1 Introduction

1.1 Normative Decision Theory: What It Is

People make dozens, maybe hundreds of decisions per day. In view of all this practice it is alarming how haphazardly it goes. Most people would drive 15 minutes to save $5 on a $15 jacket but not to save $5 on a $125 calculator.[1] A recent US President based vital decisions on the advice of an astrologer.[2] The Naskapi of Labrador decided where to hunt by the cracks and spots that appeared when they held caribou bones over fire.[3] Rome was supposedly founded on the Palatine Hill because of how many birds Romulus could see from it.[4]

But behind all this seeming arbitrariness lies a vast and ancient fabric of choices that are more (or more obviously) rational. Our survival as a species depended on people knowing how to make fire; that it burns but also cooks; that these berries are edible and those poisonous; that you can fish from this river but that those woods are best avoided, and so on. It follows from the fact that we exist at all that our ancestors acted mostly on experiences that told them these things and not on whether A is a Pisces or B dreamt of seven fat cows.

It took longer to develop a scientific basis for decision-making that applied not only when the relevant facts were known but also when they were uncertain. If you know green berries are poisonous, you shouldn't eat them. What if you know green *or* red berries are poisonous but can't remember which? A systematic approach to such cases had to await two surprisingly late discoveries.

One was probability. Probability applies most simply in some games of chance. It is clear enough what it means to say that the probability of dealing the ace and king of hearts in poker is about 1 in 332. Here we take the probability of an event to control or to arise from the frequency of other events that resemble it in some obvious way. To find the probability of, for example, 'ace and king of hearts' on this deal you look at how often they turn up in other deals.

But even many 'games of chance', like betting on a horse, turn on events that don't naturally fall into a large class of similar events. If in 2011 you wanted to know the probability that Red Cadeaux wins the Melbourne Cup, you would look at – what? How often he won it before? But he never ran it before. How often he wins against Dunaden, who is also racing? But he never ran against Dunaden. How often he wins any race at all? But he ran those other races in widely varying conditions against widely varying opposition, and so on.

[1] Kahneman and Tversky 1984: 347 (so note that these are 1980s dollars). [2] Seaman 2002.
[3] Speck 1935 ch. VI. [4] Livy *Ab Urbe Condita* 1.7.

But there is such a thing as *subjective* probability, or confidence. Your confidence in an event needn't depend on (or settle) how often anything similar happens – you might be ignorant of that. Of course, it *may* so depend; but the point is that we can measure confidence without being dogmatic about that. In a way people have known what confidence is for as long as they have felt it; but we can trace its 'discovery' to Frank Ramsey's famous paper of 1926, which also tells you how to measure it.[5]

The second discovery was utility or subjective value. The idea that some things have 'real' value, independently of what any one person thinks or wants, was central to Plato's philosophy and doubtless part of the interior decor for centuries before. *Price* is a kind of objective value, if not what Plato had in mind. The price of something may depend (in some market conditions) on the totality of people's wants, but it doesn't depend on any one person's wants. Similarly, evolutionary fitness – propensity to reproduce – is an objective value, at least wherever fitness is independent of anyone's opinions or tastes.

But what matters for decision-making is not objective value but what the decision maker wants. Diamonds cost more than water, but you wouldn't care if you were thirsty. Everyone knows that some religions implicitly or explicitly encourage their followers to reproduce, but nobody joins them for that reason. What motivates you and me is *subjective* value: what you want and how much you want it. People have known in a way what subjective value is for as long as they have wanted things. But we can trace *its* 'discovery' to Daniel Bernoulli's famous paper of 1738 on how to measure it.[6]

Normative decision theory arises from the interplay of subjective probability and subjective value. It is like a machine with inputs and outputs. Suppose that you are facing a set of options and you don't know what to do. Then the inputs to normative decision theory are (a) what you think (i.e. subjective probability); (b) what you want (i.e. subjective value). And the *output* is a recommendation from the options.[7]

For instance, suppose I must bet $1 on Dunaden or on Red Cadeaux. If I win on Dunaden, I make 25¢. My subjective value for this outcome is +10. If I win on Red Cadeaux, I make 10¢. My subjective value for that is +8. If I lose on either, I lose $1. My subjective value for that is zero. I'm 25% confident that Dunaden will win and 75% confident that Red Cadeaux will.

My options, the possible results of the race, my confidence in the latter and my values for the resulting outcomes are as in this table.

[5] Ramsey 1926. De Finetti 1937 is an independent treatment on similar lines.
[6] Bernoulli 1738.
[7] For an extended introduction to normative and other forms of decision theory see Peterson 2017.

