaced by a weighing *function* in a cardinal (non-ordinal) variant of the model
  - here, the primitives are utility functions  $u_M$ ,  $M \in \mathcal{M}$ , and another theorem characterises them by a single weighing *function*.

# Reason-based decision under risk

**reasons**



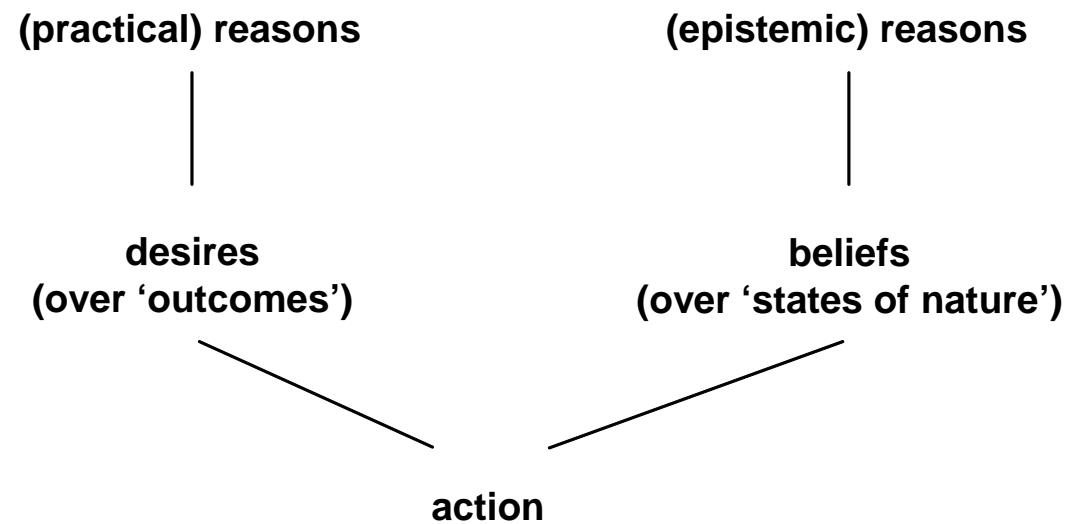
**desires**

**over *risky* prospects (lotteries over  $X$ )**



**action**

# Reason-based decision under uncertainty



# Classical rational choice as a special case

- the case of *fixed* motivating reasons for both desires and beliefs.

# Plan

1. Reasons can rationalise 'paradoxical' choice
2. The need to incorporate reasons into rational choice theory
3. A reason-based model
4. A representation theorem for reason-based preferences
- 5. Two sources of 'paradoxical' choice**
6. Atomism or holism?
7. Some future avenues

## A technical preliminary

- So far  $X$  was fixed.
- Now  $X$  varies:  $X$  is any non-empty subset of some fixed universal set  $Z$ .
- A reasons may refer to ('speak about') the context!
  - 'I act politely' is true of choosing 'big (piece of cake)' from {'big', 'huge'}, but false of choosing 'big' from {'small', 'big'}.

# Reason-based preferences

- Faced with menu  $X$  in psychological state  $M$ , the agent holds preferences given by

$$x \succsim_M^X y \Leftrightarrow M(x|X) \geq M(y|X) \text{ for all } x, y \in X,$$

where

$$M(x|X) := \{R \in M : R \text{ is true of choosing } x \underbrace{\text{from } X}_{\text{the new bit!}}\}.$$

- E.g.  $M(\text{'small'}|\{\text{'small'}, \text{'huge'}\}) = \{\text{'I act politely', I eat a small amount'}\}$ .
- *In short*: preferences go by weighing the reasons on both sides.
- N.B.: preference may depend on  $X$ :  $\succsim_M^X$  instead of just  $\succsim_M$ .
- See Theorem 1' for this characterisation of menu-dependent reason-based preferences (complementary slides)

# Two potential sources of 'paradoxical' choice

- Some reasons *refer to the context* ( $M(x|X)$  depends on  $X$ )
- Reasons are *context-dependent* ( $M \equiv M_X$ ).

Goal: Classify standard 'choice paradoxes' into these two kinds.

