What rationality is

A choice function C is rational iff: if it allows a path through a sequence of decisions with a particular outcome, then that outcome is amongst the ones that C would have chosen from amongst all the 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 can be said to *binary prefer* A to B if she chooses, or is disposed to choose, A when B is the only alternative. I binary prefer apples to oranges if I choose apples when oranges are the only alternative. More generally we can say that for a given person, A is *preferred* from a set S of options if A belongs to the set of things in S that she would choose from a menu consisting of all elements of S. Apples are *preferred* from a menu consisting of apples, oranges and pears if I am prepared to choose apples from that menu.

What preferences are rational? A traditional view of rationality is that it relates means to ends but puts no constraints on the ends themselves. *Given* your preferences over final outcomes – i.e. given what I'll call your *tastes* – it is rational to prefer means that you think conducive to the preferred ends. It is irrational to prefer means that in your own estimation frustrate those ends. But almost *any* ends are rationally permissible. 'It is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger'; but what *would* then be irrational is scratching my finger to save the world.

This essay has two aims. The first is to define means-end rationality in a way that explains its normative force. 'Why be rational?' ought to have an answer that people with *irrational* preferences would find persuasive; but many definitions of rationality violate the condition. The definition offered here is constructed to satisfy it. Anyone who is irrational in my sense could be argued out of their preferences. Anyone who is rational in my sense could *not* be argued out of their preferences. Note that I am *not* looking for a definition that respects all our intuitions about either 'rationality' or rationality. Rather, I am looking for a definition that captures this one feature of the everyday concept in terms that are both general and precise.

The second aim is to show that on this definition, some *tastes* are irrational, in the sense that anyone whose preferences include them could be argued out of their preferences. In outline this result may seem surprisingly strong. One might expect that if rationality is just instrumental or means-end rationality then *de gustibus non est disputandum*. But there is plenty of room for arguing over taste, because some tastes are structured in such a way that *no* preferences that include them could be means-end rational.

But in detail the constraints on taste are surprisingly weak. For instance, many philosophers and economists have thought that rationality demands *transitivity* of binary preference over outcomes: if you choose A when B is the only alternative, and if you choose B when C is the only alternative, then you choose A when C is the only alternative. It turns out that rational binary preference need not be transitive.

The plan is as follows: §1 states my definition of rationality (1.1) informally and then (1.2) more formally. The definition covers only the case of sequential choice under conditions of certainty. §2 explains why the definition makes rationality both (2.1) necessary and (2.2) sufficient for the dialectical stability that makes it normatively compelling. §3 argues (3.1) that the definition constrains one's tastes as well as one's preferences over the means for achieving them; but (3.2) that these constraints are consistent with violations of intuitive conditions like transitivity. §4 compares the present definition with four existing approaches to rationality: (4.1) rationality = intuitive rationality; (4.2) rationality = availability of reasons; (4.3) rationality = immunity to a money pump; (4.4) rationality = Hammond-type consequentialist rationality. In the appendix I sketch an extension of the theory to cover rational choice under uncertainty.

1 Means-end rationality

This section states my definition of rationality (1.1) informally and then (1.2) more formally as applied to the simplest, deterministic case of choice under conditions od certainty.

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 salad, ham sandwiches, and tuna sandwiches. At either restaurant you get to pick one thing on the menu. You therefore have two choices – a choice of restaurant and a choice from its menu – that interact to determine the outcome in ways that this decision tree straightforwardly illustrates:

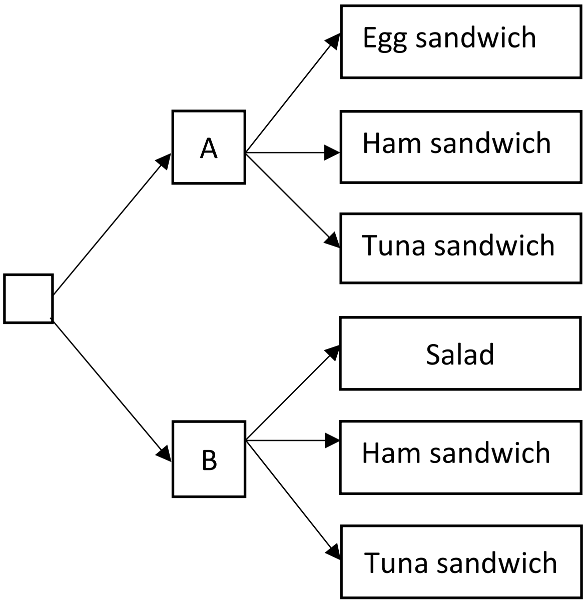

Figure 1

In this tree the boxes with arrows coming out of them are *choice nodes*, representing decision points. Reading from left to right, the first node represents the decision between restaurants and the second and third (labelled A and B) each represents a decision from some menu.

Here, and for simplicity throughout this essay, I'll be considering decision problems which, like this one, are *deterministic*: there is no uncertainty about any state of the world that is relevant to anything that matters to you, so that as far as you are concerned the outcome depends *solely* on the choices that you make. §5.1 sketches one obvious way to extend the ideas presented here to cover this case.

For instance, you might choose (a) to go up at the first node, and (b) to go up at the second node: that is, you choose restaurant A and egg sandwiches, the latter also being the outcome. If so, your choices reflect a preference for A when the alternative is B, and a preference for egg sandwiches when the alternatives are ham sandwiches and tuna sandwiches.

Nothing about this combination of preferences is irrational. But we *do* find irrationality if in addition (c) you prefer (would be prepared to choose) salad given a straight choice between it and all available outcomes – i.e. egg sandwiches, ham sandwiches, tuna sandwiches – and (d) *dis-prefer* (would *not* be prepared to choose) any of those other outcomes given a straight choice between all of them and salad.

Anyone who combines all of (a)-(d) finds himself in the following position: his preferences over the *means* to getting a salad or a sandwich (that is, his preferences over the restaurants) prevent him from realizing his preferences over the *ends* (the salad itself) to which those means are directed. By choosing restaurant A he prevents himself from getting the one outcome that he would have chosen from all of the outcomes (egg sandwich, ham sandwich, tuna sandwich, salad) that were available.

My definition of means-end rationality is that it is just the absence of the kind of self-frustration that (a)-(d) jointly involve. Informally, preferences, over means and ends taken together, are rational if and only if nobody with those preferences could face a sequential decision situation in which the preferences for means frustrate attainment of the preferred ends.

The idea behind this definition is that anyone who is irrational in my sense can in principle see that his preferences are unsatisfactory by his own lights. To get him to see this, we show him a decision tree that witnesses this irrationality. Given his preferences over *ends* – over the set of possible outcomes of the tree – he can see that his own preferences over *means* – over the options at each node of the tree – frustrate the attainment of those ends. Somebody whose preferences are irrational in my sense can therefore be brought to see what is wrong with his preferences.

In this respect rationality as I'll define it resembles Gilboa's conception of rationality as a stability condition:

> 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.[1]

But the *extension* of the concept differs sharply from what either Gilboa (in this work) or other writers on rationality have taken it to be. This should become clear following the formal exposition.

1.2 Formal definition

The technical notions behind the basic approach are very simple and familiar. Here I'll divide them into three categories: outcomes, decision trees and choice functions.

1.2.1 Outcomes. Let there be a finite set Z of possible **outcomes** or 'prizes'. Let there be a distinguished subset Y of the power set of Z i.e. a set of subsets of Z representing all possible choices from outcomes that the agent might face. I'll call Y the set of **menus**. I'll focus mainly on the case where Y is the full power set of Z. For instance, suppose Z is the set of all possible lunch options: bacon sandwich, cheese sandwich, egg sandwich etc. Then Y is the set of all lunch menus that I might face. For instance, 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 Define the **level** L of an element of Z or 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.

[1] Gilboa 2010: 5.

A **deterministic decision tree** is a set T of finite level. A **node** of a tree T is any tree T' such that $T' \in^* T$, where for any relation R I write R^* for the ancestral of R. So $z \in^* T$ means that either z is an element of T, or it is an element of an element of T, or it is 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. If Z' is a set of possible outcomes then a **decision tree over $\boldsymbol{Z'}$** is a decision tree T such that the set of its terminal nodes is Z'. If T is any 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 I'll write $\boldsymbol{\Delta(Z)}$ for the set of all deterministic decision trees over non-empty subsets of Z.

In effect this definition treats each non-terminal node of a decision tree as a set whose elements are its successor nodes, and each terminal choice node as an element of Y. Any element of a node is itself a tree as well as a node. I shall say that the elements of any node are the **available actions** at that node.

For instance, suppose that one day you can choose whether to dine at restaurant A, where the menu is egg sandwiches, ham sandwiches and tuna sandwiches, which we can write as $A = \{e, h, t\}$, or at restaurant B where the menu is salad, ham sandwiches and tuna sandwiches, which we can write as $B = \{s, h, t\}$. So initially, you are facing a decision tree over $\{s, e, h, t\}$ of level 2: this is the tree $T_1 = \{A, B\}$. We can write this out in full as the set:

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

Let me emphasize that this definition of decision trees only covers trees in which all non-terminal nodes are choice nodes. There are no chance nodes. Confining attention to this simplest type of case helps me to convey the central idea as clearly as I can; in §5.1 I'll sketch how the model of rationality presented here might naturally extent to cover trees at some nodes of which nature reveals something relevant about its state.

