Rational choice

Word count: 15443

Abstract. Your choice behaviour is rational iff: if it permits a path through a sequence of decisions with a particular outcome, then that outcome is amongst the ones that you would have chosen directly from all possible outcomes of the sequence. This implies, and it is the strongest definition that implies, that anyone who is irrational could be talked out of their own preferences. It also implies weak but non-vacuous constraints on choices over ends. These do not include alpha or beta.

A person *X prefers A* to *B* if she's disposed to take *A* when *B* is the only alternative. *A* is *chosen* by *X* from a set *S* of options if *A* is among the options that *X* would be prepared to take from *S* when constrained to take exactly one. Apples and pears are chosen by me from a menu consisting of apples, oranges and pears if I am prepared to take apples or pears from that menu.

When is choice *rational*? A traditional view identifies rational choice with *means-end* rationality. Given your choices from the possible final outcomes – given your *ends* or *tastes* – it is rational to choose means that conduce to them, irrational to choose means that frustrate them. A definition of rational choice spells out this means-end connection. This essay offers such a definition.

I want a theory of rationality that makes it normatively *engaging*. 'Why should I be rational?' A theory *T* of rationality should say why, and in a way that gives people whose choices are (according to *T*) *not* rational a reason to *change*: a reason that *those people* can appreciate as such, if they are intelligent enough.[1]

> An irrational mode of behaviour is one that I can hope to change by talking to the decision maker, by explaining the theory to him, and so forth. A rational mode of behaviour is one that is likely to remain in the data despite my preaching and teaching.[2]

This very internalist approach to rationality is obviously open to question, but I won't defend it here. If you like, read this essay as tracing the consequences of one natural way, out of many possible ways, of looking at rationality. I am not trying to capture all our intuitions about rationality or 'rationality'. I want a definition that captures this one feature of rationality, its normative engagement, in general and precise terms.[3]

My definition is *constructed* to capture that feature. Anyone who is *ir*rational in my sense could be argued out of their choices. Anyone who is rational in my sense could *not* be argued out of them, not without additional resources.[4] Like Gilboa in the quotation, I treat rational choice as a mode of behaviour (a pattern of dispositions) that is dialectically stable. The main proposal is the *stability* definition of means-end rationality.

My guiding conception looks familiar, but the *extension* of the stability concept turns out to differ sharply from what Gilboa and most other writers have taken rationality to be. It turns out that one might rationally: prefer apples to bananas, bananas to cherries and cherries to apples. One might rationally: choose apples when bananas and cherries are on the menu but not when cherries have been taken off. One might rationally: choose apples or bananas indifferently when they alone are on the menu, but choose apples or cherries, but not bananas, when all three are on the menu. So if means-end rationality has normative force, it is less demanding than you'd think.

Here is the plan. §1 states the stability definition informally (1.1), then more formally (1.2). §2 explains why stability is necessary (2.1) and sufficient (2.2) for normative engagement. §3 argues that stability constrains ends as well as means (3.1). But it allows violations of intuitive conditions like transitivity of preference (3.2). §4 compares stability with four other approaches to characterising rational choice: intuition (4.1), availability of reasons (4.2), immunity to money pumps (4.3) and consequentialism (4.4). §5 sketches some applications. The Appendix extends the definition to cover choice under *uncertainty*.

1 Means-end rationality

This section states the stability definition informally (1.1) and then more formally (1.2). My overall aim is to show how the stability definition capture the idea of normative engagement, and to explain its startling consequences. To do this, I need only consider *deterministic* situations: sequential choices involving no material uncertainty, whose outcome (in so far as you care) depends *solely* on your choices. So the main definition is stated, and its main consequences drawn, for this type of case. The Appendix extends the definition to decision situations involving material uncertainty.

1.1 Informal definition

Tonight you will visit one of two restaurants, *A* and *B*. *A* offers egg sandwiches, ham sandwiches and tuna sandwiches. *B* offers ham sandwiches, salad and tuna sandwiches. At either restaurant you pick *one* thing from the menu. So you have two successive

choices, a choice of restaurant and a choice from its menu, that jointly determine the outcome.

Figure 1 represents this situation. Boxes with arrows coming out of them are *choice nodes* representing decision points. Boxes with no arrows coming out are *outcomes*. Reading left to right: the first node represents the decision between restaurants. The second and third (A and B) each represents a decision from a menu. (Ignore the bold arrows and text for now.)

For instance, you might (a) go up at the first node (b) go up at the second node: restaurant *A* and egg sandwiches. If so, you are choosing *A* when the alternative is *B*, and egg sandwiches when the alternatives are ham sandwiches and tuna sandwiches. The choices (a) and (b) are marked as bold arrows in Figure 1. There is nothing irrational about this combination.

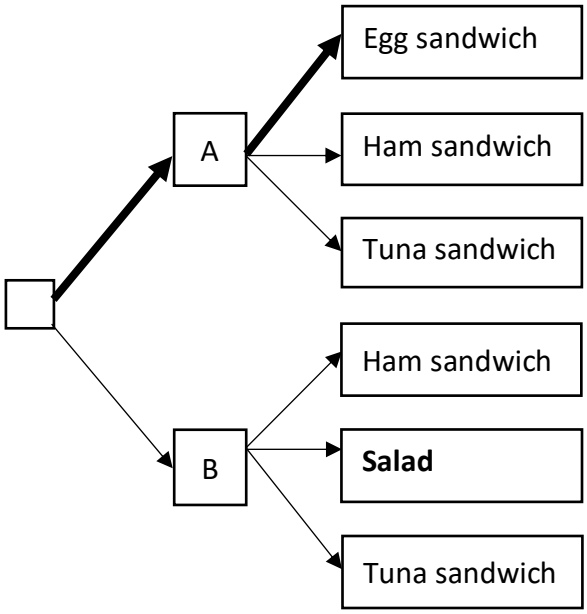
Egg sandwich
A
Ham sandwich
Tuna sandwich
Ham sandwich
B
Salad
Tuna sandwich

Figure 1

We *do* find irrationality if in addition (c) you are *only* prepared to choose salad given a straight choice from all available outcomes i.e. from egg sandwiches, ham sandwiches, salad and tuna sandwiches. 'Salad' is bold in Figure 1 to indicate that this is your favoured outcome.

Combining the dispositions (a)-(c) puts you in this position: your choice of *means* for getting dinner (restaurant *A*) prevents you from realizing your chosen *end* (salad). By choosing *A* you exclude any outcome *that you yourself would have chosen* from the four outcomes that were available *ex ante*.

My stability definition exactly rules out the kind of self-frustration that (a)-(c) involve. Informally: your choice dispositions over means and ends are rational if and only if nobody there could be no sequential decision situation in which your chosen means frustrate your chosen ends.

The idea behind it is that anyone who is irrational in my sense can see by his own lights that his choices are unsatisfactory. To show him this, we show him a sequential decision problem that witnesses this irrationality. Given his choices over *ends* – over the outcomes – he can see that his own choices over *means* – over the pre-terminal options – frustrate those ends. Anyone whose preferences are irrational can therefore *see* what's wrong with them.[5]

1.2 Formal definition

The technical preliminaries, which are simple and familiar, fall under three headings: outcomes, decision trees and choice functions.

1.2.1 Outcomes. Assume a finite set Z of possible **outcomes** or **prizes**. Let Y be a distinguished subset of the power set of Z: a set of subsets of Z representing all possible menus over outcomes. I'll call Y the set of **menus**. I focus on cases where $Y = \wp(Z)$.

For instance, Z may be the set of all possible lunches: bacon sandwich, cheese sandwich, egg sandwich etc. Then Y is the set of all lunch menus that are available at some restaurant. E.g. at a restaurant which offers only cheese sandwiches and egg sandwiches I face the menu $y_1 \in Y$, where $y_1 = \{\text{Cheese sandwich, Egg sandwich}\}$.

1.2.2 Decision trees. We turn now to sequential choice problems and their representation as decision trees.

Define the **level** L of an element z of Z or of a non-empty set S as follows:

(i) $L(z) = 0 \equiv_{\text{def.}} z \in Z$

(ii) If all elements of S have finite level, $L(S) = 1 + \max\{L(S')|S' \in S\}$

(iii) Nothing else has a level.

A **deterministic decision tree** (usually just 'tree') is a set T of finite level. A **node** of such a tree T is any T' such that $T' \in^* T$, where for any relation R I write R^* for the ancestral of R. So $\in^*$ is the ancestral of set-membership: $T' \in^* T$ means that z is an element of T, or an element of an element of T, or an element of an element of an element of T, or… A **terminal node** of a tree is any node of that tree of level 0 i.e. an outcome. If Z' is a set of possible outcomes, then a **tree over Z'** is a tree T such that the set of its terminal nodes is Z'. If T is a tree, then I'll write $\boldsymbol{T}^*$ for the set $Z' \subseteq Z$ that it is a tree over i.e. the set of its

terminal nodes. In other words $T^* = \{z \in Z | z \in^* T\}$. For a given set Z of outcomes, $\boldsymbol{\Delta(Z)}$ is the set of all deterministic trees over non-empty subsets of Z.

In effect this definition treats each non-terminal node of a tree as a set whose elements are its successor nodes, and each terminal choice node as an element of Y. Such a set represents a choice from its members. Any element of a node is either an outcome or itself a tree as well as a node. The elements of any node are the **options** at that node.

For instance (see again Figure 1 above) suppose you can choose whether to dine at A, where the menu is egg sandwiches, ham sandwiches and tuna sandwiches, which we write $A = \{e, h, t\}$, or at B where the menu is ham sandwiches, salad and tuna sandwiches, which we write $B = \{h, s, t\}$. So initially, you are facing a tree over $\{e, h, s, t\}$ of level 2: this is the tree $T_1 =_{\text{def.}} \{A, B\}$. We can write this out in full as:

$$T_1 = \big\{\{e, h, t\}, \{h, s, t\}\big\}$$

I emphasize that this definition only covers trees whose non-terminal nodes are all *choice* nodes. There are no nodes representing resolution of uncertainty. Confining attention to this simplest case lets me convey the central idea more clearly. The Appendix gives the corresponding definitions for choice under uncertainty.

1.2.3 Choice function. A **choice function** C on $\Delta(Z)$ is any function taking non-empty elements of $\Delta(Z)$ – non-empty trees – to non-empty subsets of themselves.

The choice function C encodes the agent's choice dispositions. C **represents** an agent if for any T in its domain, the elements of $C(T)$ are exactly the elements of T that

the agent might take from T if she had to take exactly one.[6] I'll write 'a's choice function' for the C that represents a. We call $C(T)$ the set of options that C **permits** from T.

Preference is the relation revealed by applying C to binary sets:

- $\boldsymbol{C}$ **weakly prefers** $\boldsymbol{a}$ **to** $\boldsymbol{b}$, written $\boldsymbol{a \succsim_C b}$, if $a \in C(\{a, b\})$
- $\boldsymbol{C}$ **strictly prefers** $\boldsymbol{a}$ **to** $\boldsymbol{b}$, written $\boldsymbol{a \succ_C b}$, if $b \notin C(\{a, b\})$
- $\boldsymbol{C}$ **is indifferent between** $\boldsymbol{a}$ **and** $\boldsymbol{b}$**, written** $\boldsymbol{a \sim_C b}$, if $C(\{a, b\}) = \{a, b\}$

When $T \subseteq Z$ these definitions specify the subset of C that constitutes the choice function, and the subsets of $\succsim_C$ and $\succ_C$ that constitute the preference relations, over *outcomes* or ends. But *fully* defined choice functions and preferences allow higher-level nodes (e.g. sets of restaurants as well as menus), to fall under their scope.

Lastly, write $\boldsymbol{T \rightarrow_C T'}$ for $T' \in C(T)$. The set of **outcomes that** $\boldsymbol{C}$ **permits in** $\boldsymbol{T}$ is $\boldsymbol{C^*(T)} = \{z \in Z \cap C(X) | T \rightarrow_C^* X\}$, $\rightarrow_C^*$ the ancestral of the $\rightarrow_C$ relation. Informally, $C^*(T)$ defines the *outcomes that one could reach* by applying choice function C to tree T.

Altering the previous example: suppose you always choose randomly between restaurants, that you always choose ham sandwiches if possible when salad is on the menu, and that you always choose tuna sandwiches if possible when salad is *not* on the menu. Applying the choice function C that represents you, to the tree T_1 and to its nodes, gives these results:

$$C(T_1) = \{A, B\}$$

$$C(A) = \{t\}$$

$$C(B) = \{h\}$$

$$C^*(T_1) = \{h, t\}$$

See the bold *arrows* as marked in Figure 2. (Ignore the bold *outcomes* for now.) $C^*(T_1) = \{h, t\}$ means that your choice dispositions will, if acted upon, realize the outcome in which you get ham sandwiches or the outcome in which you get tuna sandwiches.