# Sen's example: a case of fixed context-referring reasons

Alternatives:  $s, l, h$  ('small/large/huge piece of cake')

*Set of motivating reasons* :  $M = \{P, S, L, H\}$ , where

- $P$  : I act politely (true of choosing  $s$  from  $\{s, l, h\}$ ,  $l$  from  $\{s, l, h\}$ ,  $s$  from  $\{s, l\}$ , ...)
- $S$  : I eat a small amount (true of  $s$ , context-independently)
- $L$  : I eat a large amount (true of  $l$ , context-independently)
- $H$  : I eat a huge amount (true of  $h$ , context-independently)

*Weighing relation*  $\geq$  :

$$\{P, H\} > \{P, L\} > \{P, S\} > \{H\} > \{L\} > \{S\}$$

## Sen's example rationalised

The above  $M$  and  $\succeq$  rationalise the choice behaviour:

- $C(\{s, l, h\}) = \{l\}$  since  $\begin{cases} M(l|\{s, l, h\}) = \{P, L\} \\ > M(s|\{s, l, h\}) = \{P, S\} \\ > M(h|\{s, l, h\}) = \{H\} \end{cases}$
- $C(\{s, l\}) = \{s\}$  since  $\begin{cases} M(s|\{s, l\}) = \{P, S\} \\ > M(l|\{s, l\}) = \{L\} \end{cases}$
- ...

# A case of variable but not context-referring reasons

- Suppose someone develops motivation by  $R$  only when faced with a menu in which  $R$  is true of *some but not all* alternatives.  
→ car size becomes relevant only once cars differ in size
- ‘Dynamic inconsistency’

# Plan

1. Reasons can rationalise 'paradoxical' choice
2. The need to incorporate reasons into rational choice theory
3. A reason-based model
4. A representation theorem for reason-based preferences
5. Two sources of 'paradoxical' choice
- 6. Atomism or holism?**
7. Some future avenues

# The model at work

- The model can serve as a platform for addressing philosophical problems.
- I now give an example

# Holism or atomism?

- Holists about reasons claim that what counts as reasons in favour of something varies with the context.
  - that I am being honest is a reason in favour of *some* honest actions, but against some others
- Atomists insists on the invariability of (well-specified) reasons.

The two positions can be spelled out precisely in our model: holism rejects certain formal ('separability') properties of the weighing relation.

# Moral particularism or generalism?

- Within ethics, holism is called *particularism*: there are no general moral principles; moral judgment is (highly) case-sensitive.
- ... and atomism is called *generalism*: morality is governed by general principles.

# What exactly is atomism?

- Different versions of atomism seem to coexist in the atomism/holism debate.
- They come to light once construed as different forms of separability of the weighing relation:
  - weak atomism/separability
  - strong atomism/separability
  - additive atomism/separability (even stronger)

# Weak atomism/separability

- Informally: whether a reason counts positively does not depend on which other reasons are present.
- Formally: for every reason  $R$ , we have  $\{R\} \cup C \geq C$  for either all or no combinations of other reasons  $C$ .

*Recall our holiday choice example:*

- reasons: it's exotic ( $E$ ), it's safe ( $S$ ), it's french-speaking ( $F$ ).
- $\{E, S, F\} > \{E, S\} > \{S, F\} > \{E, F\} > \{F\} > \{E\} > \{S\} > \emptyset$
- Weakly atomic! Any reason is better present than absent *regardless* of what else is present.

# Why weak atomism isn't full atomism

- Recall our holiday example:

$$\{E, S, F\} > \{E, S\} > \{S, F\} > \{E, F\} > \{F\} > \{E\} > \{S\} > \emptyset$$

- Is it better that it is exotic or that it is safe?
- The answer this time depends on what other reasons are present:
  - nothing else present:  $\{E\} > \{S\}$
  - $F$  present:  $\{S, F\} > \{E, F\}$ .
- Such context-dependence violates the spirit of atomism.
- The example is weakly but not strongly atomic.