1.2.3 Choice function. A **choice function** C on $\Delta(Z)$ is any function taking non-empty elements of $\Delta(Z)$ – non-empty deterministic trees – to non-empty subsets of themselves. If $T \in \Delta(Z)$ then $C(T)$ is the set of all elements of T that the choice function **permits** you to select from T. If a and b are (possibly identical) elements of $T = \{a, b\}$ then:

- $\boldsymbol{C}$ **weakly prefers $\boldsymbol{a}$ to $\boldsymbol{b}$**, written $\boldsymbol{a \succsim_C b}$ if $a \in C(T)$
- $\boldsymbol{C}$ **strictly prefers $\boldsymbol{a}$ to $\boldsymbol{b}$**, written $\boldsymbol{a \succ_C b}$, if C doesn't weakly prefer b to a
- $\boldsymbol{C}$ **is indifferent between $\boldsymbol{a}$ and $\boldsymbol{b}$, written $\boldsymbol{a \sim_C b}$**, if C weakly prefers a to b *and* weakly prefers b to a.

In particular, the definitions apply when T is a subset of Z, and when restricted to all such cases, they define the subset of C that constitutes the choice function, and the subset of $\succsim$ that constitutes the weak preference relation, over outcomes or ends. But the full definition of weak and strict preference puts trees of level >1, as well as those of level 1, in the fields of these relations. Finally, writing $\boldsymbol{T \rightarrow_C T'}$ for $T' \in C(T)$, we can define the set of **outcomes that $\boldsymbol{C}$ permits in $\boldsymbol{T}$** to be $\boldsymbol{C^*(T)} = \{z \in Z \cap C(X) | T \rightarrow_C^* X\}$. Informally, $C^*(T)$ defines the *outcomes* that one is liable to reach by applying the choice function C to the tree T.

Continuing the previous example (see also Figure 1), suppose that you always choose to go north when the only alternative is going south, that you choose at random between any two restaurants, and that you always choose ham sandwiches if salad is available and tuna sandwiches if salad is *not* available. This means that applying your choice function C to the tree T_1 and to its nodes gives the following results:

$$
\begin{gathered}
C(T_1) = \{A, B\} \\
C(A) = \{t\} \\
C(B) = \{h\} \\
C^*(T_1) = \{t, h\}
\end{gathered}
$$

We can now state in formal terms the difference between choice functions over trees that are rational and those that are not.

1.2.4 Rational choice function. The formal definition is very simple: if C is a choice function defined on a set $\Delta(Z)$ of trees over a set Z of possible prizes then rationality imposes on C exactly the following constraint:

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

Informally and as already indicated, what this definition says is that if you let your choice function C guide you through any tree T, you are guaranteed to end up with something that you would have been prepared to choose (i.e. that your choice function permits) from all of the outcomes that were available at the outset.

2 Dialectical stability

This section explains why the definition makes means-end rationality (as I'll call it) both necessary (2.1) and sufficient (2.2) for the dialectical stability that makes it normatively compelling. The discussion here will be more informal than elsewhere, in part because the notion of dialectical stability is informal: a choice function is said to be dialectically stable if someone in the grip of it cannot be persuaded that anything is wrong with it.

2.1 Dialectical stability implies means-end rationality

Suppose that an agent's choice function C is in my sense irrational: that is, there is a tree to which applying the choice function is liable to yield an outcome that would not have been chosen from all those available at the outset i.e. there is a T such that $z \in C^*(T) - C(T^*)$ for some $z \in T^*$.

We can then explain things to the agent as follows. The only thing about this tree that matter to you are the outcomes: for any given outcome, what matters is *that* you achieve it, not *how* you achieve it. The outcomes that you want from this tree T^* are just the elements of $C(T^*)$, everything else in T^* being an outcome that you want to avoid. So in particular z is an outcome that you want to avoid, because $z \notin C(T^*)$. And nothing is stopping you from avoiding it: the outcome that you get from this tree depends entirely on the choices that you make at each node. And yet your own choices are liable to issue in z when applied to this tree, because $z \in C(T^*)$. So clearly your choice function is unsatisfactory by your own lights.

The agent cannot be indifferent to this argument if its conclusion is true; nor can she resist any step on the way to it. It just does follow from my definition of irrationality that there are situations in which an irrational choice function is liable to lead the agent into outcomes that she wants to avoid. There are therefore situations in which the agent herself will, if she is minimally foresighted, prefer to abandon her own choice function: given a choice between following her own choice function down a tree and being *forced* to take one of the options that she would like, she will prefer the latter i.e. abrogation of her own power of choice.

We can put this slightly more formally. Call C *foresighted* if $C(T) \subseteq \{T' \in T | C^*(T') \subseteq C(T^*)\}$ whenever the latter is non-empty. A foresighted choice function 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. Now suppose that C is foresighted and irrational i.e. there is a tree T such that $z \in C^*(T) - C(T^*)$ for some $z \in T^*$ and choose some $z^* \in C(T^*)$. Now let $S = \{T, \{z^*\}\}$: this corresponds to a choice between following one's own choice function along T and being forced to take one of the outcomes z^*. It follows from $z^* \in T^*$ that $S^* = T^*$; therefore $C(S^*) = C(T^*)$ so $z^* \in C(S^*)$. Therefore $C^*(\{z\}) \subseteq C(S^*)$. Given that T witnesses the irrationality of C it must be the case that $C^*(T) \nsubseteq C(T^*)$ and therefore $C^*(T) \nsubseteq C(S^*)$. It follows from the foresightedness of C that $C(S) = \{\{z^*\}\}$, so that $\{z^*\} \succ_C T$. In English: the foresighted agent whose choice function is irrational will strictly prefer being forced into an option over letting her own choice function select via some sequence of choices from amongst it and others. This is the sense in which rationality as defined here is normatively attractive, or equally in which irrationality is normatively repulsive.

Suppose for instance that when offered a binary choice between more or fewer glasses of wine, you always take more; but from all the available amounts of wine you most want to take just one glass. Imagine now that you are attending a party at which you are first offered a glass of wine and then if you accept are offered a second glass. If we write z_i for the outcome in which you take i glasses of wine in total, then we can represent this situation as a tree $T = \{z_0, \{z_1, z_2\}\}$. See Figure 2.

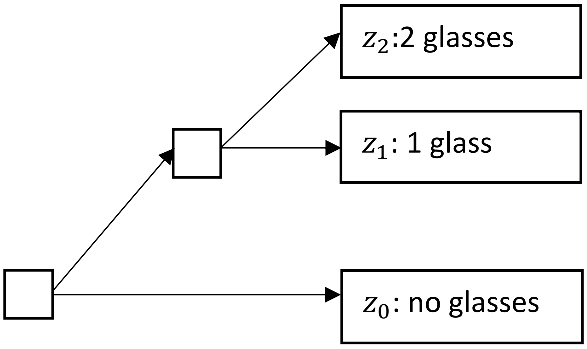

Figure 2

The fact that your optimal consumption is one glass implies that your choice function satisfies $C(T^*) = \{z_1\}$. But the fact that you always prefer more wine to less implies that if you ever get to the point in this tree where one glass is an option, you will always choose two glasses instead. So by applying your choice function C you get either no wine or two glasses, but either way it isn't the outcome that C itself regards as optimal out of those available, so that $C^*(T) \nsubseteq C(T^*)$. So your choice function is irrational on the present definition.

Now suppose that in advance of the party the host offers a self-binding option. If you choose this option, then when you get to the party you will get exactly one glass of wine. This option doesn't make available to you an outcome that wasn't already available to you. (Nothing was stopping you from having just one glass of wine when you got to the party). But it does inevitably lead to that outcome, whereas trusting your own choice function when you get to the party inevitably leads away from it. So the only way for you to get the outcome that your choice function regards as optimal is to acquiesce in the restriction of that choice function, and if you are (i.e. your choice function is) foresighted then that is what you (or it) will choose to do.

I believe that this definition of rationality accounts for its normative grip: that is, it answers the question ‘Why be rational?’ that would appeal even to the *ir*rational. More precisely, the definition guarantees that to any person equipped with an irrational choice function we can present an argument for abandoning it that is compelling by *his own* lights. Exactly *what* the argument is will vary from one irrational C to another, because the tree T satisfying $C^*(T) \nsubseteq C(T^*)$ will vary from one such C to another: but the definition of irrationality guarantees that some such argument can always be found. Means-end irrationality in my sense is therefore incompatible with dialectical stability; equivalently, means-end rationality is necessary for it.

2.2 Means-end rationality implies dialectical stability

Suppose that an agent’s choice function is in my sense rational: there is *no* tree to which applying the choice function is liable to yield an outcome that would not have been chosen from all those available at the outset i.e. every T satisfies $C^*(T) \subseteq C(T^*)$. Then by definition it is impossible to confront its bearer with any hypothetical decision sequence – e.g. any ‘money pump’ – through which C might generate any outcome that she would consider sub-optimal. Whether *we* consider it sub-optimal is beside the point: *she* can always shrug her shoulders at such devices. ‘By following my own choices I might end up with an outcome that seems bad to *you*; but I’ll *always* end up with something that satisfies *me*.’