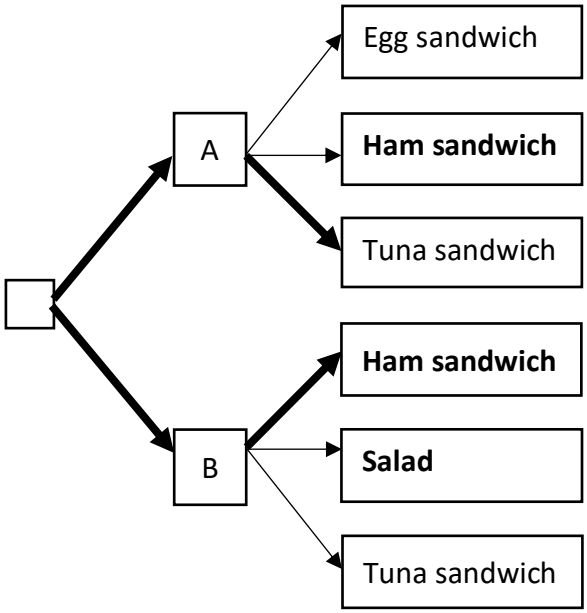
Egg sandwich
A
Ham sandwich
Tuna sandwich
Ham sandwich
B
Salad
Tuna sandwich

Figure 2

1.2.4 Rational choice function. Now the main definition. If C is a choice function on a set $\Delta(Z)$ of trees over a set Z of possible prizes, then:

> **Means-end rationality (stability definition)**: C is a rational choice function over Z if and only if for any $T \in \Delta(Z)$, $C^*(T) \subseteq C(T^*)$.

This means that in any sequential choice C always leads to an outcome that you (i.e. C) would have accepted from those available at the outset.

For instance: still on Figure 2, suppose again that you are indifferent between any two restaurants, that you choose ham sandwiches if possible when salad is on the menu, and that you choose tuna sandwiches if possible whenever salad is *not*. But now imagine that the only foods you would choose in a *straight* choice from the available items are (i) ham sandwiches (ii) salad. So:

$$C(T_1^*) = C(\{e, h, s, t\}) = \{h, s\}$$

These are marked as bold *outcomes* in Figure 2. We know that this choice function, when applied to T_1, leads to either ham sandwiches or tuna sandwiches, i.e. $C^*(T_1) = \{h, t\}$. $C^*(T_1)$ is *not* a subset of $C(T_1^*) = \{h, s\}$. So $C$is means-end *irrational* on the stability definition.

Because it is irrational in that way, *you* can see what is wrong with it. Suppose you contemplate T_1. This tree has four possible outcomes: egg sandwich, ham sandwich, salad, tuna sandwich i.e. $T_1^* = \{e, h, s, t\}$. Of these you would be happy with ham sandwiches or salad, because $C(\{e, h, s, t\}) = \{h, s\}$. But your choice function is liable to deliver tuna sandwiches instead, because $t \in C^*(T_1)$, and this is something you do *not* want *ex ante*.

So you cannot guarantee getting what you want: not because of your ignorance, or because of anything outside your control, but because of your own choices. *Your* choice function is failing to get you where *you* would choose to be! You are as well-placed as anyone else to appreciate this fact. And since *your* ends are being frustrated, you are more likely than anyone else to care.

2 Dialectical stability

This section puts that last point more generally: the stability definition makes means-end rationality both necessary (2.1) and sufficient (2.2) for the dialectical stability that makes it normatively compelling.

2.1 Dialectical stability implies means-end rationality

Let some agent's choice function C be irrational in my sense. There is then a tree T from which C may yield an outcome that C would not have chosen from those available at the outset i.e. $C^*(T) \not\subseteq C(T^*)$. Therefore, there is an outcome $z \in C^*(T) - C(T^*)$ for some $z \in T^*$.

We address the agent as follows. 'The outcomes that you want from this tree T are just the elements of $C(T^*)$. You want to avoid every other outcome in T^*. So z is an outcome that you want to avoid, because $z \notin C(T^*)$. But nothing is *stopping* you from avoiding it. What you get from this tree depends entirely on your choices. But your choices are liable to yield z, because $z \in C(T^*)$. So clearly your choice function is unsatisfactory by your own lights.'

The agent cannot be indifferent to this argument; nor can he resist it. It just does follow from my definition that an irrational choice function is liable to get the agent outcomes that he wants to avoid. The agent can see this. So he can see what is wrong with his choice function. That is why the stability concept is normatively gripping.

But it is gripping in a second way. There are situations where the agent herself will, if she can follow this argument, be *motivated to bypass* her choice function. That is, given a choice between following her own choice function down a tree and being *forced* into an outcome, she strictly prefers the latter. Informally – she would pay to bind herself.

We can put it more precisely using this definition.

Foresightedness. A choice function C on $\Delta(Z)$ is *foresighted* if for any tree $T \in \Delta(Z)$, $C(T) \subseteq \{T' \in T | C^*(T') \subseteq C(T^*)\}$ *whenever the latter is non-empty*.

Foresightedness captures something like *sophistication*: a foresighted choice function represents an agent who makes present choices that will *if possible* get her what she now wants, given her future choices conditional on this or that present choice.[7] It chooses at any node n those successor nodes *if any* to which its own application would result in outcomes that it wanted from those available at n.

Let C be foresighted *and* irrational. Because C is irrational there is a tree T such that $z \in C^*(T) - C(T^*)$ for some (unwanted) $z \in T^*$. Choose some (desirable) $z^* \in C(T^*)$. Let $S = \{T, \{z^*\}\}$: this corresponds to a choice between following one's own choice function along T and being forced into the outcome z^*. Since $z^* \in T^*, S^* = T^*$. So $C(S^*) = C(T^*)$, so $z^* \in C(S^*)$, therefore $C^*(\{z^*\}) \subseteq C(S^*)$. Since T witnesses the irrationality of C it must be that $C^*(T) \nsubseteq C(T^*)$, so $C^*(T) \nsubseteq C(S^*)$. So because C is foresighted, $C(S) = \{\{z^*\}\}$, so $\{z^*\} \succ_C T$. That is: the foresighted but irrational agent strictly prefers being

bound to an outcome of some tree (in this case, z^*) over applying her own choice function to it.

This is a second sense in which stability is normatively attractive, or rather in which instability is normatively repulsive. It is not just that a sufficiently intelligent but practically irrational agent will *regret* his choice function. But also: a foresighted but practically irrational agent will *choose to bypass* her choice function.[8]

To illustrate: suppose that in any binary choice between m and n glasses of wine, you always take m glasses whenever $m > n$; but from *all* available amounts of wine you only want exactly *one* glass. You are at a party where first you get offered one glass; if you accept you get offered a second. Write z_i for the outcome in which you take i glasses total. You are facing the tree $T_3 = \{z_0, \{z_1, z_2\}\}$, Figure 3.

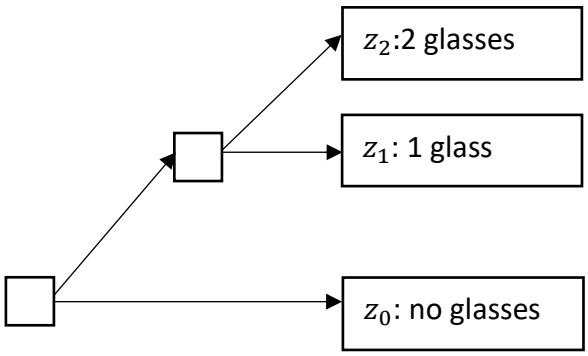

Figure 3

Since optimal consumption is one glass, your choice function satisfies $C(T_3^*) = \{z_1\}$. But since you always prefer more wine to less, if you get to the point where one glass is an option, you will always choose two. So in T_3 your C gets either no wine or two glasses. Either way, it isn't what C itself considers optimal out of those available: $C^*(T_3) \nsubseteq C(T_3^*)$. Your C is irrational.

Imagine now that before the party, the host offers self-binding to one glass of wine but no more. This option doesn't make available any outcome that wasn't already available. (Nothing was stopping you from having just one glass when you got to the party.) But it *does* inevitably yield that outcome, whereas your own choice function leads away from it. So the only way to get what you want is to acquiesce in this self-restriction. And, if your C is foresighted then that is what you (i.e. it) will do.

The stability definition of rational choice thus explains its normativity. It says why you should be rational in a way that moves the *ir*rational. More precisely, it guarantees that we can present to any agent whose C is irrational an argument against it that is compelling by *his own* lights. Exactly *what* the argument is will vary from one irrational C to another, because the T satisfying $C^*(T) \nsubseteq C(T^*)$ must vary from one such C to another. But the definition guarantees that it exists. Means-end irrationality in my sense implies dialectical instability. Equivalently, dialectical stability implies rationality.

2.2 Means-end rationality implies dialectical stability

Suppose an agent's choice function C is rational. Then in *every* tree, C must yield an outcome that it might have chosen from those available *ex ante*: every T satisfies $C^*(T) \subseteq C(T^*)$. We cannot confront the agent with any decision sequence – e.g. a 'money pump' – through which C might generate outcomes she does not want. *We* may consider her

outcomes sub-optimal, but *she* can shrug her shoulders. 'The outcome of my choices may seem bad to *you*; but I'll *always* end up with something that satisfies *me*.'

Imagine e.g. that C makes preference intransitive because you accept small increments of pain for small monetary compensations but not big increments for a big compensation. Specifically: let the set Z of outcomes include vectors (x, y), where x is wealth in dollars, y pain in volts, x and y integers in $[0, N]$ for $N \geq 3$. You weakly prefer (x_2, y_2) to (x_1, y_1) iff: the cube root of the difference between their first components weakly exceeds the difference between their second:

$$(x_2, y_2) \succsim_C (x_1, y_1) \leftrightarrow \sqrt[3]{x_2 - x_1} \geq (y_2 - y_1)$$

(Recall C weakly prefers a to b if and only if $a \in C(\{a, b\})$.) The choice function induces intransitive preferences between these vectors:

$$z_0 = (0,0)$$
$$z_1 = (2,1)$$
$$z_2 = (4,2)$$

For it follows from the definitions of weak and strict preference that C induces a *strict* preference cycle over these vectors (i.e. $z_2 \succ_C z_1 \succ_C z_0 \succ_C z_2$). Intransitivity of strict (and weak) preference is a straightforward consequence.

Now imagine that you start with zero units of money and pain and are twice offered two extra units of money and one extra unit of pain. So you are facing $T_4 = \{\{z_0, z_1\}, \{z_1, z_2\}\}$ (Figure 4).

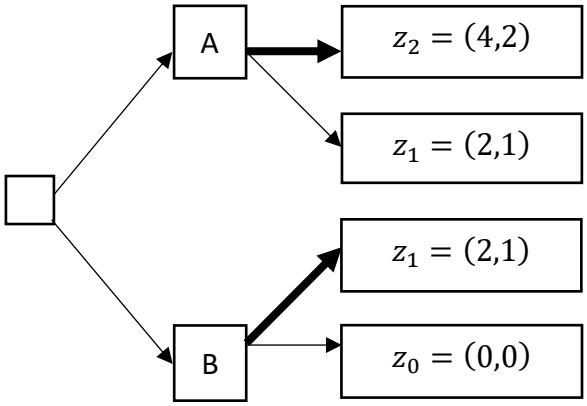

Figure 4

Given $z_2 \succ_C z_1 \succ_C z_0 \succ_C z_2$, you inevitably end up with z_1 or z_2 when you face T_4, because you will inevitably go straight along at A (if you reach it) and up at B (if you reach it): see the bold arrows. Both outcomes are consistent with your preferences as described, because which one you reach depends not only on your preferences between z_0, z_1 and z_2, but also on your preference between the level 1 nodes $A = \{z_2, z_1\}$ and $B = \{z_1, z_0\}$, which were unspecified.[9] Whatever that other preference is, you *will* end up with an outcome to which you strictly prefer another: if you end up with z_1 then you strictly prefer z_2 to what you get, and if you end up with z_2 you strictly prefer z_0. Suppose for definiteness that $C(\{A, B\}) = \{A, B\}$, so both outcomes are possible: $C^*(T_4) = \{z_1, z_2\}$.

But that is no reason for *concern* over how C deals with T. Whether you should be concerned depends on what you – what C – wanted out of T_4 in the first place: that is, on $C(T_4^*) = C(\{z_0, z_1, z_2\})$. Rationality implies $C^*(T_4) \subseteq C(T_4^*)$. Suppose e.g. that $C(T_4^*) = \{z_1, z_2\}$: you regard both vectors as acceptable outcomes of the present adventure. Then your being liable to get z_1 or z_2 needn't worry you at all. Applying C to T_4 yields an outcome that you find acceptable from *all* available outcomes, even though some other outcome is *pairwise* preferred to it. T_4 is therefore no reason either to rue your choice function or to restrict it by self-binding.

In short: means-end rationality implies that there is *no* possible tree in which the outcomes of your choices would not have been acceptable *ex ante*. There is no tree, of which the contemplation could motivate you to abandon, alter, or bind of your own choice function. So rationality implies dialectical stability.

3 Rational taste

Means-end rationality might seem not to constrain preferences over ends as opposed to preferences over means. If rationality is just choosing means that suit our ends, it might seem that any *ends* are as rational as any other. 'It is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger.' But even if rationality is the suiting of means to ends, there could still be rational criticism of ends. A combination of ends can be irrational if there *could* not be means that are suited to them. And although means-end rationality is a constraint on the relation of means (choices over nodes) to ends (choices over outcomes), it *does* exclude some combination of ends.