# Strong atomism/separability

- Informally: whether one reason combination weighs more than another does not depend on which other reasons are present.
- Formally: for every pair of reason combinations  $C_1, C_2$ , we have  $C_1 \cup C \geq C_2 \cup C$  for either all or no sets of other reasons  $C$ .
- Strong atomism implies weak atomism: take  $C_1 = \{R\}$  and  $C_2 = \emptyset$ .
- Our holiday example isn't strongly separable: take  $C_1 = \{E\}$ ,  $C_2 = \{S\}$ .
- Strong atomism holds *if and only if* a stronger variant of Axiom 2 holds.

# Why even strong atomism isn't the end

- Some atomists go beyond strong separability.
- They want to 'weighing a whole by adding the weights of its parts'
- Let's call this 'additive atomism'.

# Additive atomism/separability

- Informally: the total weight/force of a reason combination is the sum of the weights/forces of its atomic components.
- Formally: there is an assignment of a weight  $w(R)$  to every possible reason  $R$  such that for every pair of reason combinations  $C, C'$  we have  $C \geq C'$  if and only if  $\sum_{R \in C} w(R) \geq \sum_{R' \in C'} w(R')$ .
- E.g.,  $w(\text{'you're happy'}) = 3$ ,  $w(\text{'I'm hungry'}) = -1$ , ...
- Additive separability implies strong separability (because in an inequality we can cancel identical terms on both sides).

# Should we be holists or atomists?

- Possibly depends on the variant of atomism.
- Worth exploring the intermediate atomistic positions (non-additive strong atomism, or weak but not strong atomism)
  - weak atomism definitely seems more defensible than full-blown additive atomism
  - classical utilitarianism: adding the pleasures and subtracting the pains.
- I won't give a definite answer.
- But I have a few views, if there is time...

# Plan

1. Reasons can rationalise 'paradoxical' choice
2. The need to incorporate reasons into rational choice theory
3. A reason-based model
4. A representation theorem for reason-based preferences
5. Two sources of 'paradoxical' choice
6. Atomism or holism?
7. **Some future avenues**

# Some future avenues (1)

- Reason-based rationalisability/explicability in an explanation frame
- Consequentialist vs. deontological reasons
- Atomism/holism debate, moral generalism/particularism debate
  - is the weighing relation separable? (And to what degree?)
- Subjective vs. objective ontology

## Some future avenues (2)

- Is the normative (moral or rational) weighing relation universal?
- Is the actual weighing relation universal?
  - Do people disagree because of different motivating reasons based on a *shared* (universal) weighing relation?
  - → compare with Gary Becker's Thesis
  - Does group deliberation lead to convergence of motivating reasons?

## Some future avenues (3)

- Game theory with changes in players' motivations (e.g., emerging sympathy for players who cooperate).
- Amartya Sen's distinction between sympathy and commitment as a distinction on the level of motivation
- 'Paternalism' and 'meddlesomeness' as morally wrong motivations (→ may lead into the Liberal Paradox)
- Motivation through *qualitative understanding* (→ qualia)  
Frank Jackson's thought experiment, Mary

Complementary slides

# Example of a cyclical weighing relation

- $X = \{110, 101, 011\}$  (each altern. describes which of three facts/reasons hold; e.g., 110 means that the first two hold).
  - $\mathcal{P} = \{R_1, R_2, R_3\}$ , where  $R_i$  is the proposition 'fact  $i$  holds' (i.e.,  $R_1$  is true of 110 and 101; and similarly for  $R_2$  and  $R_3$ )
  - Consider orders  $\succeq_M$ ,  $M \subseteq \mathcal{P}$ , given as follows:
    - (i) if  $M = \{R_1, R_2\}$  then  $101 \succ_M 011 \succ_M 110$ ;
    - (ii) if  $M = \{R_2, R_3\}$  then  $110 \succ_M 101 \succ_M 011$ ;
    - (iii) if  $M = \{R_1, R_3\}$  then  $011 \succ_M 110 \succ_M 101$ ;
    - (iv) if  $\#M \in \{0, 1, 3\}$  then  $\succeq_M$  is the full-indifference order.
- Axiom 1 holds (easy to check).
- Axiom 2 holds because for all  $x, y \in X$ ,  $M \subseteq \mathcal{P}$ , and  $R \in \mathcal{P} \setminus M$  false of  $x, y$ , we have  $x = y$ , whence  $x \sim_M y$  and  $x \sim_{M \cup \{R\}} y$ .