To illustrate, consider a choice function C that violates transitivity of binary preference, because you are willing to accept small increments of pain for a small monetary compensation but not willing to accept big increments of pain for a big monetary compensation. Specifically: let the set Z of outcomes include all vectors (x, y), where x specifies a monetary holding, y specifies a level of pain, x and y both integers in $[0, N]$ for some large positive N. And suppose your choice function weakly prefers (x_2, y_2) to (x_1, y_1) just in case $x_2 - x_1 \geq (y_2 - y_1)^3$. (Recall that a choice function weakly prefers a to b if and only if $a \in C(\{a, b\})$.) Then this choice function induces intransitive preferences, as follows if we consider these three vectors:

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

It follows from the definition of weak preference that this choice function induces a cycle of *strict* preference over these vectors (i.e. $z_2 \succ_C z_1 \succ_C z_0 \succ_C z_2$); and the failure of transitivity of both strict and weak preference is a straightforward consequence of this.

Now imagine a decision situation in which you start with zero units of money and zero units of pain and are twice offered two units of money in exchange for an increment of one unit of pain. You therefore face the tree $T = \{\{z_0, z_1\}, \{z_1, z_2\}\}$. See Figure 3.

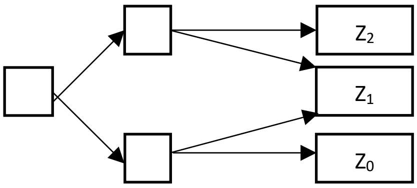

Figure 3

Given your preferences you will inevitably end up with either z_1 or z_2. *Both* outcomes are consistent with your preferences as described, because which one you reach will depend not only on your preferences between z_0 and z_1 and between z_1 and z_2, but also on your preference between $\{z_0, z_1\}$ and $\{z_1, z_2\}$, which I have so far left unspecified.[2] But whatever that other preference is, you *will* end up with an outcome to which you strictly binary prefer some alternative: if you end up with z_1 then you strictly binary prefer z_2 to what you get, and if you end up with z_2 you strictly binary prefer z_0.

But it doesn't follow that you have any reason for *concern* at how your choice function deals with T. That all depends on what you – what your choice function – wanted to get out of T in the first place: that is, on what is in $C(T^*)$. Suppose e.g. that $C(T^*) = T^*$ i.e. that you regard all the possible outcomes of T as acceptable outcomes from that field. Then the fact that you end up with z_1 (or z_2) needn't disturb you at all. Applying your choice function to this tree will leave you with an outcome that you fund acceptable.

Having a means-end rational choice function implies that there is *no* possible tree to which the application of your choice function is liable to eventuate in something that you would not have found acceptable at the outset. Therefore, there is no tree the contemplation of which generates any dialectical pressure on you to abandon, to alter, or to attempt to restrict the application of your own choice function. Rationality implies dialectical stability.

3 Rational taste

Means-end rationality might seem not to constrain preferences over ends, or tastes, as opposed to preferences over the means to achieve them. If rationality *is* just a matter of choosing means that suit our ends, it can look as though any set of ends is as rational as any other. But even if rationality is just a matter of suiting means to ends ,there may still be grounds for criticism of ends from this perspective. A combination of *ends* can irrational if there is no way to select *means* that *could* relate to them in the way that rationality demands. As I'll now argue, it turns out that means-end rationality, although explicitly a constraint on the relation of means (choices over nodes) to ends (choices over outcomes), *does* rule out some ends as irrational by themselves.

3.1 Rational tastes

The intuitive idea is that one's ends are rationally permissible when they form part of a means-end rational choice function. Anyone whose ends are irrational in this sense therefore does *not* have a rational choice function. Irrational ends by themselves imply that one's means, whatever they are, are ill-suited to them.

To formalize this idea we use the topological concept of a *cover*. A **cover** of a set x is just a collection of sets that between them include all the elements of x. An **exact cover** of a set x is just a collection of sets that between them include *all and only* the elements of x. More formally, if Y is a set of menus and m is a menu then Y is a cover of m if $m \subseteq \bigcup Y$. And Y is an exact cover of m iff $\bigcup Y = m$. For instance, suppose:

[2] '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 would select from each of $\{z_0, z_1\}$ and $\{z_1, z_2\}$ if you were to select *it*. But rationality as defined here 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 choice function C chooses at a node 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.

$$m = \{\text{Cheese sandwich, Egg sandwich, Ham sandwich}\}$$
$$y_1 = \{\text{Cheese sandwich, Egg sandwich}\}$$
$$y_2 = \{\text{Cheese sandwich, Ham sandwich}\}$$

Then if $Y = \{y_1, y_2\}$ then Y' is an exact cover of m. In what follows, $K(Y, m)$ will mean that Y is an exact cover of m.

The idea behind the definition of rational tastes (rational choices over outcomes) is as follows. Suppose that we have a set m of outcomes. Suppose there is some way of dividing m into possibly overlapping subsets $m_1, m_2, \ldots, m_n$ from *each* of which C permits the choice of something that it would *not* have chosen from the overall set 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 end up with anything that you wanted from m, namely by confronting you with the tree $\{m_1, m_2, \ldots, m_n\}$. So this pattern of choices over outcomes enough to show that your overall choice function is means-end irrational. Ruling out such a pattern should therefore be a necessary condition on rationality of choices over outcomes. As we shall see it is also sufficient.

Now for the formal definition. For any choice function C let $\overline{C} = C \upharpoonright \wp(Z)$ i.e. the restriction of that choice function to sets of outcomes – informally, the corresponding set of tastes. Then we define:

> **Rationality of ends**: $\overline{C}$ is *outcome-rational* if: for any $Z' \subseteq Z$ and any perfect cover K of Z', $\exists k \in K \left(\overline{C}(k) \subseteq \overline{C}(Z')\right)$[3]

As we just saw, the justification for this definition is that anyone whose choices over ends are *not* outcome-rational must have an irrational choice function i.e. if $\overline{C}$ is outcome-irrational then C is means-end irrational. Equivalently, if $\overline{C}$ is outcome-irrational then *any* choice function D such that $\overline{D} = \overline{C}$ is means-end irrational. That is, if your ends are outcome-irrational then it is not always possible to choose means that will achieve an available outcome that you want.

But the converse is also true: if $\overline{C}$ is outcome-*rational* then *some* choice function D is such that $\overline{D} = \overline{C}$ is means-end rational. That is, if your ends are outcome-rational then it is always possible (in a deterministic setting) to choose means that will achieve an available outcome that you want.

The formal statement of this connection is as follows:

> **Rationality and outcome-rationality**: If C is means-end rational then $\overline{C}$ is outcome-rational. Conversely, any $\overline{C}$ that is outcome-rational has a rational extension i.e. there is some means-end rational D s.t. $\overline{C} = \overline{D}$.

The proof is in a footnote.[4]

[3] Cf. Hammond's definition of metastatic consistency (1977: 344). In the 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', $\forall k \in K \left(\overline{C}(k) = \overline{C}(Z')\right)$. Metastatic consistency strengthens outcome-rationality in just the way that Hammond's better-known consequentialist consistency requirement strengthens means-end rationality. For further discussion see also n. 9 and §4.4.

[4] First, suppose $\overline{C}$ is outcome-irrational. Then there is a perfect cover K of some $Z' \subseteq Z$ such that $\forall k \in K \left(\overline{C}(k) \nsubseteq C(Z')\right)$. 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 $C^*(K) \nsubseteq$

Outcome-rationality is therefore as normatively compelling as means-end rationality. Anyone whose preferences $\overline{C}$ over outcomes are outcome-*ir*rational can be brought to see that her chosen means are inadequate to her own ends (because C is means-end irrational). Moreover, she cannot fix this defect by adjusting her *means* to those ends, because for *any* choice function D such that $\overline{D} = \overline{C}$, D is also means-end irrational. In contrast, anyone whose preferences $\overline{C}$ over outcomes are outcome-*rational* either cannot be brought to see that her means are inadequate to her ends (because they never are), or if she can be brought to see this then she can get around the difficulty by adjusting her means, because if $\overline{C}$ is outcome-rational then there is *some* means-end rational D such that $\overline{C} = \overline{D}$.

This definition of rational taste is somewhat abstract. It doesn't tell us explicitly whether a person with rational tastes must satisfy any substantive conditions, by which I mean constraints like transitivity of binary preference over ends (if you prefer z_1 to z_2 and you prefer $z_2$to z_3 then you prefer z_1 to z_3) or α (if you choose z from A then you choose z from any subset of A that also contains z). It turns out that rationality of taste demands surprisingly little – or so I now argue.

3.2 What rational taste demands

Economists and others have devised a large class of constraints on the choice function for various descriptive or normative purposes. In particular many people think that if C is rational then the following two principles hold for any menus A, B:

(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 $x \in C(B)$[5]

But outcome-rationality as defined here (and therefore also means-end rationality) is consistent with violation of α and β. Overall the situation is as follows: outcome-rationality 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*-entailment claims (for the proof of the entailment claim see this footnote[6]). There are choice functions C_1, $C_2 \ldots$ whose outputs,

$C(Z') = C(K^*)$ so C is not rational. Conversely, suppose $\overline{C}$ is outcome-rational. Define D as follows. If $L(T) = 1$, $D(T) =_{\text{def.}} \overline{C}(T)$ (so $\overline{D}(T) = \overline{C}(T)$). If $L(T) \geq 2$, $D(T) =_{\text{def.}} \{S \in T \mid D^*(S) \subseteq \overline{C}(T^*)\}$. 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 rational choice function. It remains to show that for any (finite) tree T, if T is non-empty then so is $D(T)$. Proof by induction on $L(T)$. The base step is straightforward: if $L(T) = 1$ then $D(T) = \overline{C}(T)$ and this is non-empty. Inductive step: suppose that if $L(T) < n$ then if T is non-empty then $D(T)$ is non-empty. We now have to consider two cases: (i) the case where $n = 2$ (ii) the case where $n > 2$. (i) Suppose $L(T) = 2$ and let $T = \{S_1 \ldots S_m\}$. So T itself is a perfect cover of T^*. By outcome-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 $D(S_j) \subseteq \overline{C}(T^*)$ 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 $T = \{S_1 \ldots S_m\}$. So $\{S_i^*\}_{i=1}^m$ is a perfect cover of T^*. By outcome-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 $D(T)$ is non-empty.