Table 1.1 Horse race

	RC wins	**D wins**
Bet on RC	+8, 75%	0, 25%
Bet on D	0, 75%	+10, 25%

In each cell there are two numbers: first, the *value* of the corresponding outcome if I take the corresponding option; second, my *confidence* in the outcome if I take the option. For example, the top left-hand cell (+8, 75%) says (a) that I am 75% confident that if I bet on RC, then RC wins; (b) that this is worth +8 to me. Similarly with the other three entries. Now, how should I bet?

A simple approach calculates the *expected value* of each option. For each option this is got by adding the subjective value of each outcome if you choose that option, multiplied by the subjective probability of that outcome if you choose that option. One simple normative decision theory then says: choose any option with the highest expected value. I'll call this theory **MEU** ('Maximize Expected Utility').

Thus in Table 1.1 the expected value of a bet on Red Cadeaux is 6. The expected value of a bet on Dunaden is 2.5. So MEU advises betting $1 on Red Cadeaux.[8]

MEU is one of many normative theories. There is a theory that advises you to choose the (or any) option whose best possible outcome you like most ('maximax'): here, a bet on Dunaden. There is a theory that advises you to choose any option whose *worst* possible outcome you like most ('maximin'): this theory finds both bets acceptable. There are many others.

But something like MEU is appealing. The connection between value and expectation is a consequence of plausible assumptions.[9] And generalizing addition and multiplication in various natural ways reveals a correspondingly generalized idea of expectation within many approaches to decision-making.[10] The theory is simple and gives correct advice where the right decision is obvious.

One version of MEU is the subject of this Element: *Evidential Decision Theory* or EDT. EDT is the normative theory which (according to me) gets choice right: given what you think and want, it gives rational advice about what to do.

[8] Expected value for a bet on RC is 8(75%) + 0(25%) = 6; for a bet on D: 0(75%) + 10(25%) = 2.5.
[9] Milne and Oddie 1991: 54–8. For the merits of MEU-style 'linear pooling' see Pettigrew 2019 ch. 9.
[10] Chu and Halpern 2004.

But what this involves is stranger and more austere than you might expect. Most people think that any general account of how to behave ought to mention the *effects* of your behaviour: what it causes or brings about. But EDT has no special place for that relation. What matters about an option is what it *indicates* – whether by bringing it about or by being symptomatic of it in other ways. As we'll see, this has unsettling practical and philosophical consequences.

1.2 What It Is Not

Before getting into all that, I should distinguish my topic from two others.

First: descriptive decision theory. In one way, a descriptive theory is like a normative theory: it takes beliefs and desires as inputs and gives options as outputs. The difference is in what it is *meant* to do: the normative theory tells you what to do, but the descriptive theory is supposed to predict what you will, in fact, do. Normative MEU theory advises you to maximize expected utility; but descriptive MEU predicts that you will. Given, for example, beliefs and value as in Table 1.1, descriptive MEU predicts that you *will in fact* bet on RC (but it doesn't say that you *should*).

Descriptive decision theory belongs to ethology, whereas normative decision theory belongs to ethics. Facts about actual behaviour might refute some descriptive decision theory but not its normative counterpart. For instance, the well-known 'paradoxes' of Allais and Ellsberg seem to refute *descriptive* MEU. They present situations where subjects apparently make choices that are not maximizing the expectation of anything.[11] But they don't refute *normative* MEU, not if the Allais and Ellsberg subjects are behaving irrationally. And this combination – accepting normative but rejecting descriptive MEU – was a popular reaction to Allais's findings.[12] In any case, EDT itself has descriptive and normative versions. The normative interpretation takes centre stage here. The point is not to describe your behaviour but to guide it.[13]

To explain the second thing I won't discuss, I distinguish *behaviouristic* from *psychological* decision theory.

Behaviouristic decision theory states principles of predicted or recommended behaviour that don't mention anything mental – what you think or want. They just interrelate choices. One such principle is 'transitivity of preference': if you'd choose A over B ('you prefer A to B'), and B over C ('you prefer B to C'), then you'd choose A over C ('you prefer A to C'). 'Preference' here is just behaviour: preferring A to B, on this reading, *means* being disposed to choose A when B is the alternative.

[11] Allais 1953; Ellsberg 1961. [12] Moscati 2019: 190.
[13] EDT has potential as a descriptive theory: see Grafstein 1991, 1999.

Psychological decision theory specifies behaviour as a function of explicitly psychological parameters. MEU is a psychological decision theory. It specifies behaviour as a function of what you think (subjective probability) and what you want (subjective value).

The behaviouristic/psychological and normative/descriptive distinctions create four possibilities for decision theory:

- behaviouristic and normative
- behaviouristic and descriptive
- psychological and normative
- psychological and descriptive.