## Example of a cyclical weighing relation (cont.)

The preferences  $\succeq_M$ ,  $M \subseteq \mathcal{P}$ , are generated by the (cyclic!) weighing relation  $\succeq$  which

- (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 weighing relation must obey (i)-(iii).

How can the weighing relation be cyclical given that the preferences are not?

- In the cyclical example certain reason combinations  $C \subseteq \mathcal{P}$  are not *instantiated* in  $X$ , that is, there is no alternative  $x$  in  $X$  of which exactly the reasons in  $C$  are true.
- If each reason combination could be instantiated, then cycles on the level of reason combinations would lead to cycles on the level of preferences, and hence are impossible.
- So...

## Theorem 2: a single weighing order

When is the weighing relation  $\geq$  an *order* (i.e. a complete and transitive relation)?

**Weakly Independent Reasons.** For any set of possible reasons  $\mathcal{P}^* \subseteq \mathcal{P}$ , if some alternative in  $X$  satisfies all reasons in  $\mathcal{P}^*$  then some alternative in  $X$  satisfies all these reasons *and no others*.

**Theorem 2.** Assume Weakly Independent Reasons. The agent's preference orders  $\succsim_M$  across all the psychologically possible motivating reason sets  $M$  satisfy Axioms 1 and 2 if and only if these preferences are given by

$$x \succsim_M y \Leftrightarrow \{R \in M : R \text{ is true of } x\} \geq \{R \in M : R \text{ is true of } y\}$$

for some fixed weighing *order* (!)  $\geq$ .

# Varying the feasible set: the axioms

**Axiom 1' (consistency across feasible sets).** For any psychological state  $M$ , feasible sets  $X, X'$ , and alternatives  $x, y \in X$ , and  $x', y' \in X'$ , if  $\{R \in M : R \text{ is true of choosing } x \text{ from } X\} = \{R \in M : R \text{ is true of choosing } x' \text{ from } X'\}$  and  $\{R \in M : R \text{ is true of choosing } y \text{ from } X\} = \{R \in M : R \text{ is true of choosing } y' \text{ from } X'\}$ , then  $x \succsim_M^X y \Leftrightarrow x' \succsim_M^{X'} y'$ .

**Axiom 2' (consistency across motivational states).** For any feasible set  $X$ , alternatives  $x, y \in X$  and psychological states  $M, M^+ \in \mathcal{M}$  with  $M \subseteq M^+$ , if all  $R \in M^+ \setminus M$  are false of choosing  $x$  from  $X$  and of choosing  $y$  from  $X$ , then  $x \succsim_{M^+}^X y \Leftrightarrow x \succsim_M^X y$ .

# Varying the feasible set: representation theorem

**Theorem 1'.** Let  $\mathcal{M}$  be closed under finite intersection. The agent's preference orders  $\succsim_M^X$  across all the psychologically possible motivating reason sets  $M$  and feasible sets  $X$  satisfy Axioms 1' and 2' if and only if these preferences are given by

$$x \succsim_M^X y \Leftrightarrow \{R \in M : R \text{ is true of choosing } x \text{ from } X\} \geq \{R \in M : R \text{ is true of choosing } y \text{ from } X\}$$

for some fixed weighing relation  $\geq$ .

- *In short:* The axioms hold if and only if preferences go by weighing reasons: an alternatives are ranked according to the weight of their true motivating reasons.

# Proofs

# A preliminary to the proof of Theorem 1

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 Axiom 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  $\{R \in M : y \in R\} = \{R \in M : y' \in R\}$  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  $\{R \in M : y \in R\} = \{R \in M : y' \in R\}$ . By Axiom 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,  $M_x := \{R \in \mathcal{M} : x \in R\}$  for the set of possible reasons in  $M \subseteq \mathcal{P}$  true of  $x \in X$ , and  $\mathbf{C} := \{C \subseteq \mathcal{P} : \bigcap_{R \in C} R \neq \emptyset\}$  for the set of consistent reason sets.