[5] See Sen 1971. α is sometimes also called the Chernoff Axiom or contraction consistency.

[6] Suppose $\overline{C}$ is outcome-irrational. 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

when applied to subsets X of a set Z of outcomes a, b and c, are defined in the table. For instance, the choice function C_1 selects either of a and b when choosing from $\{a, b, c\}$.[7]

$\boldsymbol{X}$	$\boldsymbol{C_1(X)}$	$\boldsymbol{C_2(X)}$	$\boldsymbol{C_3(X)}$	$\boldsymbol{C_4(X)}$	$\boldsymbol{C_5(X)}$
a, b, c	a, b	a, b	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	b	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

We can now see e.g. that C_1 violates α but is outcome-rational. It violates α because although it permits the selection of a and of b from $\{a, b, c\}$, it only permits the selection of a from a and b. (You can think of C_1 as permitting the choice of an element from any set iff it is preferred to some element in that set.) But it is, or rather its restriction $\overline{C_1}$ is, outcome-rational: any perfect cover K of any subset X of $Z = \{a, b, c\}$ has an element k from which C_1 chooses only elements that it would have chosen from the full set X of available outcomes. For instance, suppose $X = Z = \{a, b, c\}$. It follows that if K is a perfect cover of X then one of its elements must be a subset of X containing b. So one of its elements is $\{b\}$, or $\{a, b\}$, or $\{b, c\}$, or $\{a, b, c\}$. But (as we can see from the first column in the body of the table), given any of *these* sets C_1 always selects only items that it would select from Z itself. So $\overline{C_1}$ is outcome-rational; and that means that there is a rational choice function that agrees with C_1 over ends, that is, over subsets of Z.

But isn't there *something* intuitively irrational about selecting b from $\{a, b, c\}$ but not from $\{a, b\}$, as any extension of $\overline{C_1}$ does? After all, it would be strange if you were willing to take (say) either fruit or ice cream from a desert menu that also included cheese, but suddenly became averse to ice cream once cheese was off the menu. What does the presence or absence of cheese from the menu have to do with whether you prefer fruit to ice cream?

But although such a pattern of choice behaviour is *unusual* there is nothing *irrational* about it. What can we say that might convince you to try to change it, or to consent to be forced out of it, or at least see that something is wrong with it by your own lights? There may be *no* sequential-choice situation in which your own choices *ever* lead you to an outcome that *you* would not have wanted from that situation. And, if your choice function is means-end rational in my sense, then there *could* be no such situation. For any tree whose terminal points are exactly a, b and c, following this rational extension of C_1 leads to the outcome a or b So why

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 outcome-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 outcome-irrational, then it violates either α or β.

[7] Note also that given any singleton e.g. $\{a\}$ as input each choice function returns that set as output.

should you change it?[8] The same points apply mutatis mutandis to any intuitive objections to C_2 and C_3 on grounds that they violate β.[9]

It would be possible to say much more about the relationship between outcome-rationality and the many other conditions on choice functions that philosophers and economists have studied, such as γ, the Nash Axiom and various kinds of 'path-independence'.[10] For instance, it is clear that outcome-rationality entails the γ axiom (sometimes called expansion-consistency), which says that if $a \in \bigcap_{i\in I} C(X_i)$ then $a \in C(\bigcup_{i\in I} X_i)$, but the converse is false, since (e.g.) C_4 in Table 1 satisfies the γ axiom but is not outcome-rational.[11]

But I hope to have said enough to make the point. If constraints on choice are rational if and only if normatively compelling, then rationality is much less demanding than it appears on the standard picture. Even so widely accepted a principle as α isn't a demand of *rationality*, because there are ways of violating it according to which you means are by your own lights unimprovably suited to your ends, and from which therefore you could not be persuaded to diverge without coercion. On the other hand, neither is the notion of choice-functional rationality completely empty: there are some conditions, like γ, that make legitimate demands on the harmonization of means and ends.

4 Existing theories of rationality

There are four main ways to think about choice-functional rationality. (i) rational preferences are those that intuition classes as rational. (ii) Rational preferences are those for which one can give, or for which there exist, good reasons. (iii) Rational preferences are those that avoid the possibility of a money pump. (iv) Rational preferences are those that are consequentialist in Hammond's sense. I don't intend to *reject* these approaches; and indeed the third and fourth are closely related to mean-end rationality. But I will argue that none of them do what I have sought to do, namely, to isolate constraints on choice that exert a normative grip.

4.1 Intuitive constraints on rationality

Philosophers often defend some putative norm of rational choice on the grounds that it is 'intuitive', by which is meant something like this: it seems pre-reflectively reasonable. For

[8] Of course if you think that rational choice must in a fairly strong sense *maximize* something then violation of α is obviously irrational. There is a long tradition particularly in economics that a rational chooser does maximize: she chooses what is in some sense best (Simon 1978: 2). But nothing in the concept of rationality demands maximization, in the sense that there are ways of choosing (e.g. in accordance with C_1) that (a) do not involve the maximizing of anything but (b) are normatively stable in the sense that failure to be talked out of them needn't involve any intellectual deficiency. Note also that one might violate α whilst being in some weaker sense a 'maximizer' – for discussion see Sen 1993: 500f.

[9] Note also that C_1, C_2 and C_3 are also all metastatically inconsistent in the sense of Hammond 1977 (see n. 3 above). Since (for reasons given in the text) they are all rationally defensible, I believe that this shows that metastatic consistency is too strong as a criterion of rationality of a choice function over Z.

[10] Suzumura 1983 ch. 2 discusses these and other principles of choice. It is also worth mentioning the principles of rationality that Cantwell identifies in a paper that attempts, like this one, to spell out the connection between normative force and internal coherence (2003). He actually 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 $Y \subseteq X$ s.t. $a \in Y$. Neither of these conditions (a) entails or (b) is entailed by outcome-rationality. Proof: (a) C_1 is outcome-rational but neither strongly nor weakly coherent; (b) C_5 is outcome-irrational but both weakly *and* strongly coherent. And clearly there is something wrong with C_5: anyone whose choice-function it is can see that 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.

[11] Proof that outcome-irrationality entails γ: suppose C outcome-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 $a \in C(\bigcup_{i\in I} X_i)$.

example, Egan's basic case against 'Causal Decision Theory' consists of two examples: *The Murder Lesion*, in which Causal Decision Theory recommends an option of 'shooting'; and *The Psychopath Button*, in which it recommends an option of 'pressing' (a button). The argument that Causal Decision Theory is wrong is the assertion that these recommendations are intuitively irrational. 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 *The Psychopath Button*, however. That's enough for my purposes.'[12] Egan is clearly presupposing that we settle questions of rationality not by measuring this or that choice or preference against some pre-defined technical notion but rather that we do so by measuring it against our intuitions about what is rationally acceptable.[13]

This characterization of rationality lacks any normative grip, by which I mean that there is nothing that we can say to persuade anyone who diverges from it. Suppose that I consciously follow Causal Decision Theory and so endorse the 'pressing' option in Egan's *Psychopath Button*. Suppose that you, or Egan, then upbraids me for 'irrationality' *in the sense of* violating people's intuitions about what to do in this example. My reply is that yes, it is indeed irrational in *that* sense, but why should I *care* about being irrational in that sense? If I have a choice between the option that seems sensible to the vast majority of people, and the option that has optimal effects (by my lights and in my expectation) then as a follower of Causal Decision Theory I'll always choose the latter. It isn't clear what more one might say to object to this.

Of course intuition *does* play an important role if we are engaging in conceptual analysis – trying to give an account of the *word* 'rationality', in terms of necessary conditions, paradigms or anything else, that tracks its actual everyday uses. In that case there is some point in respecting the *endoxa*; or at least there would be, if – as I doubt – all or most of our uses of that word correspond to just one concept.

But if my project could be described as conceptual anything then it is not conceptual analysis but conceptual *refinement*. I am taking *one* feature that we often associate with rationality, namely its normative compulsion, and asking what *it* demands of choice behaviour. It is hardly surprising that the outcome, means-end rationality, is a purified concept that diverges in many places from our diverse, unsystematic, historically contingent and frequently contestable intuitions about 'rationality'.

4.2 The existence of reasons

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

To illustrate what this rules out, consider what Street calls the *Future-Tuesday Indifferent*: a person who on any day cares equally about his welfare on that day and on all future days *except* for future Tuesdays. He is completely indifferent to any fortune or suffering that he might incur on any future Tuesday.

For instance, if he is scheduled for a dental operation on a *Monday*, then he may willingly pay the dentist in advance to ensure that anaesthetic is used during the operation. But if the operation is scheduled for a *Tuesday*, then he will *not* put down any money now, however painful he expects the operation to be and however inexpensive the anaesthetic is.[14]

[12] Egan 2007: 97.