Psychological normative and psychological descriptive decision theories include the readings of MEU described above. The behaviouristic principle of transitivity might be understood descriptively: people do, in fact, typically behave this way: if a person chooses A over B and B over C, then she will, in fact, also choose A over C. Or it might be understood normatively: if you choose A over B, and B over C, but *not* A over C, then you are choosing irrationally.

Table 1.2 lists principles illustrating all four kinds of theory. For instance, the entries in the top row are behaviouristic: neither specifies what you think or want. But the top-left entry *prescribes* behaviour, whereas the top-right entry *predicts* it.

Given a behaviouristic theory *B* and a psychological theory *P*, there may be a *representation theorem* connecting them. This says that *if* your behaviour conforms to *B*, then we could simulate ('represent') your behaviour by

Table 1.2 Four kinds of decision theory

	Normative	**Descriptive**
Behaviouristic	If you choose apples over pears, and pears over bananas, then you should choose apples over bananas.	If you choose apples over pears, and pears over bananas, then you will choose apples over bananas.
Psychological	If you think apples more nourishing than bananas and only want nourishing food, then you should choose apples over bananas.	If you think apples more nourishing than bananas and only want nourishing food, then you will choose apples over bananas.

programming any of a suitable range *m* of mental states into someone who conformed to *P*. For instance, Savage showed that anyone whose choices satisfied his behaviouristic theory could be represented as having beliefs and desires from a given range and conforming to a version of MEU.

Representation theorems are contributions to philosophy of mind. We can speculate about whether dogs or spiders (or plants or cricket bats) are conscious. But simple behaviour makes it empirically pointless to postulate a complex mentality to things that behave simply. 'We say a dog is afraid his master will beat him; but not: he is afraid his master will beat him tomorrow. Why not?'[14] Because the dog's behaviour is not so complex that any explanation of it would have to distinguish thoughts about tomorrow from thoughts about today.

Representation theorems say precisely what patterns of behaviour *would* give empirical point to attributing this or that belief-desire mentality to the thing doing the behaving. And by specifying what *range* of attributions would explain the behaviour, they also say *how much* mentality it compels us to attribute. This is philosophically interesting from any perspective. Given even moderate behaviouristic sympathies, it takes on special importance: it tells us what it takes to have a mind.[15]

Standard expositions of decision theory typically offer (a) a behaviouristic theory, (b) a psychological theory, and (c) a representation theorem. Jeffrey's classic exposition of Evidential Decision Theory involved (a)–(c). And much of the mathematical and philosophical ingenuity in his and in Bolker's work lay in their discovery of the behaviouristic axioms and the representation theorem.[16] Because of its philosophical importance and interest, Appendix C tries to spell out the intuition behind (c), the representation theorem.

But for the main part I focus on (b). That is, this Element mainly concerns EDT considered as a *normative psychological* thesis. If you tell it what you think and want, it tells you what to do.

1.3 Plan of This Element

Section 1 explains subjective probability – how confident you are that something is true, and subjective or *news* value – how much you want it to be true. Then I introduce Evidential Decision Theory, which recommends maximizing

[14] Wittgenstein 1953: §650.

[15] An analogy from philosophy of language: Quine's argument for inscrutability of reference is a representation theorem connecting linguistic behaviour with the theory assigning references to the speaker's terms (Quine 1981: 19–20). Quine shows that different assignments represent the speaker's behaviour equally well. He infers that there is no fact about which assignment is correct. As we'll see, the Bolker–Jeffrey representation theorem may not get us this far, because it is dubious just how behaviouristic their 'behaviouristic axioms' are. See Appendix C, Section 3.1.

[16] Bolker 1966, 1967; Jeffrey 1983.

news value. Then I sketch the orthodoxy to which EDT is a challenge: Causal Decision Theory.

Sections 2–4 display the content of EDT via its unorthodox recommendations. Each section highlights a different feature of EDT:

- It recommends options that signal good news but do nothing to cause it.
- It recommends options that scramble their own signals of bad news.
- It evaluates future options in the same way as present ones.

In all these cases I find EDT defensible, each section arguing briefly to that effect. But their main aim is less to convince you that EDT is true than to make vivid what it means.

What emerges is not just advice but a vision of decision-making. Your choices don't flow from some part of you that has, or that you for some reason believe to have, a power to intervene in the external world without itself being subject to that world. On the contrary, your decision-making processes are as subject to external influence as your digestive processes. Strange to say, only Evidential Decision Theory fully accommodates this fact.

2 News Value

Evidential Decision Theory says, 'Do what you most want to learn that you will do.'