3.1 Rational tastes

Ends are rationally permissible if they *could* form part of a means-end rational choice function. Anyone whose ends are not rational in this sense does *not* have a rational choice function in the stability sense. If one has irrational ends then one's means, whatever they are, are ill-suited to them.

Here's how I'll cash that out. Let m be a set of outcomes. Suppose we can divide m into possibly overlapping subsets $m_1, m_2, \ldots, m_n$ from *each* of which your choice function C permits something that C would *not* have chosen from m itself. Then *whatever* your choice function over nodes, it is possible to structure your choices over m in such a way that you are liable not to get what you wanted from m, namely the tree $\{m_1, m_2, \ldots, m_n\}$. So this pattern of choices over outcomes implies that C is means-end irrational. Ruling out such a pattern is therefore a necessary condition on rationality of ends. We'll see that it's also sufficient.

We can formalize this with the topological concept of *cover*. A cover of a set S is just a collection of sets that between them include all elements of S. An *exact cover* of S is a collection of sets that between them include *all and only* elements of S. More formally, if Y is a set of menus (a set of sets) and m a menu:

Y is a **cover** of m iff $m \subseteq \bigcup Y$.

Y is an **exact cover** of m iff $\bigcup Y = m$.

For instance, suppose:

$$m = \{\text{Cheese sandwich, Egg sandwich, Ham sandwich}\}$$

$$y_1 = \{\text{Cheese sandwich, Egg sandwich}\}$$

$$y_2 = \{\text{Cheese sandwich, Ham sandwich}\}$$

Then $\{y_1, y_2\}$ is an exact cover of m.

For any choice function we can define its restriction to ends or outcomes, or equivalently the *tastes* that it induces:

If $C$is a choice function over $\Delta(Z)$ then its **taste function** is $\overline{C} =_{\text{def.}} C \restriction \wp(Z)$

The taste function of C is simply its restriction to sets of outcomes. And this is what it is for tastes or end to be rational:

Rationality of ends / tastes: $\overline{C}$ is a rational taste function if: for any $Z' \subseteq Z$ and any perfect cover K of Z', $\exists k \in K \left(\overline{C}(k) \subseteq \overline{C}(Z')\right)$.[10]

As we just saw informally, what justifies this definition is that anyone whose tastes are *not* rational must have an irrational overall choice function i.e. if $\overline{C}$ is an irrational taste then C is means-end irrational. Equivalently, if $\overline{C}$ is an irrational taste then *any* choice function D such that $\overline{D} = \overline{C}$ is means-end irrational. That is, if your ends are irrational then it is not always possible to choose means that will get what you want out of what is available.

The converse also holds. If a taste function $\overline{C}$ is *rational* then *there is some* means-end rational choice function D such that $\overline{D} = \overline{C}$. So if you have rational ends then it is always possible to choose means that will achieve an available outcome that you want.

The formal statement of this connection is as follows:

Means-end rationality and rationality of taste: If C is means-end rational then $\overline{C}$ is a rational taste function. Conversely, if $\overline{C}$ is a rational taste function then it has a means-end rational extension i.e. there is some means-end rational D s.t. $\overline{C} = \overline{D}$.

For proof see footnote.[11]

Rationality of ends is therefore as normatively compelling as means-end rationality. Anyone whose C induces irrational ends $\overline{C}$ can be made to see that her means are inadequate to her ends (because C is means-end irrational). Moreover, she cannot fix this by *adjusting* her means, because *any* choice function D, such that $\overline{D} = \overline{C}$, is *also* means-end irrational. In contrast, anyone whose taste function $\overline{C}$ is *rational* either cannot

be shown that her means are inadequate to her ends, because they never are; or she *can* be brought to see this but can get around it by adjusting her means, for if $\overline{C}$ is a rational taste then *some* means-end rational D is such that $\overline{C} = \overline{D}$.

This definition is very abstract: it doesn't say whether a person with rational tastes must satisfy substantive conditions like transitive preference. It turns out that rational taste (hence rationality) demands very little, as I now argue.

3.2 What rational taste demands

Here are two famous principles of rational choice:

(α): if $A \subseteq B$ and $x \in A \cap C(B)$ then $x \in C(A)$

(β): if $A \subseteq B, x, y \in C(A)$ and $y \in C(B)$ then $x \in C(B)$.[12]

(α) says that nothing chosen from a menu becomes unchosen when you remove other options. If you are prepared to choose apples when bananas and cherries are on the menu then you are prepared to choose them when cherries have been taken off. (β) says that if an originally chosen item remains chosen when you add others, so too do *any* originally chosen items. If you would choose apples or bananas indifferently when they alone are on the menu, then you would not choose apples and *reject* bananas when cherries are also on the menu.

Both principles are intuitively plausible. But rational taste as defined here (hence also means-end rationality) allows *violations* of α and β. More precisely: rational taste

neither entails nor is entailed by α. It neither entails nor is entailed by β. It doesn't even entail their disjunction $\alpha \vee \beta$. But it *does* follow from their conjunction $\alpha\beta$.

This table sets out and justifies the foregoing *non*-entailments. (For proof of the entailment see footnote.[13]) There are choice functions C_1, C_2… whose outputs, when applied to subsets X of a set Z of distinct outcomes a, b, c, are defined in the table. E.g. C_1 selects either of a and b when choosing from $\{a, b, c\}$.[14]

$\boldsymbol{X}$	$\boldsymbol{C_1(X)}$	$\boldsymbol{C_2(X)}$	$\boldsymbol{C_3(X)}$	$\boldsymbol{C_4(X)}$	$\boldsymbol{C_5(X)}$
a, b, c	a, b	a, c	a, b	a	a
a, b	a	a, b	a	a	a, b
a, c	c	a, c	a, c	c	a
b, c	b	c	b	b	b
$\boldsymbol{\alpha}$	No	Yes	No	No	Yes
$\boldsymbol{\beta}$	Yes	No	No	Yes	No
Rational	Yes	Yes	Yes	No	No

Table 1

For instance, C_1 is rational but violates α. It violates α because it permits a and b from $\{a, b, c\}$, but only a from $\{a, b\}$. And yet it is rational: any perfect cover K of any subset X of $Z = \{a, b, c\}$ has an element k from which C_1 chooses only what it would have chosen from X. For instance, suppose $X = Z = \{a, b, c\}$. If K is a perfect cover of X then one of its elements must contain b. So one of its elements is $\{b\}$, $\{a, b\}$, $\{b, c\}$, or $\{a, b, c\}$. But given any of these C_1 always selects only what it would select from X. The same goes for all

subsets of X: so C_1 is a rational taste. So there some means-end rational choice function that extends C_1 to trees.

This violation of α also shows that rational taste does not entail transitivity of weak (or strict) preference. Since $C_1(\{a,b\}) = \{a\}$ and $C_1(\{b,c\}) = \{b\}$ we have $a \succsim_{C_1} b$ and $b \succsim_{C_1} c$ (and $a \succ_{C_1} b$ and $b \succ_{C_1} c$); but since $C_1(\{a,c\}) = \{c\}$ we *don't* have $a \succsim_{C_1} c$ (or $a \succ_{C_1} c$). So transitivity of preference, on the stability view of rationality, is not a demand of rationality.

But isn't it intuitively incoherent to select b from $\{a,b,c\}$ but not from $\{a,b\}$, as C_1 does? It would be odd if you accepted either fruit or ice cream from a dessert menu that also included cheese, but suddenly became averse to ice cream once cheese was off the menu. What does the availability of cheese have to do with whether you prefer fruit to ice cream?

Odd yes – irrational no. Since C_1 is a rational taste on my definition, it follows by §3.1 that some means-end rational choice function C over $\Delta(\{a,b,c\})$ exhibits the – α-violating and intransitive – behaviour of C_1 over those ends, i.e. $\overline{C} = C_1$. If that C is your choice function then we cannot persuade you that anything is wrong with it *by your lights*, however strange it seems to us. For since C is means-end rational, for any tree T, following C through T inevitably yields an outcome that you would have found acceptable at the outset. So why should you see a problem?[15] The same goes for any objections to C_2 and C_3 on similar grounds.[16]

There is more to say about rationality of taste and other conditions, including γ, the Nash Axiom and various kinds of 'path-independence'.[17] For instance, rationality of taste entails γ, which says that if I is a collection of sets of options, if $a \in \bigcap_{i \in I} C(X_i)$ then $a \in C(\bigcup_{i \in I} X_i)$. (If you choose apples when the alternative is bananas, and when the

alternative is cherries, then you choose apples when the alternatives are bananas *and* cherries.)[18] The converse fails – C_4 in Table 1 satisfies γ but is not rational.

But I hope to have said enough to make the main point. If constraints on choice are rational if and only if normatively compelling, then rationality is much less demanding than everyone seems to think. Even a widely accepted principle like α is *not* a demand of rationality, because there are ways to violate α on which your means are unimprovably suited to your ends, and from which therefore *you* could not be persuaded to diverge. Still, rationality is not *empty*. Some conditions, like γ, are legitimate demands on the harmonization of means and ends.[19]

4 Existing theories of rationality

There are four main alternative ways to think about rational choice. (i) Rational choices are those that intuition classes as rational. (ii) They are those for which one can give, or for which there exist, reasons. (iii) They are those that avoid a money pump. (iv) They are those that are consequentialist in the sense of Hammond. This section briefly discusses those approaches. I don't quite *reject* all of them: rather, the stability concept refines both (iii) and (iv). But none of (i)-(iv) does what stability does: isolate the constraints on choice that exert a normative grip.

4.1 Intuitive constraints on rationality

Philosophers often defend norms of rational choice as 'intuitive', by which they mean that they are pre-reflectively reasonable.

For instance, Egan's well-known argument against 'Causal Decision Theory' (CDT) involves two main examples: *Murder Lesion*, where CDT recommends 'shooting'; and *Psychopath Button*, where it recommends 'pressing' (a button). The argument against CDT is that these recommendations are unintuitive. He adds: 'Some people lack the clear intuition of irrationality for the *Murder Lesion* case. Pretty much everyone seems to have the requisite intuition for *Psychopath Button*, however. That's enough for my purposes.'[20] That presupposes that we settle rationality of choice by measuring it not against some pre-defined technical notion but rather against our intuitions.[21]

This characterization of rationality lacks normative grip. There may be nothing we can say to persuade someone who is in this sense 'irrational'. Suppose I consciously follow CDT and endorse 'pressing' in *Psychopath Button*. You upbraid me for 'irrationality' *in the sense of* violating intuition. Yes, I'm irrational in *that* sense. But so what? Given a choice between what seems 'intuitive' to most people and what has optimal effects (by my lights and in my expectation), I as a follower of CDT will choose the latter. It is unclear how you might reply.

Of course intuition *is* important for conceptual analysis – for trying to define the *word* 'rational', in terms of necessary conditions, paradigms or anything else, in a way that tracks its actual everyday uses. Here there *is* some point in respecting the *endoxa*.

But if my project is conceptual anything, it's not conceptual analysis but conceptual *refinement*: taking *one* thing we associate with rationality, namely its normative compulsion, asking what *it* demands, and then constructing a concept that meets these demands. Given the artificiality of this procedure, it is unsurprising (but hardly unwelcome) that its upshot, stability, diverges from our unsystematic, contingent and frequently contestable intuitions about 'rationality'.

4.2 The existence of reasons

The second approach identifies rational choices with those for which one has or can find some reason.

To see what that rules out, consider the *Future-Tuesday Indifferent*: a person who on any day cares equally about his welfare then and on all future days *except* for future Tuesdays. He is completely indifferent to any fortune or suffering on any future Tuesday. For instance, if he has a dental operation on a future Monday, he may willingly pay in advance for anaesthetic to be used. But if the operation is on a future Tuesday then he will *not* pay now, however painful the operation to be and however cheap the anaesthetic.[22]

Parfit, who invented the example, comments:

> This man's pattern of concern is irrational. Why does he prefer agony on Tuesday to mild pain on any other day? Simply because the agony will be on a Tuesday. *This is no reason*. If someone must choose between suffering agony on Tuesday or mild pain on Wednesday, the fact that the agony will be on a Tuesday is no reason for preferring it. Preferring the worse of two pains, for no reason, *is* irrational.[23]

But on the stability definition, choices may be rational even if made for *no* reason: there is nothing irrational about preferring (say) a pain next Tuesday to a pain next Wednesday, even if there is *no* difference between the pains that could rationalize the preference. What stability demands is that the preference is not part of an overall profile of choice-dispositions, over nodes and outcomes, that could in some tree frustrate its own ends.