[13] For other examples of the 'intuitive' approach 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 simply 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); and Peterson's endorsement of Egan's judgment about the cases discussed in the main text (2017: 212).

[14] For this example, and for illuminating discussion of Future-Tuesday Indifference from a slightly different perspective, see Street 2009: 284ff.

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.[15]

The definition of rationality in this essay is far more relaxed than what Parfit seems to mean, because it allows that preferences can be rational even if you can't give *any* reasons for holding them. There is, on my definition, nothing irrational about your preference (say) for a pain tomorrow over a qualitatively identical pain today, even though you can't point to *any* difference between them that could rationalize it. All that is required is that this preference not belong to an overall profile of preferences, over nodes and over outcomes, that could in some tree frustrate its own ends i.e. its own preferences over outcomes.

But there is nothing normatively compelling about choice functions for which one has reasons (or 'good' reasons). Equivalently, there is nothing normatively repellent per se about choice functions that lack them. When it comes to matters of what we call taste, we often *do* tolerate variations in preference for which nobody has a reason. As a matter of basic preference I choose avocado over broccoli or cauliflower, and similarly you choose broccoli over avocado or cauliflower. More to the point, pointing out to you or me that there is no reason for preferring one of these over the others will not change anyone's mind: that is just what is meant by saying that these are matters of taste. Perhaps Parfit will say that the timing of one's own future pain *ought* not to be a matter of taste; but then this is a sense of 'ought' that one can stably violate. Lacking a reason for holding onto this choice function isn't the same thing as *having* a reason for switching to another one. Thus e.g., if I don't care about Tuesdays but I do care equally about all other days, it isn't clear how Parfit could show me how I have gone wrong by my own lights: that is, how my means are maladjusted to *my* ends.[16]

[15] 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).

[16] Although Parfit doesn't say this, one might think that if on *Tuesdays* this character is neutral between present and all future pains, then he *is* in fact open to a kind of exploitation; and foreseeing this fact might be a way to talk him out of his preferences.

The argument for this is as follows. If the subject in a given week is facing 5 units of pain on Tuesday afternoon and 5 on Wednesday, we first offer on Monday to exchange 1 unit of pain on Wednesday afternoon for an additional 1+2Δ on Tuesday afternoon, so that he is then facing 6+2Δ on Tuesday afternoon and 4 on Wednesday. And then on Tuesday morning we offer to exchange 1+Δ on Tuesday afternoon for an additional 1+Δ on Wednesday. He will accept both offers for a final consumption schedule of 5+Δ on Tuesday afternoon and 5+Δ on Wednesday, which by his own lights on *any* day is a deterioration.

It therefore looks as though Parfit's character *is* irrational according to my formal definition. Moreover, one might think that pointing this out to the subject will persuade him both that something is wrong with his choice function and also that it would be a good idea to bind (or to pay someone else to bind) his future choices in advance.

I have two points to make in response. (a) What this argument shows to be irrational (if anything) is not any one choice function but a *combination* of two choice functions: one (which he has now) prefers any good on any future non-Tuesday to the same good next Tuesday, and one (which he has next Tuesday) that is indifferent between them. One might perhaps say on that this 'pain pump' shows the *subject* to be irrational, but it does not reveal any irrationality in the choice function that gives no weight to Tuesdays – what Street (2009: 285) calls *Consistent Tuesday Indifference*. Indeed if the subject were to adhere to it consistently – if on all days, *including Tuesdays themselves,* he were indifferent to everything that happened on *any* Tuesday – there would be no possibility of exploitation. (b) Besides, *this* argument could not vindicate Parfit's idea that it is a lack of reasons that makes a choice function irrational, because the source of the exploitability that it reveals is not any lack of

The two approaches described here are both at odds with the spirit of the present proposal. I turn now to two other definitions of rationality to which it owes much more, namely (4.3) rationality as immunity to money pumps and (4.4) rationality as consequentialism.

4.3 Money pumps

The third approach identifies rational preferences with the unavailability of a money pump for those preferences, where this is standardly taken to mean: a sequence of choices that lead to an outcome that is either (a) objectively worse than, or (b) binary dis-preferred to, some available outcome, when those preferences are applied to it.[17] Note that on this way of understanding 'money pumps', the supposedly disastrous outcome need not involve a literal loss of *money*, although it is common to describe it in such terms.[18]

For instance, a money pump that has been alleged to affect the cyclic preference structure $A \succ B \succ C \succ A$ is the Rabinowicz Money Pump (RMP). This works as follows[19]: the agent (call her Alice) starts with A and an arbitrary endowment of money. But this time we *repeatedly* offer her the following trade:

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

Imagine that we display (*) in our shop window, so Alice sees it every morning on her way to work. Alice knows that this morning (Monday), and the next two mornings, she has an opportunity to take up the offer. For instance, she might refuse the offer today, but then take it up on Tuesday and Wednesday. In that case she ends up with B (swapping A for C on Tuesday, and C for B on Wednesday) but is 2¢ worse off. Or she might refuse all three offers. In that case she ends up with her original A and her original wealth. We can represent her choices as follows:

reasons for indifference to any future Tuesdays but rather (what is quite different) the inconsistency between this and his concern for present Tuesdays.

Defining irrationality in terms of money pumps is similar idea to the definition of rationality as means-end rationality (certainly they are closer than either is to Parfit's approach); but see §4.3 for some indication of the differences and for my reservations about the 'money pump' approach.

[17] Objectively worse: Cubitt and Sugden 2001. Binary dis-preferred: McClennen 1988: 89-90.

[18] See e.g. the classic account of Davidson, McKinsey and Suppes 1955: 145-6. There may be a particular problem with using money in the present context, because almost nobody apart from Scrooge treats the accumulation of money as an end in itself: rather, people want more money because it expands the range of options. But it may not be unquestionable that one should *want* a larger range of options. It *is* plausible if one's choice function is maximizing i.e. if one's weak preference relation is transitive and complete, but of course the rationality of transitive preference is among the things that the money-pump argument was supposed to *establish*. Having said all that, one could easily replace monetary losses with some other good (for instance, seconds of pain) the avoidance of which *could* be taken as a plausible end.

[19] Rabinowicz 2000.

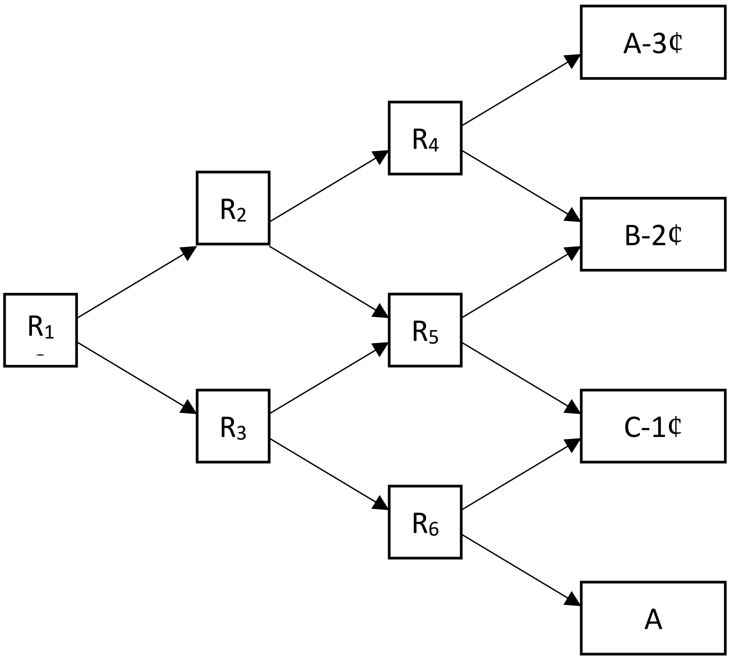

Figure 4: Rabinowicz Money Pump (RMP)

In Figure 4, an upward arrow means ‘accept’ and a downward arrow means ‘reject’. For instance, if Alice accepts (*) on Monday morning and rejects it on Tuesday morning (‘up’ then ‘down’) then just before Wednesday she has the same holding as if she had rejected (*) on Monday morning and accepted it on Tuesday (*), namely $C - 1¢$ (note that ‘$A - 3¢$’ denotes a final outcome in which Alice pays 3¢ for A, and similarly for ‘$B - 2¢$’ and ‘$C - 1¢$’.