1. First, suppose a relation  $\geq$  on  $\mathbf{C}$  generates all orders  $\succeq_M$ ,  $M \in \mathcal{M}$ . Axiom 2 holds obviously. As for Axiom 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  $\geq$  generates  $\succeq_M$ ,  $M_x \equiv M_x$ , which by  $M_x = M_y$  implies  $M_x \equiv M_y$ . Hence, again using that  $\geq$  generates  $\succeq_M$ ,  $x \sim_M y$ , as desired.

# Proof of Theorem 1 (cont.)

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

*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 finite intersection,  $M \cap M' \in \mathcal{M}$ . We first show that

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

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  $(M \cap M')_y = M_y$ , Axiom 2 yields

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

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

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

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

Lemma 1 yields

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

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  $\geq$  on  $\mathbf{C}$  as follows: for all  $C, D \in \mathbf{C}$ ,  $C \geq D$  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  $M \in \mathcal{M}$ ,  $\geq$  generates  $\succeq_M$ , i.e.  $x \succeq_M y \Leftrightarrow M_x \geq M_y$  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  $M_x \geq M_y$ , 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  $M_x \geq M_y$ . Then, by  $\geq$ '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 1'

Necessity of the axioms is obvious. As for sufficiency, suppose the three axioms hold.

*Claim 1.* For each fixed feasible set  $X$  there is a weighing relation  $\geq^X$  such that for each  $M \in \mathcal{M}$  the order  $\succsim_M^X$  is given by  $x \succsim_M^X y \Leftrightarrow M(x|X) \geq^X M(y|X)$ .

This follows from Theorem 1 by noting that Axioms 1' and 2' imply Axioms 1 and 2 for each fixed  $X$ . To see why Axiom 1' indeed implies Axiom 1 for fixed  $X$ , note that by Axiom 1', applied in the special case that  $X = X'$ ,  $x' = y$  and  $y' = x$ , if  $M(x|X) = M(y|X)$ , then we have  $x \succsim_M^X y \Leftrightarrow y \succsim_M^X x$ ; which implies that  $x \sim_M^X y$  by the completeness of  $\succsim_M^X$ .

## Proof of Theorem 1' (cont.)

*Claim 2.* For all feasible sets  $X, X'$ , alternatives  $x, y \in X$  and  $x', y' \in X'$ , and states  $M, M' \in \mathcal{M}$ , if  $M(x|X) = M'(x'|X') =: S$  and  $M(y|X) = M'(y'|X') =: T$  then  $S \geq^X T \Leftrightarrow S \geq^{X'} T$ . Consider  $X, X', x, y, x', y', M, M'$  such that  $M(x|X) = M'(x'|X') =: S$  and  $M(y|X) = M'(y'|X') =: T$ . As  $\mathcal{M}$  is closed under finite intersection, it contains  $M'' := M \cap M'$ . One easily checks that  $M''(x|X) = M''(x'|X')$  and  $M''(y|X) = M''(y'|X')$ . So, by Axiom 1',  $x \succsim_{M''}^X y \Leftrightarrow x' \succsim_{M''}^{X'} y'$ . Hence,  $M''(x|X) \geq^X M''(y|X) \Leftrightarrow M''(x'|X') \geq^{X'} M''(y'|X')$ . In other words,  $S \geq^X T \Leftrightarrow S \geq^{X'} T$ , q.e.d.

## Proof of Theorem 1' (cont.)

Claim 2 allows us to define a binary relation  $\geq$  on  $\mathbf{C}$  as follows: for all  $S, T \in \mathbf{C}$ ,  $C \geq D$  holds if and only if  $C \succeq^X D$  for *some* (hence by Claim 1 *all*) feasible sets  $X$  such that  $M(x|X) = C$  and  $M(y|X) = D$ .

*Claim 3* (which completes the proof). For each  $M \in \mathcal{M}$  and feasible set  $X$ , the order  $\succeq_M^X$  is given by  $x \succeq_M^X y \Leftrightarrow M(x|X) \geq M(y|X)$ .

This follows from Claim 1 and the definition of  $\geq$ . ■