But there is nothing normatively compelling about the idea that one's choice function ought to be backed by reasons, for there is nothing normatively repellent about choice functions that are not. On questions of taste in the everyday sense, we often *do* tolerate variations for which nobody could give a reason. As a matter of basic preference I prefer the taste of avocado to that of broccoli, but you might be the opposite. You couldn't change my preferences (nor could I change yours) by explaining that there is no *reason* for them. Parfit might say that the timing of future pain *ought* not to be such a matter of taste; but then this is a sense of 'ought' that one can stably violate. Lacking a reason to hold onto this choice function isn't the same thing as *having* a reason to switch to another. So if I don't care about Tuesdays but I do care equally about all other days, Parfit could not show me that I have gone wrong by my own lights: that is, that my means are maladjusted to *my* ends.[24]

4.3 Money pumps

The third approach identifies rationality of a choice function C with the impossibility of a *money pump* for it. A money pump for C is a tree where C yields an outcome that is (a) objectively worse than, or (b) binary dis-preferred to, some available alternative.[25] The bad outcome needn't involve literal *money* loss though typically it does.[26] On the proposed definition, C is rational over Z if and only if no tree in $\Delta(Z)$ is a money pump for C.

For instance, an alleged money pump for the cyclic preference structure $A \succ B \succ C \succ A$ is the Rabinowicz Money Pump (RMP).[27] It works like this: the agent (call her Alice) starts with A and some monetary endowment. We repeatedly offer the trade:

(*) I will give you C for A, B for C or A for B at a charge of 1¢

Alice knows that this morning (Monday), and the next two mornings, she has an opportunity to take up (*). For instance, she might refuse (*) today but accept on Tuesday and Wednesday: so, she gets B (swapping A for C on Tuesday, and C for B on Wednesday) but is 2¢ worse off. Or she might always refuse, leaving her A and her original wealth.

See Figure 5. An upward arrow means 'accept'; a downward arrow means 'reject' (ignore the bold for now). If e.g. Alice accepts (*) on Monday morning and rejects on Tuesday morning ('up' then 'down') then just before Wednesday she has the same as if she had rejected (*) on Monday morning and accepted on Tuesday, namely $C - 1¢$. ('$A - 3¢$' denotes that Alice has paid 3¢ and now holds A. Similarly for '$B - 2¢$' and '$C - 1¢$').

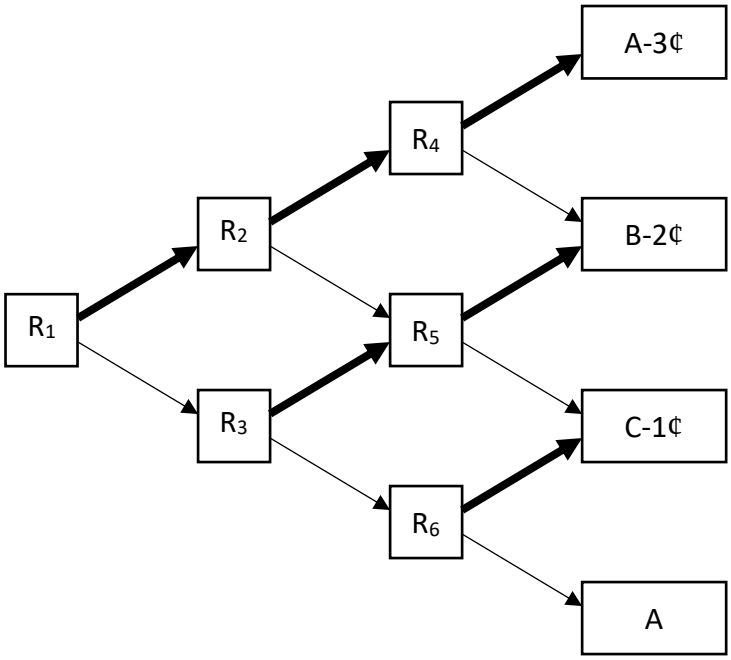

Fig. 5: Rabinowicz Money Pump

What will Alice do? Assume she only cares about her final holding (on Wednesday afternoon) and that addition or subtraction of a single cent makes no difference to her cyclic preferences, so: $A - 3¢ \succ B - 2¢ \succ C - 1¢ \succ A$. Backwards induction shows that she accepts all three offers and ends up with $A - 3¢$. The formal details are tedious, but the bolded arrows give the basic idea. They indicate options that she foreseeably takes if she can. For instance, at R_4 she would go up because she prefers $A - 3$ to $B - 2$, and at R_5 she would go up because she prefers $B - 2$ to $C - 1$ (from now I omit '¢'). Foreseeing this, she would go up at R_2 because she prefers $A - 3$ to $B - 2$. This kind of reasoning motivates her to go up at *every* stage, with outcome $A - 3$, which she binary dis-prefers to an available alternative A. The fact that it yields this outcome seems to show that the cyclic $A \succ B \succ C \succ A$ irrational.

One objection to the money-pump definition concerns its normative force. Suppose either (a) that your C creates an outcome that is in some sense *objectively worse* than an available alternative, or (b) that it creates an outcome in that you *binary dis-prefer* to an available alternative. Is that a reason by *your* lights to change or to bind your choices? Probably not. If you are e.g. Oblomov then you are indifferent about (a); and if you are e.g. Satan then you welcome it.[28] As for (b): even if you strictly *binary prefer* e.g. A over $A - 3$, the latter might still be choice-worthy, for you, from a set of options that includes A, $A - 3$ *and the other available outcomes*.[29]

There is a simple response: amend the definition of a money pump. What a pump gives its victim is not an outcome that she *binary dis-prefers* to some available alternative, but one that she would not have *chosen* from *all* available alternatives. Such a pump *does* exert normative grip. Seeing that her choice function C creates an outcome that she (i.e. C) would not have chosen *ex ante*, she sees that C is frustrating her (i.e. its) own ends. This gives *her* a reason to change it or to bind it.

A second objection is that money pump arguments aim to show the irrationality of choice functions over *outcomes*. Thus the RMP addresses cyclic preferences $A \succ B \succ C \succ A$. But what it really shows is the irrationality of a C that includes not just these preference over outcomes, but also preferences over *nodes* that yield *ex ante* unchosen outcomes. In the RMP, the supposedly disastrous $A-3$ arises not just because of the cyclic preferences; Alice must also have preferences over *nodes* $R_4 \succ R_5$, $R_2 \succ R_3$. The total package is what is responsible for the supposed disaster.

What obscures this is that the preferences $R_4 \succ R_5$ and $R_2 \succ R_3$ can look unquestionable given $A \succ B \succ C \succ A$, because they emerge from the 'sophisticated' or backwards-inductive reasoning summarized in Figure 5. Such sophisticated reasoning itself looks compelling.

But sophisticated reasoning lacks normative grip. There could be somebody who makes unsophisticated choices over the nodes of a tree; but we cannot persuade her that she is going wrong by her own lights.

For instance, suppose (in addition to the foregoing) that Alice's choice function C has these properties:

(i) $C(\{A, B-2, C-1, A-3\}) = \{A, B-2, C-1\}$

(ii) $C(\{B-2, C-1, A-3\}) = \{B-2, C-1, A-3\}$

'Sophisticated' choice demands that at R_1 she prefers R_2 to R_3. But at R_1 she may reason thus: 'Looking at all possible outcomes, I'd be happy with any except $A-3$, by (i). If I go up now, I'm liable to end up with $A-3$ (because at R_2 I'll be indifferent between R_4 and R_5, because of (ii)). If I go down now, I'll certainly avoid it. So I'll go down.'

This 'holistic' reasoning yields different outcomes from 'sophisticated' reasoning. But what makes sophistication better? If Alice, at R_1, proposes to reason holistically, what can we say to persuade her out of it? Stability validates holistic and not sophisticated reasoning: holistic reasoning yields an acceptable outcome, whereas sophisticated reasoning does not.

Sophisticated reasoning looks plausible because if going up yields an outcome that you binary prefer to what you get by going down, then it seems you should go up. But instead of comparing these two outcomes as if they were the only possibilities, you should instead ask whether either of them is choice-worthy given the full field of outcomes that remain possible at that point. Thus at R_2, Alice should ask not whether she *prefers* $A-3$ (which she will get if she now goes up) to $B-2$ (which she will get if she now goes down). She should ask which of those outcomes *is choice-worthy from the whole set* of outcomes available at R_2 i.e. from $\{A-3, B-2, C-1\}$.

There is a reason it is hard to distinguish these questions. If Alice satisfies α and β (see §3.2) then any outcome that is choice-worthy from the set of all outcomes available at R_2 is also strictly preferred to any that is not. And any outcome that is strictly dis-preferred to some other available outcome is *not* choice-worthy from the set of all available outcomes.[30] So α and β make sophisticated reasoning effectively indistinguishable from holistic reasoning.

But if (as here) Alice violates one, sophistication and holism come apart, and it isn't clear why sophistication is more rational. From a stability perspective it *isn't*, at least not here, because it yields the one outcome that everyone agrees is sub-optimal i.e. $A-3$.

All this motivates another modification to the standard interpretation of money pumps. A money-pump establishes the irrationality of a choice function *over nodes as well as outcomes*, not its restriction to outcomes. If we add this modification to the first (that

the outcome of a money pump would not have been chosen from all available outcomes *ex ante*), what we get is equivalent to stability. Stability, and the associated definition of rational taste, could therefore be interpreted as improvements on the 'money pump' approach that retain its pragmatist spirit. What *is* surprising is that (as I argued at §3.2) when we define rational choice in this way, α and β no longer constrain it.

4.4 Consequentialist rationality

The fourth alternative is Hammond's consequentialism. The idea is that the consequentialist cares *only* about outcomes, not her route to them. So for any tree T, the outcomes that a consequentialist choice function C permits should depend only on the possible outcomes of T, not on its shape. The outcomes C permits from T should therefore be the same as it permits from any *other* tree with the same possible outcomes.

We can state this as follows:

> C is a consequentialist-rational (or just 'consequentialist') choice function over Z if and only if, for any $T, S \in \Delta(Z)$, if $T^* = S^*$ then $C^*(T) = C^*(S)$.

It will be convenient to use the following formulation, which is obviously equivalent:

> **Consequentialist rationality**: C is a consequentialist choice function over Z if and only if for any $T \in \Delta(Z)$, $C^*(T) = C(T^*)$.[31]

The difference between consequentialist rationality and stability is simple. On the stability definition, means-end rationality requires that in any tree your choice function

permits *only* the outcomes you might have taken in a straight choice from all available outcomes. Consequentialist rationality requires that your choice function permits *all and* only the outcomes that you might have taken in a straight choice from amongst all available outcomes. Formally, Hammond's definition has $C^*(T) = C(T^*)$ where mine has $C^*(T) \subseteq C(T^*)$.

Consequentialist rationality is not normatively gripping. There is no reason why anyone whose C is stable, but not consequentialist in Hammond's sense, should *care* that it's not consequentialist in Hammond's sense.

For instance, suppose let the set of possible outcomes be $Z = \{a, b, c\}$ and let C be defined on $\Delta(Z)$ as follows:

- $C(Y) = Y$ if $Y \subseteq Z$
- Failing that, $C(Y) = \{y \in Y - \{c\} | c \notin y^*\}$ if the latter is non-empty.
- Failing that, $C(Y) = \{y \in Y - \{b\} | b \notin y^*\}$ if the latter is non-empty.
- Failing that, $C(Y) = Y$

This C doesn't care which of a, b and c is selected in a *straight* choice between them; but if it has at least one *pre*-terminal option then it will always take an option that eliminates the alphabetically last candidate that can be eliminated.

To illustrate more concretely: suppose our Search Committee has just one aim: appoint a suitable candidate. There are three suitable candidates, but we can only appoint one. We could appoint at random; but the HR Department insists that selection proceeds by two stages of elimination. That is, it imposes the tree in Fig. 6.

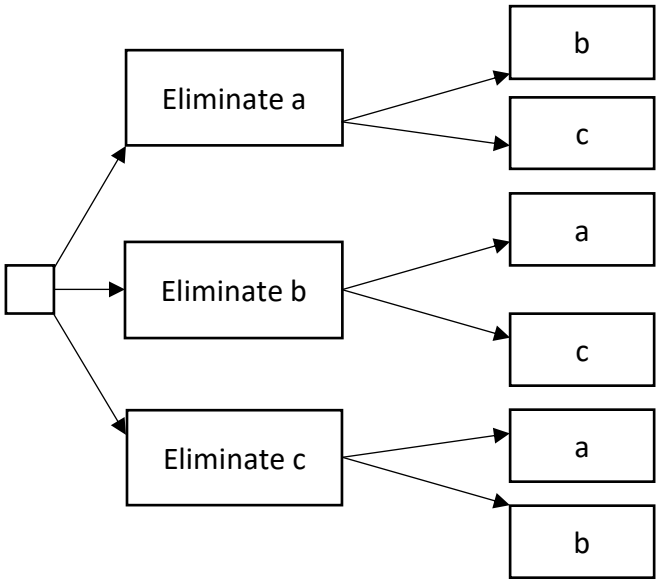

Fig. 6

Suppose we apply C to this tree, call it T_6. So first we eliminate the alphabetically last candidate, and then we choose at random. This means going down at the first node, then choosing a or b. Since all three candidates were initially in the running, $T_6^* = \{a, b, c\}$. Since in a straight choice between them, we'd choose any candidate, $C(T_6^*) = \{a, b, c\}$. But since C goes *down* at the start of T_6, then allows either a or b, $C^*(T_6) = \{a, b\}$. So $C^*(T_6) \neq C(T_6^*)$. Therefore C is *not* consequentialist rational.