What will Alice do? Assume that she only cares about her final holding (i.e. on Wednesday afternoon) and that throughout the procedure the addition or subtraction of a cent makes no difference to her cyclic preferences, so that she has these preferences: $A - 3¢ \succ B - 2¢ \succ C - 1¢ \succ A$. Then a very simple backwards induction argument shows that she accepts all three offers and ends up with $A - 3¢$. I won’t go into the formal details of that, but Figure 5 – in which I have marked in bold the choices that she foreseeably makes if in a position to make them – conveys the basic idea:

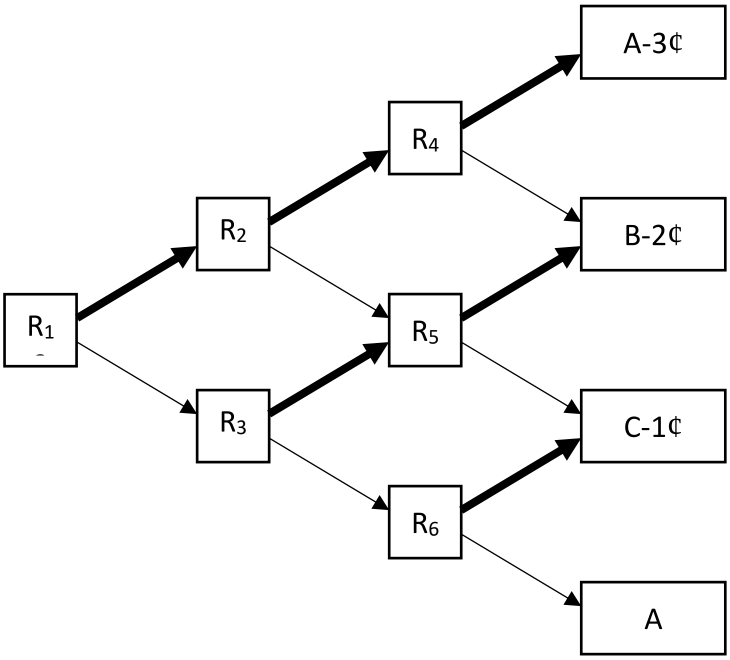

Fig. 5: Alice's choices in RMP

For instance, at R4 she will go up because she prefers $A - 3$ to $B - 2$, and at R5 she will go up because she prefers $B - 2$ to $C - 1$ (from now on I omit the omit the '¢' sign). Foreseeing this, she chooses to go up at R2 because she prefers $A - 3$ to $B - 2$. This and similar reasonings motivate her to go up at every stage, for a final outcome $A - 3$, which she certainly binary dis-prefers to an available alternative A. The fact that this arrangement leads Alice to this outcome is supposed to be grounds on which to find the cyclic pattern $A \succ B \succ C \succ A$ irrational.

There are two main objections to the use of money-pumps as criteria of rationality. The first concerns the normative force of the criterion. Knowing that they will lead her to an outcome that is either (a) in some sense objectively worse than, or (b) strictly dis-preferred to, an available alternative, need not convince an agent that there is anything wrong with her choice dispositions. (a) If her attitude towards the relevant kind of objective goodness is like that of Milton's Satan towards the moral kind of objective goodness, then she will recognize but simply not care that her choices lead to an outcome that is objectively worse.[20] (b) She might even acknowledge that she strictly *binary prefers* another outcome to the one that her preferences realize but insist that it does *not* follow, from the fact that she has a strict *binary* preference for e.g. A over $A - 3$, that $A - 3$ is not choice-worthy from a set of options that includes A, $A - 3$ *and other things*. Why should she, unless she already accepts α?

The right response to these concerns is to amend the definition of what a money pump does. What a pump leads its victim to achieve is not an outcome that she *binary dis-prefers* to *some* available alternative, but rather one that she would not have chosen from the set of *all* available alternatives. Vulnerability to such a pump *does* have a normative grip on its victim. Seeing that her own choice function leads her through the structure of the pump to an outcome that she would not have chosen *ex ante*, she sees that that choice function is in this case frustrating its own ends. And this gives her a reason to change it.

[20] Cf. Railton's (1986: 167-8) discussion of the 'Sensible Knave'. Of course this is no objection to the use of objective notions in Cubitt and Sugden's paper on money pumps, because that paper uses money pumps for predictive not normative ends.

The second criticism is that money pump arguments purport to show the irrationality of some choice function over *outcomes*. For instance, the RMP purports to show the irrationality of any choice function that involves the cyclic preferences $A \succ B \succ C \succ A$. But all that it really shows is the irrationality of a choice function that includes not only these preference over outcomes, but also the preferences over *nodes* that lead to the *ex ante* dis-preferred or unchosen outcomes. In the RMP, reaching the supposedly disastrous outcome $A - 3$ involves not only the cyclic preferences that are the supposed target but also the preferences $R_4 \succ R_5$ and $R_2 \succ R_3$: it is only this total package of preferences that leads to disaster.

What tends to obscure this point is that those other preferences, namely $R_4 \succ R_5$ and $R_2 \succ R_3$, are supposed to be unquestionably rational given $A \succ B \succ C \succ A$, because they emerge from the 'sophisticated' or backwards-inductive reasoning as summarized in Figure 5; and sophisticated reasoning itself is supposed to be normatively compelling.

But sophisticated reasoning *lacks* normative grip. Somebody might make unsophisticated choices over the nodes of a sequential decision problem; and yet it would be impossible to persuade her that there is anything wrong with this by her own lights.

For instance, suppose Alice's preferences are $A \succ B \succ C \succ A$ and that the addition or subtraction of a cent makes no difference to these preferences, but also that her choice function C has these properties:

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

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

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

As we saw, 'sophisticated' choice demands that at R_1 she prefers R_2 to R_3. But she may instead have the opposite preference, on the following grounds: 'Looking at all the outcomes that remain possible, I'd be happy with any of them except $A - 3$. If I go up at this point I am liable to end up with $A - 3$ (because when I'm at R_2 I'll be indifferent between R_4 and R_5, because of (iii)). If I go down at this point I'll certainly avoid it. So I'll go down.'

This 'holistic' consequentialist reasoning leads to different outcomes from 'sophisticated' reasoning in this case. But what makes sophisticated reasoning more compelling? If Alice, at R_1, proposes to reason holistically, what can we say to persuade her out of it? A crudely consequentialist perspective would validate holistic reasoning, because it leads to an outcome that Alice finds acceptable, whereas sophisticated reasoning does not.

I suppose what makes sophisticated reasoning plausible is that if going up leads to an outcome that you *binary prefer* to the outcome that you will get by going down, then it seems that you should go up.

But why not say instead that rather than comparing these outcomes as if they were the only two possibilities, you should instead ask whether either of them is *choice-worthy given the full field of possible outcomes*? For instance at R_2, Alice should ask not whether she *binary 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 instead 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 intuitively hard to distinguish these questions. *If* Alice satisfies both α and β, then any outcome that is choice-worthy from the set of all outcomes available at R_2 will be strictly binary preferred to any that is not. And any such outcome that is strictly binary dis-preferred to some other is *not* choice-worthy from the set of all available outcomes.[21]

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

So given α and β, sophisticated reasoning would seem to rule out the same choices, at any node, as holistic reasoning.

But if, as here, Alice violates at least one of these principles, then sophisticated and holistic reasoning come apart, and it is not clear why the sophisticated methods have any special claim to rationality. Certainly from a purely consequentialist perspective they don't, because as this example shows their use leads Alice to the one outcome that everyone agrees to be sub-optimal i.e. $A-3$. So it is hardly clear that there is anything mandatory about 'sophisticated' choice at non-terminal nodes.

Therefore, it is hardly clear that if anyone *is* vulnerable to a money pump then what this shows to be irrational is their preferences or choice-dispositions over *outcomes* rather than the total package of preferences, including preferences over nodes, that leads to this outcome.

So I propose to modify the standard interpretation of money pumps. A money-pump establishes the irrationality of a choice function *as a whole*, not its restriction to outcomes. We can then define irrationality of choice functions over outcomes in terms of the notion of a perfect cover, as at §3.1. These modifications, in conjunction with the amendment necessary to cover the preceding objection, the result is therefore exactly the definition of rationality defended in this essay. Means-end and outcome rationality could therefore be interpreted as improvements on the 'money pump' approach with which they share an essentially pragmatist spirit.

4.4 Consequentialist rationality

The final approach to rationality is Hammond's notion of *consequentialist rationality*. The basic idea is that a rational consequentialist ought to care *only* about the outcomes of one's choices and not about the route through which one reached those outcomes. Therefore as applied to any tree T, the outcomes that a consequentialist choice function C permits should depend only on the total set of outcomes of the tree and not on its shape. The outcomes that C permits from any tree should be the same as the outcomes that C permits from any other tree with the same outcomes, in particular from the tree consisting of a single straight choice between all of them.

We can state this idea using the present formalism as follows:

> **Consequentialist rationality**: C is a consequentialist choice function over Z if and only if for any $T \in \Delta(Z)$, $C^*(T) = C(T^*)$.[22]

The difference between consequentialist rationality, and the means-end rationality defended here, is therefore very simple. Means-end rationality requires that in application to any tree your choice function permits *only* outcomes that you might have chosen, given a straight choice from amongst the outcomes. Consequentialist rationality is more demanding: it requires that in application to any tree your choice function permits *all and only* outcomes that you might have chosen, given a straight choice from amongst the outcomes.

[22] Here is how Hammond puts it. 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 key 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 also that Hammond introduces further parameters to represent uncertainty about the state of nature and about the outcomes of chance processes, whereas here I am dealing only with the simplest case of 'deterministic' choice.

I do not see why consequentialist rationality is normatively gripping: I don't see why anyone whose choice function is means-end rational but not consequentialist, will care that it fails to be consequentialist.

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

- If $Y \subseteq Z$ then $C(Y) = Y$.
- If Y is a tree of level ≥ 1 then $C(Y) = \{y \in Y | c \notin y^*\}$ if the latter is non-empty;
- Failing that, $C(Y) = \{y \in Y | b \notin y^*\}$ if the latter is non-empty;
- Failing that, $C(Y) = Y$

This choice function doesn't care which of a, b and c is selected in a straight choice between any of them; but given the option to eliminate one of these, it will always eliminate the alphabetically last of the remaining possible outcomes.