But C *does* satisfy the stability definition of rationality. Since $C(Y) = Y$ for any $Y \subseteq Z$, $C^*(T) \subseteq C(T^*) = T^*$ for any $T \in \Delta(Z)$. Specifically, $C^*(T_6)$ is a proper subset of $C(T_6^*)$. So if my definition of rationality is a normatively illegitimate weakening of consequentialism, then it ought to be possible to talk us out of C.

How could anyone do that? You, or HR, might say that the procedure eliminates c unfairly. His being alphabetically last has got nothing to do with his suitability. – Maybe it *is* unfair. But all *we* cared about was appointing a suitable person, not doing so fairly. And we did that. Appeals to fairness cannot show that we are getting anything wrong by our own lights.

Another complaint is that we are inconsistent. Initially we regarded c as an optimal choice from these three candidates. But it is implicit in our procedure that he is *not* optimal, because that procedure prefers a and b to c. Is c optimal or not? – Answer: c *is* optimal. But why should that stop us from eliminating c? What matters is that we eliminate every *sub*-optimal. It doesn't matter if, having done that, we also eliminate a few optimal ones.

That point illustrates the most basic reason to prefer stability to consequentialism. Stability is more consequentialist than consequentialism itself! Stability lets C care about

'more' than the consequences, in that C needn't always return the same outcomes from the same possible outcomes. But it *prioritizes* the consequences: any means-end rational C must subordinate any other principles by restricting them to operate only on what consequentialism allows i.e. on $C(T^*)$. In other words, the difference between stability and consequentialist rationality is that the former makes consequentialism a *side-constraint* on choice, whereas the latter makes it the *sole* determinant of choice.

But a truly thoroughgoing consequentialism *should* regard itself as 'merely' a side-constraint. For as the example shows, the *consequences* of treating consequentialism as a side-constraint are as acceptable as the consequences of admitting no other determinant of choice. Someone who ranks (say) welfare policy *solely* by the number of quality-adjusted life years that it saves, cannot object to a policymaker who always maximizes this quantity, but in case of a tie always chooses e.g. to benefit the materially worst-off.

Hence by treating consequentialism as *more* than a side-constraint, Hammond's 'consequentialist' rationality only imperfectly expresses consequentialism itself. When we correct for this, the upshot is stability.

The situation therefore resembles that at the end of §4.3, where we saw the 'money-pump' criterion of rationality as a flawed expression of the idea behind it; when we correct the flaws, the result is stability. It is interesting that both lines of thought – one starting from money pumps, the other from consequentialism – converge on stability. Perhaps it is an indirect argument for the stability definition. But my main argument for it is that it is the strongest condition that you cannot violate without going wrong by your own lights.[32]

5 Applications

The obvious next steps are (a) to extend the definition to cover uncertainty; (b) to apply it. The appendix makes a start on (a). As for (b): a thorough treatment of any of these cases would double the length of this essay. Here I sketch three applications and mention a few others.

5.1 Supererogation

The first problem case is as follows.

> Suppose that two children are about to be crushed by a collapsing building. You have three options:
>
> [a] do nothing
>
> [b] save one child by allowing your arms to be crushed
>
> [c] save both children by allowing your arms to be crushed.
>
> Here are two very plausible claims about this case:
>
> (1) It is morally permissible for you not to save the children.
>
> (2) It is morally wrong for you to save only one child.[33]

The puzzle arises because:

(3) In a *binary* choice between doing nothing and saving one child, both options are morally permissible.

A function C that selects exactly the morally permissible options from any set of options will therefore satisfy:

- $a \in C(\{a, b, c\})$
- $b \notin C(\{a, b, c\})$
- $C(\{a, b\}) = \{a, b\}$

Any such choice function violates β (see §3.2): it chooses a *and* b from a subset of a set from which it chooses a but *not* b. So if rationality implies β then we must either (i) revise our view about what is morally permissible, or (ii) admit that it would be irrational to be prepared to choose, in any situation, all and only the options that are morally permissible in that situation.

Our definitions of rational choice and rational taste get around the problem, because according to them β is not a demand of rationality. There is a *rational* taste function that selects, from each $X \subseteq \{a, b, c\}$, the options that are intuitively permissible choices from X: in fact it is C_2 in Table 1. The stability definition can therefore show that, and why, these choices are rational.

There is obviously more to say from the perspective of *moral* permissibility. For instance, the argument only shows that the choices (1)-(3) are *rationally* permissible; but there may be other reasons to doubt they are morally permissible. Conversely, and more ambitiously, we might try to show that moral permissibility is structurally no more demanding than rational permissibility, so that the choice function that always selects

what is morally permissible will sometimes violate α, for instance by permitting cycles of strict preference.[34]

5.2 The self-torturer

The next case is fictitious but has wide applications, because 'the safest road to hell is the gradual one'. It concerns a 'self-torturer' who, once a week for the next 100 weeks, has an option to accept an irreversible but indiscernible increment of pain in exchange for \$10,000. After 100 weeks, having accepted the option on *every* week, he is a millionaire but in constant agony.[35] What went wrong?

The set of possible outcomes is $Z = \{z_i | i = 0, \dots 100\}$, z_i indicating pain level of i units and wealth level of $\$10{,}000i$. The terminal nodes – the choices he might face on the last week – can be labelled $n_k^1 = \{z_k, z_{k+1}\}$, $k = 0{,}1 \dots 99$. n_k^1 is the (level 1) node that the agent reaches in the final week iff he has accepted exactly k increments in the previous 99 weeks. We can inductively define $n_k^{j+1} = \{n_k^j, n_{k+1}^j\}$, where $k = 0, \dots 100 - j$, where n_k^j is the level j node that the self-torturer reaches after $100 - j$ weeks if and only if he has already accepted exactly k increments. At the outset he faces the tree $T = n_0^{100}$.

We can argue that the outcome z_{100} is made both inevitable and disastrous by three facts about the self-torturer's choice function C on $\Delta(Z)$.

(1) $0 \leq i \leq 99 \rightarrow z_{i+1} \succ_{\overline{C}} z_i$

(2) $0 \leq j \leq 99 \rightarrow C(n_k^{j+1}) = C(\{n_k^j, n_{k+1}^j\}) = \{n_{k+1}^j\}$

(3) $z_{100} \notin \overline{C}(Z) = C(Z)$

(1) says that given a binary choice between (a) the *outcome* of some final level of wealth and of pain, and (b) the outcome of $10,000 more and indiscernibly more pain, the self-torturer's choice function C (i.e. its taste function $\overline{C}$) always strictly prefers (b). (2) says that at any week in the sequence, the self-torturer is always willing to accept an indiscernible increment of pain in exchange for another $10,000. (1) and (2) imply that that the self-torturer ends up at the maximal point z_{100} i.e. $C^*(T) = \{z_{100}\}$. (3) says that the self-torturer would not have chosen this outcome from those available *ex ante*; that is why it is disastrous.[36]

On any standard view, (1) and (3) jointly imply that the self-torturer has irrational *tastes*, because they imply that $\overline{C}$ violates α.[37] On the other hand (2) looks very plausible given (1), because 'sophisticated' backwards inductive reasoning will convince the self-torturer to accept each increment that he gets offered.[38] So the standard view implies that he went wrong, not because his means ill-suit his ends, but because his ends or tastes, as specified in (1) and (3), are themselves irrational.[39] But this is something of a paradox, since (1) and (3) are, as Quinn says, quite natural.[40]

The stability theory preserves this intuition because it allows that (1) and (3) *are* rational. There is a rational taste function over Z (a choice function on $\wp(Z)$) that satisfies both (1) and (3).[41] §3.2 implies that there is a means-end rational choice function D over $\Delta(Z)$ that agrees with $\overline{C}$ over Z (i.e. such that $\overline{C} = \overline{D}$). In fact there are many such functions. And any of them will specify a route through T that is rationally defensible and does *not* terminate in the disastrous z_{100}. Any of them therefore constitutes a rationally defensible way in which anyone with the self-torturer's (natural) tastes can satisfy them.

Of course none of those functions satisfies (2): there must always some point at which the rational self-torturer *declines* an increment. But only 'sophisticated' reasoning

makes (2) seem rational. And §4.3 implies that rationality does *not* demand sophistication, not if rationality has normative force.

5.3 Causal Decision Theory

Here is a recent problem case for Causal Decision Theory.

> There are two opaque boxes, A and B, and an envelope. The agent can take A, B, or the envelope. The envelope contains $40. One of the boxes contains $100. Which one it is depends on the reliable prediction of a 'Randomizing Frustrater'. If he predicted that the agent takes A, he put $100 in B. If he predicted that the agent takes B, he put $100 in A. If he predicted that the agent takes the envelope, he put $100 in A or B based on the toss of a coin.[42]

Causal Decision Theory (CDT) always recommending taking a box, not the envelope. Letting the options be $\{a, b, e\}$, the choice function C associated with CDT has $e \notin C(\{a, b, e\})$. This is counter-intuitive; but perhaps it is defensible.[43]

However, it can also be shown that CDT recommends *pre-committing* to taking the envelope if this option is available.[44] That is: if T is the tree $\big\{\{a, b, e\}, e\big\}$ then $C(T) = \{e\}$; so $C^*(T) = \{e\}$. Since $T^* = \{a, b, e\}$ it follows that $C^*(T) \not\subseteq C(T^*)$ i.e. C is means-end irrational. So stability puts pressure on CDT from an unexpected direction. (Of course, it may matter that CDT concerns choice under *uncertainty*, whereas the stability definition applies only where no state of nature is both relevant and uncertain. So there is much more to say.)

5.4 Conclusion

Stability may apply to other problems: Kamm's intransitivity paradox for instance,[45] or problems concerning preference aggregation in light of the results of Arrow[46] and Sen.[47] But these are all speculative. I mention them only to encourage investigation into stability, for which the main advertisement remains this: it is the strongest condition that you cannot violate without going wrong by your own lights.

Appendix: Rational choice under uncertainty

The definition of rationality in this essay covers what I called *deterministic* situations i.e. where there is no relevant ignorance d. I don't think this makes it uninteresting. It's obviously interesting e.g. that α and β are not, but γ *is*, a demand of rationality in this simple setting.

Still, the obvious next question is whether uncertainty somehow brings other principles into the picture. It would be infeasible to answer that properly here, but I can sketch an extended definition of means-end rationality to cover that case.

Extending the definition means expanding the set of outcomes and the set of trees that can be built upon them. To this end, Ω be a set of possible worlds (informally: those not ruled out at the outset). Let there be a set Z of prizes. Call any subset E of Ω an *event*. Now we define terminal nodes, choice nodes, natural nodes and a height function that applies to all of them:

(i) A *terminal node* is an ordered pair (n, E) such that $E \subseteq \Omega$ and $n \in Z^E$. If (n, E) is a terminal node then its height is $H\big((n, E)\big) = 0$

(ii) A *choice node* is an ordered pair (n, E) s.t. $E \subseteq \Omega$ and n is a set of ordered pairs (n', E) of finite height. If (n, E) is a choice node then its height is $H\big((n, E)\big) = 1 + \max\big\{H\big((n', E)\big)\big|(n', E) \in n\big\}$.

(iii) *A natural node* is an ordered pair (n, E) s.t. $E \subseteq \Omega$ and n is a set of ordered pairs (n', E') of finite height such that (a) $\{E'|(n', E') \in n\}$ partitions E; (b) if $(n_1, E') \in n$ and $(n_2, E') \in n$ then $n_1 = n_2$. If (n, E) is a natural node its height is $H\big((n, E)\big) = 1 + \max\big\{H\big((n', E')\big)\big|(n', E') \in n\big\}$.

(iv) Nothing else is a node; nothing else has a height.

Intuitively, we think of (n, E) as carrying two pieces of information: n specifies where the agent is in a tree-like structure, and E expresses her knowledge at that point: the set of worlds that might (for all she knows at that point) be actual.

At a terminal node (n, E), n is a function from the set E of still-possible worlds to the set Z of possible prizes: that is, it's a gamble that returns prize $n(w) \in Z$ if the actual world is $w \in E$. At a choice node (n, E), *the agent* chooses between nodes (n', E) at which his information is still E. These nodes are analogues to non-terminal nodes in a deterministic tree. At a natural node (n, E), *nature* chooses between nodes (n', E') at which the agent learns that the actual world belongs to some cell E' of some partition of E. Natural nodes are not analogous to anything in a deterministic tree: they model the evolution of the agent's knowledge over the decision process. A *decision tree with uncertainty* is a node of finite height of the form (n, Ω).[48]

A *choice function under uncertainty* is any function taking any choice node (n, E) to a non-empty subset of n. Because the prize that a choice function C realizes in a tree T depends on which world is actual, the outcomes that C permits in T are not themselves prizes but gambles over prizes i.e. functions from Ω to Z. We shall also consider partial gambles, that is, functions from E to Z for arbitrary $E \subseteq \Omega$ (these include all terminal nodes). We now recursively define $C^*(n, E)$: the outcomes that C permits at an arbitrary node (n, E):

(i) If (n, E) is a terminal node then $C^*\big((n, E)\big) = \{n\}$

(ii) If (n, E) is a choice node and $g \in Z^E$ then $g \in C^*\big((n, E)\big)$ iff $g \in C^*\big((n', E)\big)$ for some $(n', E) \in C(n)$

(iii) If (n, E) is a natural node and $g \in Z^E$ then $g \in C^*\big((n, E)\big)$ iff: there are $E_1 \dots E_k$ that partition E and $n_1 \dots n_k$ s.t. $n = \{(n_i, E_i) | 1 \leq i \leq k\}$, and

$g_1 \ldots g_k$ s.t. for each $i = 1, \ldots k$, $g_i \in Z^{E_i}$ and $g_i \in C^*\big((n_i, E_i)\big)$, and for any world $w \in E$, if $w \in E_i$ then $g(w) = g_i(w)$.