Making this a little bit more concrete: suppose that we are selecting candidates for a job, and we have *just one* aim: to appoint a suitable candidate. As it happens there are three suitable candidates. We *could* just appoint one at random; but the HR Department insists that selection proceeds by two stages of elimination. So the following decision tree is the one that HR has imposed upon us:

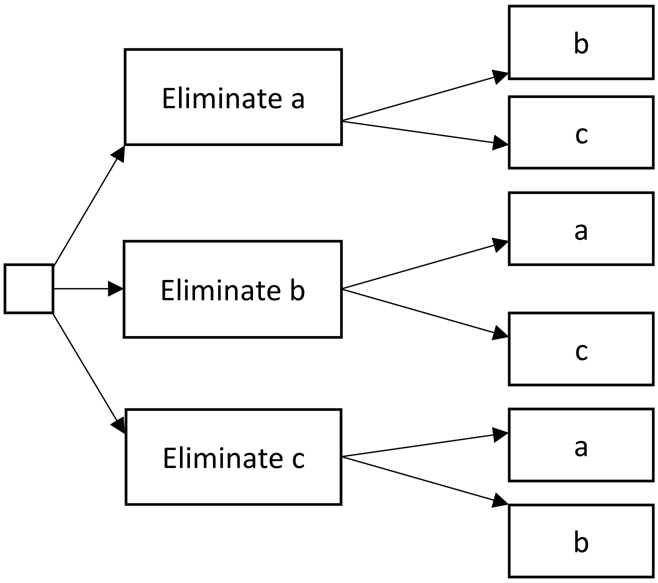

Fig. 6

Suppose that we approach this tree, call it T, in accordance with C: at the first stage we eliminate the candidate that comes last in an alphabetical list, and at the second stage we choose at random. This means that we go down at the first node and then choose either a or b. Since all three candidates were in the running at the outset, we have $T^* = \{a, b, c\}$; since in a straight choice we choose any of these at random, we also have $C(T^*) = \{a, b, c\}$. But since C dictates going *up* at the initial node of T, and allows any choice from a and b, we also have $C^*(T) = \{a, b\}$. So $C^*(T) \neq C(T^*)$. In other words, the choice function C does *not* satisfy consequentialist rationality, because this way of structuring choices makes a difference to the outcome. But it *does* satisfy means-end rationality, because it follows from the definition of C

that $C(T^*) = T^*$ and therefore $C^*(T) \subseteq C(T^*)$ for any $T \in \Delta(Z)$. So if my definition of rationality is an illegitimate weakening of the consequentialist definition then it ought to be possible to talk us out of C.

But on what grounds could anyone do that? You, or the HR representative, might say that the elimination procedure is unfair to c: after all, his being last in the alphabetical list has nothing to do with his suitability for the position. – Well, maybe it is unfair. But *ex hypothesi* all we cared about was appointing the best-qualified person, whether by a fair or by an unfair procedure. And we have done that. Appeals to fairness are not going to show us that we are getting anything wrong by our own lights.

An alternative complaint is that our approach is inconsistent. At the outset, we regard c as an optimal choice from among the three remaining candidates. But it is implicit in our elimination procedure that he is *not* optimal, because that procedure prefers a and b to c. So is c optimal or not? – Answer: c *is* optimal, but why should that imply that our elimination procedure never eliminates him? All that matters is that the procedure eliminates any candidate who is *not* optimal – as long as we do that, it simply doesn't matter that we also eliminate some optimal candidates. Almost every appointment committee on which I have sat has failed to appoint candidates who were perfectly appointable. What matters is that non-optimality is sufficient for elimination, not that it is necessary.

This point leads on to the most basic and obvious reason that the means-end criterion of rationality is preferable to the consequentialist criterion, at least from a normative perspective: the means-end criterion can be regarded as reflecting a more thoroughgoing *consequentialism* than the consequentialist criterion itself. The means-end criterion cares about 'more' than the consequences, in the sense that a means-end rational choice function needn't always return the *same* outcomes when confronted with the same consequences. But it *prioritizes* the consequences: any means-end rational choice function C can be seen as subordinating any other principles of choice by restricting them to operating on a selection from the chosen *consequences* of any tree T, i.e. from a selection from $C(T^*)$. In other words, the difference between means-end rationality and consequentialist rationality is that the former treats consequentialism as a side-constraint upon choice, whereas the latter treats it as the sole determinant of choice.

But consequentialism if it is thoroughgoing *should* regard itself as a side-constraint. For as the example shows, the *consequences* of treating consequentialism as a side-constraint cannot be any less acceptable than the consequences of admitting no other determinant of choice. Someone who ranks (say) welfare policy decisions *solely* by the number of quality-adjusted life years that they save (in expectation), cannot object to a government that always maximizes this quantity, but *in case of a tie* always chooses the policy that most benefits the materially worst-off.

So I believe that by treating consequentialism as more than a side-constraint, 'consequentialist rationality' is an imperfect expression of consequentialism. When we correct for this error, the upshot is the means-end conception of rationality that this essay has tried to defend.

The situation is therefore like that in §4.3. There we saw that the 'money-pump' criterion of rationality as initially presented was a flawed expression of the idea behind it; and when we correct for these flaws, the result is means-end rationality. And it is interesting that both lines of thought – one starting from money pumps, the other starting from consequentialism – converge on the means-end definition of rationality. But it is no part of my case for the means-end definition. My case for it is just this: it is the only definition of rationality on which there is an answer to the question 'Why be rational?' that not only satisfies those who already are but also moves those who are not.

5 Conclusion

The obvious next steps are (a) to extend the definition of rationality offered here to cover choice under uncertainty; and (b) to apply the definition to various puzzles and disputes in rational choice. The appendix makes a start on (a).

As for (b): a thorough treatment of any one of these cases would probably double the length of this essay. But I can at least mention one problem on which means-end rationality promises to shed some light. This is Quinn's famous puzzle of the self-torturer.[23]

The case involves a person who, once a week for the next 100 weeks, has an option to accept an indiscernible increment of pain in exchange for \$10,000. After 100 weeks he is a millionaire but in unendurable agony. Where did he go wrong? We can write the possible final outcomes as Z_i, where $i = 0,1 \ldots 100$ and Z_i indicates an increment of pain of i units and an increment of wealth of $\$10{,}000i$. Then the self-torturer's problem is that his choice function C satisfies these constraints: (i) for each $i < 100$, $Z_{i+1} \succ_C Z_i$, because he is always willing to accept an indiscernible increment of pain in exchange for an additional \$10,000; but also (ii) $Z_{100} \notin C(\{Z_i | i = 0,1 \ldots 100\})$, because he regards Z_{100} as an unacceptable outcome.

According to the standard view, the self-torturer is irrational, because C violates α; but $\overline{C}$ may be outcome-rational in the sense defined at 3.1. That is, it is consistent with (i) and (ii) that for any set of preferences over all sets of Z_i's, such that for any set ζ of Z_i's and any cover K of ζ, some $k \in K$ is such that $\overline{C}(k) \subseteq \overline{C}(\zeta)$. It follows that there is some means-end rational extension of $\overline{C}$ that gives sound advice about how to tackle the decision tree that Quinn describes. Comparing all such extensions with what the self-torturer would also tell us everything there is to know about where he went wrong.

Developing this inquiry is obviously a matter for further work. I also believe that means-end rationality has fruitful applications to other choice-theoretic puzzles, including the problems of supererogation and incommensurability, the dispute between Causal and Evidential Decision Theory and perhaps also Sen's argument concerning the possibility of a Paretian liberal. Independently of these possible applications though, the main advertisement for means-end rationality is that it does what we were looking for: a definition of what it means to choose rationally that makes clear to everyone by their own lights why they should.

Appendix: Rational choice under uncertainty

The definition of rationality defended here is only applicable to choices in deterministic scenarios i.e. where the agent suffers from no relevant ignorance about the state of the world. I don't think that this makes it uninteresting. It's obviously interesting that α and β are not, but γ *is*, a legitimate demand of rationality in such a setting. Still, the obvious next move would be to ask whether introducing uncertainty somehow brings those two other principles into the picture. It would take too much space to do that properly here, but §5.1 at least sketches an extended definition of means-end rationality that covers these cases.

Extending the definition of rationality requires an extension of the space of outcomes and of the set of trees that can be built upon them. To this end, let there be a set Ω of possible worlds namely those that the agent has not ruled out at the outset. And let there be a set Z of prizes. Call any subset E of Ω an *event*. Now we can define terminal nodes, choice nodes, natural nodes and a height function that applies to all of them:

[23] Quinn 1990.

(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 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 is an event 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 can think of the ordered pair (n, E) as carrying two pieces of information: n specifies where the agent is in a tree-like structure, and E expresses her knowledge: that is, E is just the set of worlds that might (for all she knows) be actual. At a terminal node (n, E), n is a function from the set E of possible worlds to the set Z of possible prizes: in other words it is a gamble that returns the prize $n(w) \in Z$ if the actual world is $w \in \Omega$. At a choice node (n, E), *the agent* has a choice between nodes (n', E) at which his state of information is still E. These nodes are the analogues of the non-terminal nodes in a deterministic tree. At a natural node (n, E), *nature* has a choice between nodes (n', E') at which the agent learns that the actual world belongs to some cell E' of some partition of E. These nodes are not analogous to anything in the deterministic case: they are meant to model the evolution of the agent's knowledge over the course of the decision procedure. We define a *decision tree with uncertainty* as a node of finite height of the form (n, Ω).[24]

A *choice function under uncertainty* is any function that takes 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 the state of nature, the outcomes that C permits in T are not themselves prizes but gambles over prizes, that is, functions from Ω to Z. We shall also be concerned with partial gambles, that is, with functions from E to Z for arbitrary $E \subseteq Z$ (which include e.g. all terminal nodes). We can now give a recursive definition of 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 $g_1 \dots g_k$ 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 $g(w) = g_i(w)$.