(iv) If T is a tree with uncertainty then the set of outcomes that C permits at T is $C^*(T) \subseteq Z^{\Omega}$.

Informally, the effect of this is that a choice function applied to a tree permits as outcomes a range of gambles over prizes, depending on which state of nature is actual. For example, consider Figure 7.

In this diagram, boxes with arrows going out are choice nodes, circles are natural nodes, and boxes with no arrows going out are terminal nodes. The labelling of the nodes indicates that $\{E_1, E_2\}$ is a partition of Ω and that $\{E_{21}, E_{22}\}$ is a partition of E_2. Let the choice function C make the selections that I indicated in bold: so $C\big((n_1, \Omega)\big) = \{(n_2, \Omega), (n_3, \Omega)\}$ etc. Then the outcomes C permits at $T = (n_1, \Omega)$ are the gambles g, h defined as follows:

- $g(w) = \begin{cases} n_6(w) \text{ if } w \in E_1 \\ n_8(w) \text{ if } w \in E_2 \end{cases}$

- $h(w) = \begin{cases} n_9(w) \text{ if } w \in E_1 \\ n_{10}(w) \text{ if } w \in E_{21} \\ n_{11}(w) \text{ if } w \in E_{22} \end{cases}$

(These definitions make sense because n_6, n_8, n_9, n_{10} and n_{11} are all themselves gambles i.e. functions from possible worlds to prizes.) So in this example, $C^*(n_1, \Omega) = \{g, h\}$.

Which outcomes of a tree are available? In contrast with the deterministic case, one cannot simply collect all terminal nodes, since which terminal nodes are available

depends on which possible world is actual. For instance, if the actual world does not belong to E_1 in Fig. 7 then the terminal node (n_6, E_1) cannot be reached through any sequence of choices.

What *is* always available, whichever world is actual, is any gamble over *all* worlds that is available from *some* sequence of choices. This motivates the following recursive definition:

(i) If (n, E) is a terminal node then $(n, E)^* = \{n\}$

(ii) If (n, E) is a choice node and $g \in Z^E$ then $g \in (n, E)^*$ iff $g \in (n', E)^*$ for some $(n', E) \in n$

(iii) If (n, E) is a natural node and $g \in Z^E$ then $g \in (n, E)^*$ iff: there are $E_1 \dots E_k$ that partition E and $n_1 \dots n_k$ s.t. $n = \{(n_i, E_i) | 1 \leq i \leq k\}$, and $g_1 \dots g_k$ s.t. for each $i = 1, \dots k$, $g_i \in Z^{E_i}$ and $g_i \in C^*\big((n_i, E_i)\big)$, and for any world $w \in E$, if $w \in E_i$ then $g(w) = g_i(w)$.

We now say: if T is a tree with uncertainty then the *set of outcomes available at* T is $T^* \subseteq Z^\Omega$

For instance, in Fig. 7, we see in addition to g and h there is available one other gamble, corresponding to the option of going straight along at (n_4, E_1). This is the gamble:

- $f(w) = \begin{cases} n_7(w) \text{ if } w \in E_1 \\ n_8(w) \text{ if } w \in E_2 \end{cases}$

So the set of gambles available at the tree $T = (n_1, \Omega)$ is $T^* = \{f, g, h\}$.

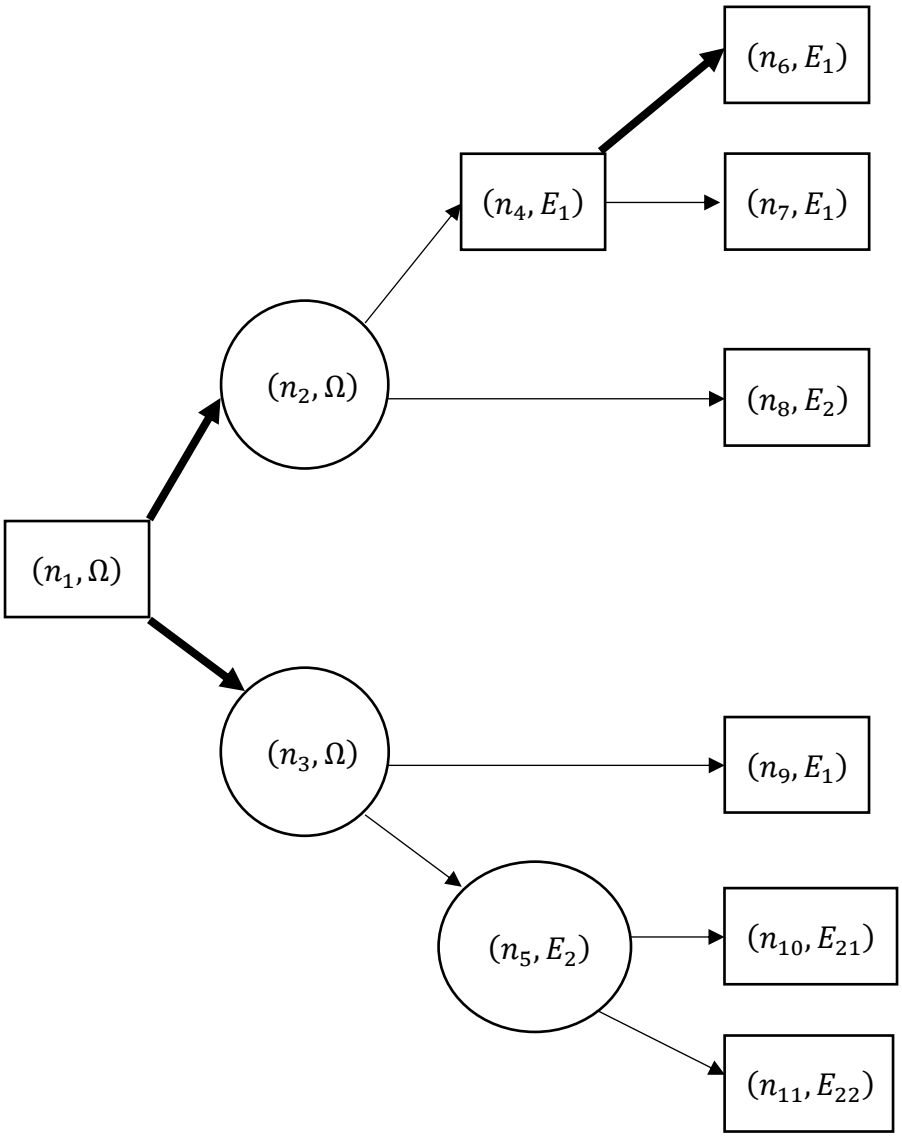
(n_1, Ω)
(n_2, Ω)
(n_3, Ω)
(n_4, E_1)
(n_5, E_2)
(n_6, E_1)
(n_7, E_1)
(n_8, E_2)
(n_9, E_1)
(n_{10}, E_{21})
(n_{11}, E_{22})

Figure 7

Rational choice

We now define means-end rationality as in the deterministic case. A choice function is irrational iff there are trees from which it is liable to select gambles which it itself would not choose, in advance, from all available gambles. This can be written in the same way as before: $C^*(T) \subseteq C(T^*)$ for any T. We can also apply the definition of rational tastes in terms of perfect covers, with tastes defined as preferences over outcomes in the sense of gambles i.e. functions from Ω to Z.

The obvious next step is to identify which 'standard' principles of choice under uncertainty are means-end rational. For instance, consider the following four gambles, where $E_1 \cup E_2 = \Omega$ and $z_1, \dots z_4 \in Z$ are possible prizes.

	E_1	E_2
f	z_1	z_3
g	z_1	z_4
h	z_2	z_3
k	z_2	z_4

Table 2

Suppose some choice function C satisfies $f \succ_C g$ and $k \succ_C h$, so that C violates an 'independence' principle close to Savage's P2. Can we show that C is irrational, on the present definition?

Well, consider the trees in Figures 8A and 8B. Note that the choice node (m_3, E_2) appears in both, so C must permit the same options at that point in each case. The terminal nodes are labelled $z_1, \dots z_4$ to abbreviate gambles that return those prizes for certain, given what one knows at that point about the state of nature.

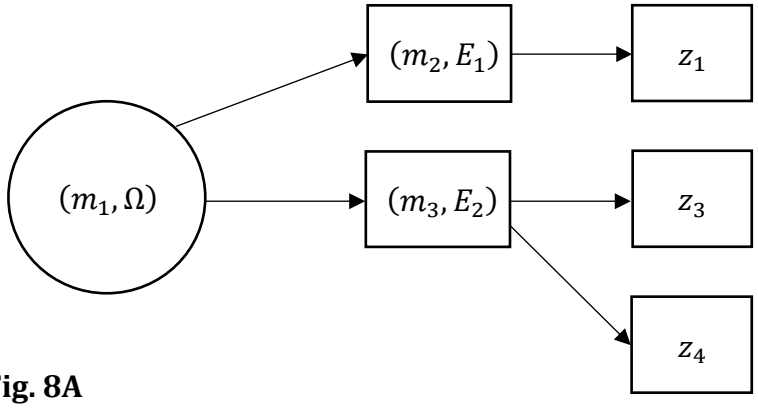

Fig. 8A

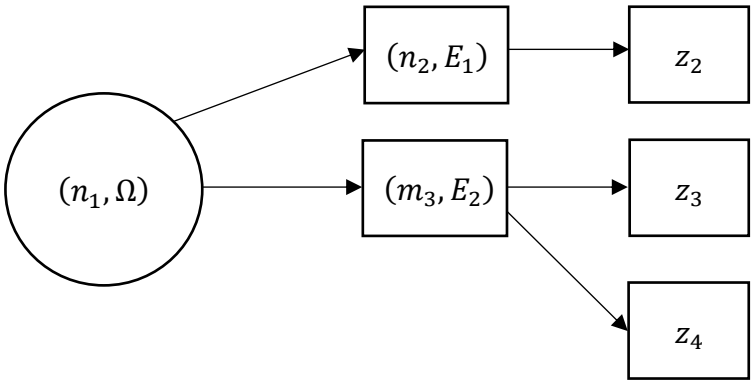

Fig. 8B

It is easy to see that $(m_1, \Omega)^* = \{f, g\}$ and $(n_1, \Omega)^* = \{h, k\}$; so given $f \succ_C g$ and $k \succ_C h$ it follows that $C((m_1, \Omega)^*) = \{f\}$ and $C((n_1, \Omega)^*) = \{k\}$. But (using our abbreviations) either $z_3 \in C\big((n_3, E_2)\big)$ or $z_4 \in C\big((n_3, E_2)\big)$. In the first case, $h \in C^*\big((n_1, \Omega)\big)$ so $C^*\big((n_1, \Omega)\big) \nsubseteq C((n_1, \Omega)^*)$. In the second case $C^*\big((m_1, \Omega)\big) \nsubseteq C((m_1, \Omega)^*)$. Either way, C is means-end irrational. So it turns out that something like Savage's P2 *is* a requirement of means-end rationality under conditions of uncertainty.[49]

It is odd that on the stability definition something as non-obvious as P2 *is* a requirement of rational choice under uncertainty, whereas an 'obvious' principle like transitivity of preference is apparently not.[50] But a systematic treatment of the subject must wait for another occasion.

References

Ahmed, Arif 2014. Dicing with death. *Analysis* 74: 587-94.

Anscombe, Frank J. and Robert J. Aumann. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34: 199-205.

Anscombe, Gertrude Elizabeth M. 2000. *Intention*. Harvard: Harvard UP.

Arrow, Kenneth J. 1951. *Social Choice and Individual Values*. New York: Wiley.

Bossert, Walter 2018. Suzumura-consistent relations: an overview. *International Journal of Economic Theory* 14: 21-34.

——— and K. Suzumura. 2010. *Consistency, Choice and Rationality*. Cambridge, Mass.: Harvard UP.

Bradley, Richard 2017. *Decision Theory with a Human Face*. Cambridge: CUP.

Buchak, Lara 2013. *Risk and Rationality*. Oxford: OUP.

Cantwell, John 2003. On the foundations of pragmatic arguments. *Journal of Philosophy* 100: 383-402.

Chapman, Bruce 2009. Leading you down the choice path: rational persuasion as collective rationality. *Queen's Law Journal* 35: 327-58.

Cubitt, Robin and Robert Sugden. 2001. On money pumps. *Games and Economic Behaviour* 37: 121-60.

Davidson, Donald, John C. C. McKinsey and Patrick Suppes. 1955. Outlines of a formal theory of value, I. *Philosophy of Science* 22: 40-60.

Egan, Andy 2007. Some counterexamples to Causal Decision Theory. *Philosophical Review* 116: 93-114.

Elson, Luke 2016. Tenenbaum and Raffman on vague projects, the self-torturer and the Sorites. *Ethics* 126: 474-88.

Gilboa, Itzhak 2010. *Rational Choice*. Cambridge, Mass.: MIT Press.

Hammond, Peter 1977. Dynamic restrictions on metastatic choice. *Economica* 176: 337-50.

———. 1988. Consequentialist foundations for expected utility theory. *Theory and Decision* 25: 25-78.

Horton, Joseph 2017. The all or nothing problem. *Journal of Philosophy* 114: 94-104.

Joyce, James M. 2018. Deliberation and stability in Newcomb problems and pseudo-Newcomb problems. In Ahmed, A. (ed.), *Newcomb's Problem*. Cambridge: CUP: 138-59.

Kamm, Frances 1985. Supererogation and obligation. *Journal of Philosophy* 82: 118-38.

Lewis, David K. 1981. Causal Decision Theory. *Australasian Journal of Philosophy* 59: 5-30. Reprinted in his *Philosophical Papers Vol. II*. Oxford: OUP (1986): 305-39.

McClennen, Edward F. 1988. *Rationality and Dynamic Choice*. Cambridge: CUP.

Muñoz Daniel 2020. Three paradoxes of supererogation. *Noûs* (online): https://doi.org/10.1111/nous.12326

Parfit, Derek 1984. *Reasons and Persons*. Oxford: OUP.

Peterson, Martin 2017. *An Introduction to Decision Theory*. 2nd ed. Cambridge: CUP.

Quinn, Philip L. 1990. The puzzle of the self-torturer. *Philosophical Studies* 59: 79-90.

Rabinowicz, Wlodek 2000. Money pump with foresight. In Almeida, M. J. (ed.), *Imperceptible Harms and Benefits*. Dordrecht: Kluwer: 123-154.

Railton, Peter 1986. Moral realism. *Philosophical Review* 95: 163-207.

Savage, Leonard J. 1972. *Foundations of Statistics*. 2nd ed. New York: Dover.

Schwartz, Thomas 1972. The myth of the maximum. *Noûs* 6: 97-117.

Sen, Amartya K. 1970. The impossibility of a Paretian liberal. *Journal of Political Economy* 78: 152-7.

———. 1971. Choice functions and revealed preference. *Review of Economic Studies* 38: 307-317.

———. 1993. Internal consistency of choice. *Econometrica* 61: 495-521.

Simon, Herbert A. 1988. Rationality as process and as product of thought. In Bell, D., H. Raiffa, and A. Tversky (ed.), *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge: CUP: 58-77.

Spencer, Jack and Ian Wells 2019. Why take both boxes? *Philosophy and Phenomenological Research* 99: 27-48.

Street, Sharon 2009. In defence of future Tuesday indifference: ideally coherent eccentrics and the contingency of what matters. *Philosophical Issues* 19: 273-98.

Suzumura Kotaro 1976. Remarks on the theory of collective choice. *Economica* 43: 381-90.

———. 1983. *Rational Choice, Collective Decisions and Social Welfare*. Cambridge: CUP.

Temkin, Larry 2012. *Rethinking the Good*. Oxford: OUP.

Tenenbaum, Sergio and Diana Raffman. 2012. Vague projects and the puzzle of the self-torturer. *Ethics* 123: 86-112.

Voorhoeve, Alex and Ken Binmore. 2016. Transitivity, the Sorites paradox and similarity-based decision making. *Erkenntnis* 2006: 101-14.

[1] This constraint is meaningless unless 'intelligent enough' mean something that does not already entail being a rational chooser. That it does, and what this is, will emerge in §§1-2. See n. 5 below.

[2] Gilboa 2010: 5.

[3] Because I start with a choice-theoretic ('revealed preference') approach, the normative question is about why one should have or adopt certain patterns of *behavioural dispositions*. A different approach, more common in philosophy than in economics, identifies preferences with *judgments* (for instance, the judgment that apples are 'all-things-considered' better for me than pears). From this perspective the normative question is about why one should affirm certain *thoughts*. That is a legitimate and interesting question, but not the one I am pursuing here. For more on these two approaches see Bradley 2017: 45-7.

[4] To illustrate the point of 'additional resources': your choices might stem from material error or ignorance. You might choose apples from a menu of apples and oranges because you think, wrongly, that apples have a higher concentration of vitamin C. We could argue you out of this choice by means of an additional resource, namely the information that oranges have a higher concentration of vitamin C. But this does not make your original choice *irrational*.

[5] Clearly one can (e.g. akratically) *see* in the intellectual sense that one will choose means that frustrate one's ends, whilst still being disposed to *make* those choices. So there could be people on whom rationality in my sense exerts a normative pull: people who are disposed to make these choices but who have a reason to suppress or bypass those dispositions (see n. 1).

It can also exert a stronger normative grip. Some people who are irrational in my sense not only have reason to change or bypass their own future choices but *would* do so given the opportunity. More precisely, they would if they are *foresighted* in the sense of §2.1 below; as I argue there, this condition does not entail rationality in the stability sense either. So there could be people who are irrational in my sense but who are still intelligent enough to see this and to do something about it given the opportunity. See n. 8 below.

[6] I understand 'might' subjunctively: the elements of T that an agent *might* take are those that there *would* have been some chance of her taking if she were to face a choice from T. There may be more than one element of T that the agent might take from T, so that $C(T)$ is not a singleton; but the agent will in fact take just one.

[7] On sophisticated choice see McClennen 1990: 11-14; on the difference between foresight and sophistication see n. 9 below and §4.3 below.

[8] Foresight does *not* entail means-end rationality. There may be nodes T at which the *only* available nodes are ones that C itself leads to outcomes that it would not have chosen from all available at T. In this case $\{T' \in T | C^*(T') \subseteq C(T^*)\}$ is empty. This is consistent with foresightedness but not with means-end rationality in the stability sense; so the former does not entail the latter. For a choice function that is foresighted but irrational, see the next example in the main text. (Ulysses would also be an example.)

It matters that foresight does not entail rationality, because it leaves open the possibility of people on whom stability can *exert motivational force*. If an agent is *not* rational but *is* foresighted, then her facing a tree that witness the irrationality will motivate her to bind herself if she can.

[9] 'Sophisticated choice' implies that if $z_2 \succ z_1 \succ z_0$ then $\{z_1, z_2\} \succ \{z_0, z_1\}$: at the first node, you treat your choice from $\{\{z_0, z_1\}, \{z_1, z_2\}\}$ as if it were between the things that your choice function selects from $\{z_0, z_1\}$ and $\{z_1, z_2\}$ respectively. Rationality on my

definition does not imply sophistication: *any* method of selection amongst nodes can be rational so long as it never yields an outcome that it would not have chosen at the outset. Nor does foresight imply sophistication, the difference being that a foresighted C chooses at n only those options that C itself takes to outcomes that are optimal from amongst *all* those available at n, whereas a sophisticated choice function chooses those options that C takes to outcomes that are optimal from amongst all those *that C could reach* from n. See further §4.3.

[10] Cf. Hammond's definition of metastatic consistency (1977: 344). In present terminology, Hammond's condition is that $\overline{C}$ is *metastatically consistent* if: for any $Z' \subseteq Z$ and any perfect cover K of Z', $\forall k \in K \left(k \cap \overline{C}(Z') \neq \emptyset \rightarrow \overline{C}(k) = \overline{C}(Z') \right)$. Metastatic consistency strengthens outcome-rationality in the same way that Hammond's better-known consequentialist consistency requirement strengthens means-end rationality. For further discussion see also n16 and §4.4.

[11] Let $\overline{C}$ be an irrational taste function. Then there is a perfect cover K of some $Z' \subseteq Z$ such that $\forall k \in K \left(\overline{C}(k) \not\subseteq C(Z') \right)$. But K is a tree of level 2 s.t. $K^* = Z'$, and for some $k \in K$, $\overline{C}(k) \subseteq C^*(K)$. Therefore $C^*(K) \not\subseteq C(Z') = C(K^*)$ so C is not rational. (This formalizes the argument outlined at the start of this section.)

Conversely let $\overline{C}$ be a rational taste function. Define D as follows. If $L(T) = 1$, $D(T) =_{\text{def.}} \overline{C}(T)$ (so $\overline{D}(T) = \overline{C}(T)$). If $L(T) \geq 2$, $D(T) =_{\text{def.}} \left\{ S \in T \middle| D^*(S) \subseteq \overline{C}(T^*) \right\}$. Plainly for every tree T we have $D^*(T) \subseteq \overline{C}(T^*) = D(T^*)$. So if $D(T)$ is non-empty for every non-empty tree T then D is a means-end rational choice function. It remains to show that for any (finite) tree T, if T is non-empty then so is $D(T)$, which we prove by induction on $L(T)$. The base step is straightforward: if $L(T) = 1$ then $D(T) = \overline{C}(T)$, which non-empty. Inductive step: suppose that if $L(T) < n$ then if T is non-empty then $D(T)$ is non-empty.

We must consider two cases: (i) the case where $n = 2$ (ii) the case where $n > 2$. (i) Suppose $L(T) = 2$ and let $T = \{S_1 \dots S_m\}$. So T itself is a perfect cover of $T^* = \bigcup_{i=1}^{m} S_i$. By rationality of $\overline{C}$, there is some $S_j \in T$ s.t. $\overline{C}(S_j) \subseteq \overline{C}(T^*)$. Since $L(S_j) = 1$ it follows from the definition of the choice function D that $D(S_j) = \overline{C}(S_j)$; and therefore $D(S_j) \subseteq \overline{C}(T^*)$ and trivially from this that $D^*(S_j) \subseteq \overline{C}(T^*)$. So $S_j \in D(T)$ i.e. $D(T)$ is non-empty. (ii) Now suppose $L(T) = n > 2$ and let $T = \{S_1 \dots S_m\}$. So $\{S_i^*\}_{i=1}^{m}$ is a perfect cover of T^*. By rationality of $\overline{C}$, there is some $S_j \in T$ s.t. $\overline{C}(S_j^*) \subseteq \overline{C}(T^*)$. Moreover $L(S_j) < n$ so by the inductive hypothesis $D(S_j)$ is non-empty i.e. $D^*(R) \subseteq \overline{C}(S_j^*)$ for some $R \in S_j$ and (by the definition of D) for all $R \in D(S_j)$. Hence $D^*(S_j) \subseteq \overline{C}(S_j^*)$. Therefore $D^*(S_j) \subseteq \overline{C}(T^*)$, so $S_j \in D(T)$ is non-empty.

[12] See Sen 1971.

[13] Suppose $\overline{C}$ is an irrational taste function. So some cover K of some set of outcomes Z' is such that for every $k \in K$, $\overline{C}(k) \nsubseteq \overline{C}(Z')$. So for every $k \in K$, $k \nsubseteq C(Z')$. But since K is a perfect cover of Z, there must be some $k \in K$ such that $k \cap \overline{C}(Z')$ is non-empty. Choose one: then either $\overline{C}(k) \cap \overline{C}(Z')$ is empty or it is not. If it is empty, then there is some $a \in k$ that is not chosen from k but is chosen from Z'; but $k \subseteq Z'$ so this violates α. On the other hand, if $k \cap \overline{C}(Z')$ is non-empty then there is some $a \in k$ that is chosen from k *and* from Z'. But since $\overline{C}$ is irrational there is some $b \in k$ that is chosen from k and is *not* chosen from Z'. This violates β. So if $\overline{C}$ is an irrational taste function, then it violates either α or β.

[14] Also given any singleton e.g. $\{a\}$ as input each choice function returns that same set as output.

[15] If you think rational choice *maximizes* something, then violation of α is obviously irrational. There is a tradition in economics that a rational chooser does maximize: she chooses what is in some sense best (Simon 1978: 2). But rationality does *not* demand maximization: there are ways of choosing (e.g. in accordance with C_1) that (a) are not maximizing anything but (b) are normatively stable in the sense that failure to be talked out of them needn't involve any intellectual deficiency (cf. Schwartz 1972). Note also that one might violate α whilst being in some weak sense a 'maximizer' – see Sen 1993: 500f.

[16] C_1, C_2 and C_3 are all metastatically inconsistent in the sense of Hammond 1977 (see n. 9 above). Since they are all rationally defensible – as I just argued – I believe that this shows metastatic consistency to be an excessively strong criterion of rationality.

[17] Suzumura 1983 ch. 2 discusses these and other principles of choice.

[18] Proof that rationality of taste entails γ: suppose C rational and that $a \in \bigcap_{i\in I} C(X_i)$ for some collection $\{X_i\}_{i\in I}$ of subsets of Z. Plainly $\{X_i\}_{i\in I}$ is a perfect cover of $\bigcup_{i\in I} X_i$. Therefore since C is outcome rational, $C(X_j) \subseteq C(\bigcup_{i\in I} X_i)$ for some $j \in I$. Since $a \in \bigcap_{i\in I} C(X_i)$, also $a \in C(X_j)$, therefore $a \in C(\bigcup_{i\in I} X_i)$.