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

[24] This definition of the relevant class of trees is somewhat simpler than Hammond's classic treatment (1988: 31-2), which includes a class of chance nodes to model processes that are genuinely indeterministic. In Hammond's model therefore, the agent's choice function ranges over *objective* gambles over subjective gambles in the manner of Anscombe and Aumann, whereas on the present approach all gambles are subjective in the manner of Savage. This makes a difference to the details of constructing a utility function given a fully consequentialist rational choice function but not (as far as I can see) to the prior issue of whether consequentialist rationality has normative grip.

Informally, the effect of this definition is that a choice function when 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.

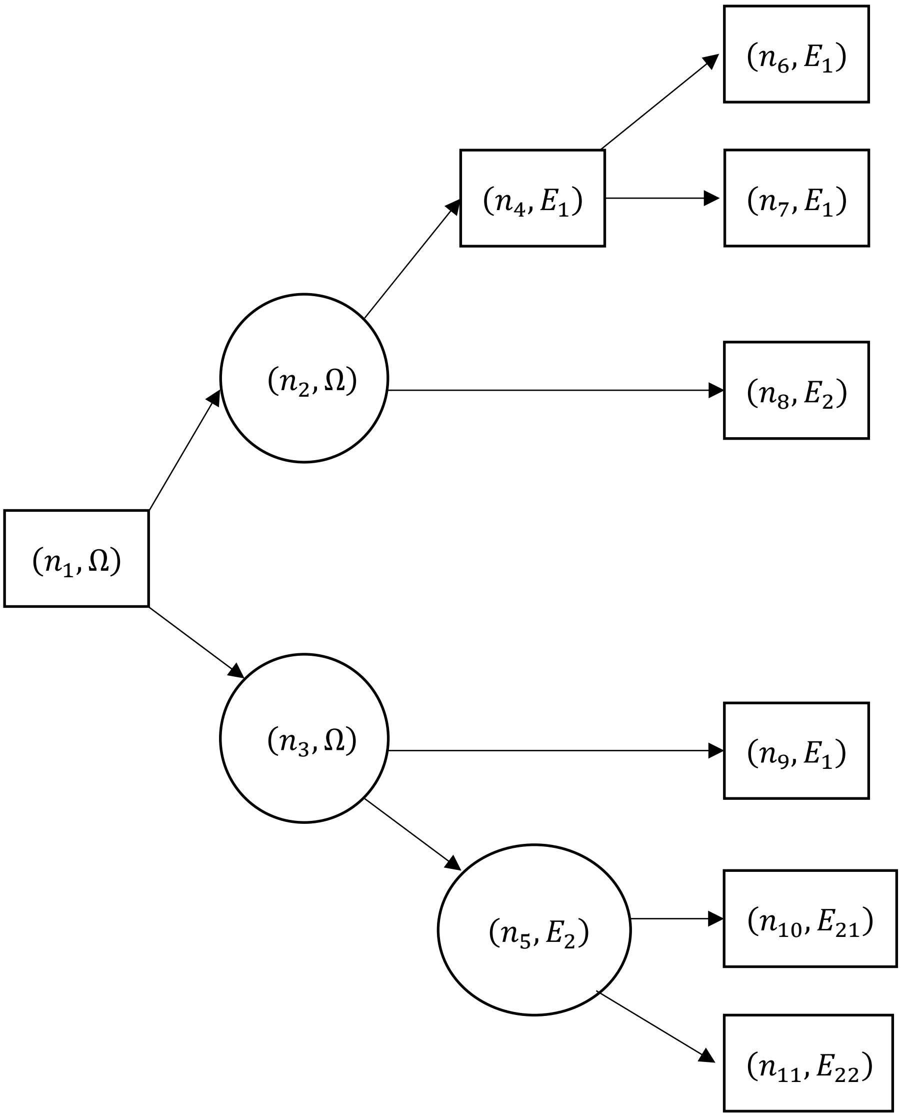

Figure 7

In this diagram, boxes with arrows going out are choice nodes, circles are natural node, 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 have indicated in bold: so $C\big((n_1, \Omega)\big) = \{(n_2, \Omega), (n_3, \Omega)\}$ etc. Then the outcomes that 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 the terminal nodes, since which terminal nodes are available will depend on which possible worlds is actual. For instance, if the actual world does not belong to E_1in 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 $E_1 \ldots E_k$ that partition E and $n_1 \ldots n_k$ 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 available at T is $T^* \subseteq Z^\Omega$.

For instance, if we consider the tree in fig. 7, we can see that in addition to g and h there is available one other gamble, corresponding to the option of going down at the choice node (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 $T^* = (f, g, h)$.

We can now define means-end rationality for choice functions under uncertainty just as we defined it in the deterministic case. A choice function is irrational just in case there are trees from which it is liable to select gambles which it itself would not choose, in advance, from all the available gambles. In other words, the condition for means-end rationality can be written in the same way as before: $C^*(T) \subseteq C(T^*)$ for any tree T. We can then 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, if any, turn out to be means-end rational. For instance, consider the following four gambles, where $E_1 \cup E_2 = \Omega$ and $z_1, \ldots z_4$ are possible prizes.

	$\boldsymbol{E_1}$	$\boldsymbol{E_2}$
$\boldsymbol{f}$	z_1	z_3
$\boldsymbol{g}$	z_1	z_4
$\boldsymbol{h}$	z_2	z_3
$\boldsymbol{k}$	z_2	z_4

Table 2

Suppose that the choice function C satisfies $f \succ_C g$ and $k \succ_C h$. Then consider the following two trees:

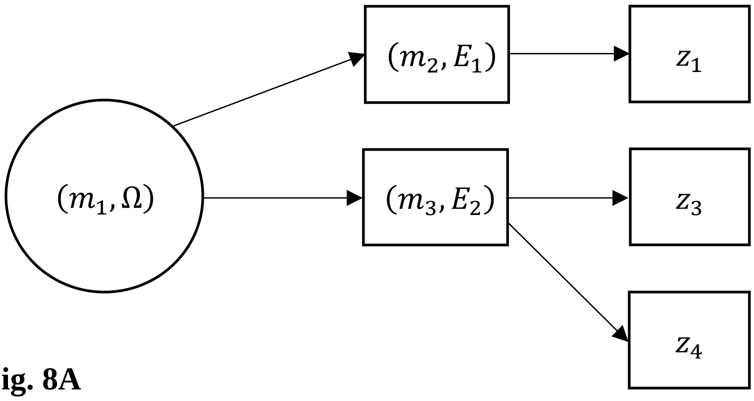

Fig. 8A

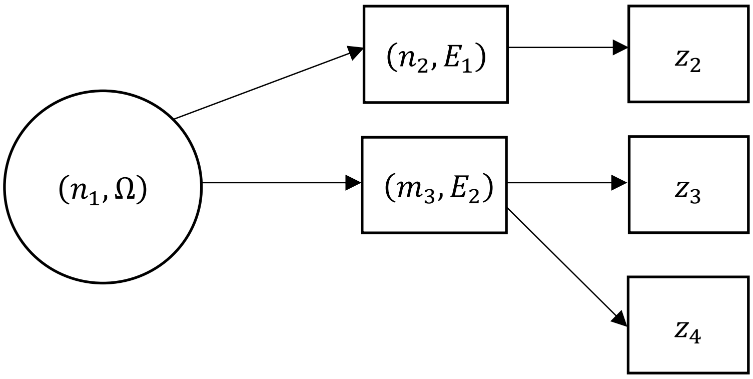

Fig. 8B

Note that the choice node (m_3, E_2) appears in both trees, so that 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 that prize for certain (given what one knows at that point about the state of nature).

Now 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 $C((n_1, \Omega)^*) = \{f\}$. 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 violates means-end irrationality. So it turns out that a principle very like Savage's P2 is a requirement of means-end rationality under conditions of uncertainty.[25]

It is interesting that 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. Clearly there is much more to be said about what else rationality requires when the state of nature is both relevant and uncertain, but this is probably not the place to do it.

[25] Savage 1972: 23. Cf. the proof of Samuelson's 'Independence' principle in the consequentialist theory: Hammond 1988: 42-4.

References

Anscombe, G. E. M. 2000. *Intention*. Harvard: Harvard UP.

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

Cantwell, J. 2003. On the foundations of pragmatic arguments. *J. Phil.* 100: 383-402.

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

Davidson, D., J. C. C. McKinsey and P. Suppes. 1955. Outlines of a formal theory of value, I. *Philosophy of Science* 22: 40-60.

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

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

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

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

Lewis, D. 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, E. F. 1988. *Rationality and Dynamic Choice*. Cambridge: CUP.

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

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

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

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

Railton, P. 1986. Moral Realism. *Philosophical Review* 95: 163-207.

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

Sen, A. K. 1971. Choice functions and revealed preference. *Review of Economic Studies* 38: 307-317.

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

Simon, H. 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.

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

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