[19] I should here mention two theories of rational choice that also violate at least one of α and β. The first is Suzumura-rationality (Suzumura 1976). Say that a relation R *rationalizes* a choice function C if $C(S) = \{x \in S | \forall y \in S: Rxy\}$ for every S in the domain of C. Say that a relation R is *Suzumura-consistent* if $\forall x \forall y\big((R^*xy . Ryx) \rightarrow Rxy\big)$, where the quantifiers range over the field of R, and R^* is the ancestral of R. A choice function C is *Suzumura-rational* if it has a Suzumura-consistent rationalization. (Roughly, it has no weak-preference cycles of any size in which one or more of the preferences is also strict.) Suzumura-rationality is weaker than the standard notion: there are Suzumura-rational choice functions that violate β and lack a transitive rationalization (see Bossert 2018: 28 for an example). But Suzumura-rationality does entail α, because any *rationalizable*

choice function satisfies α. Suzumura-rationality therefore makes demands on C that are unwarranted from the perspective of the stability theory. For instance, it rules out C_1 in Table 1, even though it is possible (as I argued) to have choice dispositions that conform to C_1 but are normatively irreproachable.

The second theory appears in a ground-breaking paper of Cantwell's that attempts, like this one, to connect normative force and internal coherence (Cantwell 2003). Cantwell identifies two principles: 'strong coherence', which is essentially equivalent to α, and 'weak coherence' which says (in my terms) that if X is a non-empty subset of Z, *some* $a \in C(X)$ is such that $a \in C(Y)$ for *every* Y such that $a \in Y$ and $Y \subseteq X$. Neither condition entails or is entailed by rationality of taste. C_1 is a rational taste function but neither strongly nor weakly coherent; C_5 is an irrational taste function but both weakly *and* strongly coherent. Clearly there *is* something wrong with C_5: when faced with the tree $\{\{a, b\}, \{b, c\}\}$, C_5 is liable to eventuate in an outcome (b) that it would *not* have chosen from those available at the outset.

[20] Egan 2007: 97.

[21] For other 'intuitive' approaches to normative questions, see e.g.: Savage's discussion of the Allais paradox, which emphasizes internal 'reflection' over deductive reasoning (1972: 101-3); Lewis's defence of Causal against Evidential Decision Theory, which just takes a stand on one side of a debate that he regards as deadlocked (1980: 309ff.); Suzumura's endorsement of the Strong and Weak Congruence Axioms (1983: 25); Peterson's endorsement of Egan's judgment about the cases discussed in the main text (2017: 212).

[22] This illustration is from Street 2009 (284ff.), which discusses Future-Tuesday Indifference from a slightly different perspective.

[23] Parfit 1984: 124. Similarly, Anscombe seems to think that one can only want things that there is some intelligible reason for wanting (2000: 70-1). Buchak appears to identify preferences that a reasonable person might have with those that have a consistent rationale (2013: 10).

[24] Although Parfit doesn't say this, one might think that if on Tuesdays the Future-Tuesday Indifferent cares about pains *on that day*, then he *is* open to exploitation, and explaining this might help to talk him out of his future-Tuesday indifference. Imagine that on Sunday the subject is facing 5 units of pain on Tuesday afternoon and 5 on Wednesday afternoon. We offer on Monday to exchange 1 unit of pain on Wednesday afternoon for an additional $1 + 2\Delta$ on Tuesday afternoon, so that he is then facing $6 + 2\Delta$ on Tuesday afternoon and 4 on Wednesday afternoon. And then on Tuesday morning we offer to exchange $1 + \Delta$ on Tuesday afternoon for an additional $1 + \Delta$ on Wednesday afternoon. He will accept both offers for a final schedule of $5 + \Delta$ on Tuesday afternoon and $5 + \Delta$ on Wednesday afternoon, which by his own lights on any day is an uncompensated loss.

So Parfit's character looks irrational on my definition. And pointing this out would persuade him both that something is wrong with his choice function and that it would be a good idea to bind it: or so it seems.

But (a) what this 'pain pump' shows irrational (if anything) is not any one choice function but a combination of two. One, which he has now, prefers any good on any future *non*-Tuesday to the same good next Tuesday. The other, which he has next Tuesday, is indifferent between them. The pain pump does not reveal irrationality in the choice function that gives no weight to Tuesdays – what Street (2009: 285) calls *Consistent Tuesday Indifference*. If the subject adhered consistently to that function – if on all days, *including Tuesdays themselves*, he were indifferent about *any* Tuesday (including the present one) – there would be no pump. (b) Besides, *this* argument could not vindicate

Parfit's idea that *lack of reasons* makes a choice function irrational, because the source of the exploitability is not any lack of reasons for future-Tuesday indifference but rather (what is quite different) the inconsistency between this and his concern for present Tuesdays. (I discuss the 'money pump' definition at §4.3 below.)

[25] Objectively worse: Cubitt and Sugden 2001. Binary dis-preferred: McClennen 1988: 89-90; Bossert and Suzumura 2010: 38.

[26] See e.g. Davidson, McKinsey and Suppes 1955: 145-6. There may be a problem with using money here because nobody (except Scrooge) treats money as an *end*: rather, people want more money because it expands their options. But one may not *want* a larger range of options. One will if one's choice function is maximizing, but the rational force of maximization (which implies α) is among the things that money-pumps purport to *establish*. Still, one could replace monetary losses with other goods (e.g. seconds of pain), avoidance of which *is* a plausible end.

[27] Rabinowicz 2000: 139f.

[28] Cf. Railton's 'Sensible Knave' (1986: 167-8). This is no objection to the use of objective notions in Cubitt and Sugden's paper, because their aim isn't to establish a normative criterion of rationality, but to assess money pumps as *predictive* instruments when the objective good is something like evolutionary fitness.

[29] That would be inconsistent with α (see §3.2); but saying that will seem question-begging to Alice: she already violates α by having cyclic strict preferences.

[30] Let R_2^* be the set of all outcomes available at R_2. Suppose $x \in C(R_2^*)$ and $y \notin C(R_2^*)$. Then by β, $y \notin C(\{x, y\})$ i.e. $x \succ y$. Conversely, suppose that $x, y \in R_2^*$ and $x \succ y$. Then by α, $y \notin C(R_2^*)$.

[31] The equivalence of the definitions holds because (a) $T \in \Delta(Z) \rightarrow T^* \in \Delta(Z)$ (b) $T^{**} = T^*$.

Here is Hammond's formulation. Let β be a *norm* on behaviour – that is, something that determines permissible choices at any point in a tree. Let the function Φ_β specify the outcomes that β allows: that is, for any tree T, $\Phi_\beta(T)$ is the set of outcomes of tree T that β permits. Also let $F(T)$ be the set of all possible outcomes of the tree T. The consequentialist thesis is that 'whenever two decision trees T, T' are *consequentially equivalent* in the sense that $F(T) = F(T')$, then behaviour in the two trees must also be *consequentially equivalent,* in the sense that $\Phi_\beta(T) = \Phi_\beta(T')$. Thus the structure of the decision tree must be irrelevant to the consequences of acceptable or recommended behaviour' (1988: 38). Note that Hammond introduces parameters governing uncertainty about the state of nature and about the outcomes of chance processes, whereas I am dealing only with the simplest case of 'deterministic' choice.

[32] Hammond's consequentialism imposes severe path-independence: the possible outcomes of a procedure for choosing from a set X of outcomes ought to be the same, whatever the structure of the selection procedure. The stability definition is more relaxed: the possible outcomes of the procedure might vary, but they must always form a subset of the outcomes that would be possible in a direct choice from X. Chapman (2009) goes further in the same direction: different selection procedures might result in completely different outcomes, long as the outcomes of a direct choice from X match the outcome of a selection procedure that suitably partitions the relevant issues. I cannot here do more justice to Chapman's discussion. But note that his criterion of rationality is not consequentialist at all, since it has essentially *procedural* elements. It may be more suitably applied to legal decisions, e.g. a finding of guilt or the determination of a sentence, than to business, economic or political ones e.g. about whom to appoint, what to consume or how to vote. (Chapman himself makes similar points, see esp. pp. 343-4.)

[33] Horton 2017: 94; labels [a]-[c] added.

[34] Such arguments would support the main thesis of Temkin 2012.

[35] Quinn 1990.

[36] Quinn does not assert (3) but rather $z_0 \succ z_{100}$; but we need (3) to expose the full force of the puzzle. See Tenenbaum and Raffman 2012: 96.

[37] Suppose $\overline{C}$ satisfies α. Since $\overline{C}(Z)$ is non-empty $z_i \in \overline{C}(Z)$ for some $i < 100$ by (3). Since $i < 100$, $\{z_i, z_{i+1}\} \subseteq Z$, so by α, $z_i \in \{z_i, z_{i+1}\}$ which contradicts (1).

[38] At the first stage the reasoning goes like this. 'I know by (1) I'll accept in week 99, whatever level of pain and wealth I have then reached. Knowing this at week 98, I know that accepting at that stage will change the outcome from z_i to z_{i+1}, for some i that is then known. So at week 98 I'll accept, again by (1). Knowing this at week 97… So I know now, at the outset, that I will accept in all future weeks. So accepting now will only change the outcome from z_{99} to z_{100}. So I'll accept now'. Increasingly truncated versions of this reasoning apply at each subsequent stage.

[39] See e.g. Voorhoeve and Binmore 2006: 103.

[40] Quinn 1990: 80. However, Elson 2016 offers an interesting argument that (1) fails because the marginal disutility of pain increases faster than the marginal utility of money.

[41] For instance, suppose that in addition to (1) and (3) C satisfies these constraints: (4) $Z_i \succ_C Z_{i+k}$ if $k > 1$, and (5) If $Y \subseteq Z$ and $|Y| > 2$ then $C(Y) = Y - \{z_n\}$, where $n = \max \{i | z_i \in Y\}$. That is: the self-torturer always chooses the lower of two settings that are at least two increments apart; and given a choice from more than two settings, he chooses any other than the highest. It can easily be shown that $\overline{C}$ is a rational taste function.

[42] Modified from Spencer and Wells 2019: 34; cf. Ahmed 2014.

[43] Joyce 2018: 155-9 defends this choice in the analogous version of the case in Ahmed 2014.

[44] See Joyce 2018: 157 for an explanation of why this holds in an analogous version of the case.

[45] Kamm 1985. Muñoz 2020 gives a unified treatment of this paradox and of Horton's puzzle of supererogation according to which violations of β, which they both involve, are harmless. The stability definition agrees. But I disagree with Muñoz over *why* they are harmless; and this affects other cases. For instance, Muñoz's explanation does nothing to disturb α (2020: 13), whereas as we saw in §3.2, the stability analysis tends to undermine it.

[46] Although no method of *preference* aggregation satisfies the Arrow conditions, there may be ways to aggregate *rational taste functions* $C_1, C_2, \ldots C_n$, which need not be rankings, to give a *rational social taste function* C, which also need not be a ranking, that is defined on the same domain Z and which satisfies analogues of the Arrow conditions that are appropriate for choice functions that needn't be rankings. For instance, in this context Arrow's independence condition ('Condition 3') would be that for $S \subseteq Z$, if $C_1, \ldots C_n$ and $C_1', \ldots, C_n'$ are two sets of individual choice functions and C and C' the corresponding social choice functions, and if $C_i(S) = C_i'(S)$ for each i, then $C(S) = C'(S)$ (cf. Arrow 1951: 27).

[47] Roughly: given individual rankings $\succsim_i$ of a set Z of outcomes, and an assignment A of rights to each individual i, we might want a social choice function C to satisfy both (i) the Pareto condition that if everyone prefers a to b then the social choice function prefers a to b and (ii) the libertarian condition that if some assignment A of rights gives individual i the right to dispose between a and b, then C agrees with i over the choice between them. It seems that no method of aggregation satisfies (i) and (ii) for arbitrary A and Z (Sen 1970). However, there may be a way to aggregate *rational taste functions* C_i over outcomes to give a *social taste function* C that satisfies analogues of (i) and (ii) that are appropriate for choice functions that needn't be rankings. For instance, in this context the

Pareto condition would be that if there are some elements of a set $X \subseteq Z$ that everyone chooses from X, then the aggregate choice function selects only such elements from X.

[48] This definition of the relevant class of trees is somewhat simpler than Hammond's (1988: 31-2), which also includes chance nodes. In Hammond's model the agent's choice function ranges over *objective* gambles over subjective gambles in the manner of Anscombe and Aumann 1963, whereas here all gambles are subjective in the manner of Savage 1972. This affects the construction of a utility function given a choice function that is consequentialist in Hammond's sense, but not (as far as I can see) the prior issue of what principles of choice are normatively compelling.

[49] Savage 1972: 23. Cf. the proof of Samuelson's 'Independence' principle at Hammond 1988: 42-4.

[50] To establish the latter I'd need to apply the argument of §3.2 to the present, extended definition of rationality. Intuitively that does not seem too hard, but there may be unforeseen difficulties here.