It can be stated and can sometimes be applied without specifying any measure of *how much* you want to learn something. But the clearest way to explain it is via a measurable quantity called *news value*, the background to which the next two sections explain.

2.1 Possible Worlds and Propositions

To model decision-making under uncertainty, we must represent what you are uncertain about. That means considering possible ways things might be, given what you know when choosing. For example, if you are betting on horses, there is a possible situation where you bet $1 on Kelso and he wins, another where you bet $30 on Trigger and Swaps wins, and so on.

Possible worlds (sometimes 'worlds') are what I'll call the possible ways things might be. Each world settles everything: by choosing which world to realize, God settled all of history. There is then a vast, maybe infinite range of possible worlds, one for each possible history. I'll mostly treat worlds as mathematical points, but nothing important is lost – and some vividness is gained – by imagining them as concrete universes, spatio-temporally isolated

Figure 2.1 The set of possible worlds

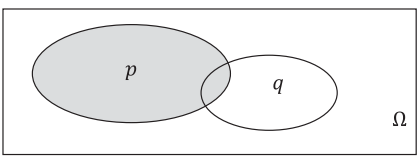

Figure 2.2 The propositions *p* and *q*

from ours, at which these histories really occur.[17] Here I'll write **@** for our possible world (the 'actual world'), and ***Ω*** for the set of all possible worlds.

Figure 2.1 represents Ω and the points and regions within it. The rectangle is the set Ω of worlds (i.e. the set of points inside it). One of these points is @, the actual world.

Every *set* of worlds corresponds to some condition that @ may meet or fail to meet. If *p* is the set of all worlds where it rains in Tokyo on New Year's Day 2020, *p* corresponds to the condition that it rains in Tokyo on New Year's Day 2020. If *q* is the set of worlds where somebody one day runs 100 m in less than 9.5 seconds, then *q* corresponds to the condition that somebody one day runs 100 m in less than 9.5 seconds. Any such set is a **proposition**. Each such set corresponds to some set of points in the rectangle. I'll indicate these sets as shaded regions of Ω: see Figure 2.2.

I'll use concisely stateable conditions to label the corresponding sets. For instance, *p* is the proposition *that it rains in Tokyo on New Year's Day 2020*. A proposition *p* is **true at a world** *w* if *w* meets the corresponding condition: that is, if *w* belongs to the set *p*, written $w \in p$. Otherwise *p* is **false at** *w*. A proposition is **true** or **false** simpliciter if it is true or false at @.

The familiar set-theoretic operations on propositions correspond to logical operations in the obvious way. For propositions *p* and *q*:

- $\boldsymbol{p} \cup \boldsymbol{q}$ ('*p* or *q*') is the proposition that is true at a world *w* if and only if either *p* is true at *w or q* is(or both).

[17] Following Lewis 1986.

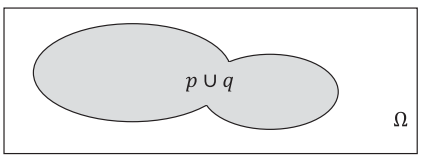

Figure 2.3 The proposition p or q

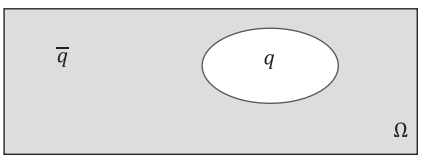

Figure 2.4 The proposition *not* q

- $\boldsymbol{p \cap q}$ (‘ p and q’) is the proposition that is true at w if and only if p is true at w *and* q is.
- $\boldsymbol{\overline{p}} = \Omega - p$ (‘not p’) is the proposition that is true if and only if p is false at w.

In our example, $p \cup q$ is the proposition that *either* it rained in Tokyo on New Year’s Day 2020 *or* somebody one day runs 100 m in less than 9.5 s: see Figure 2.3. And $\overline{q}$ is the proposition that nobody ever did or will run 100 m in less than 9.5 s: see Figure 2.4.

It may happen that p and q are never true at the same world. For instance, let p be the proposition that Tiger Roll wins the Grand National in 2019 and q the proposition that Magic of Light wins it. In that case they are **incompatible**. And $p \cap q$ is the set with *no* elements, the **empty set** $\varnothing$.

A **partition** of Ω, finally, is a set of non-empty propositions $\{q_1, q_2 \ldots q_n\}$ such that every world lies in exactly one element or **cell** of the set.[18] For instance, if in all possible worlds the Grand National was won in 2019 by exactly one of the 40 starters, but it might have been any of them, then there is a partition $\{q_1, q_2 \ldots q_{40}\}$ where each cell corresponds to one of the starters: q_1 is the proposition that Tiger Roll won it, q_2 is the proposition that Magic of Light won it, and so on. See Figure 2.5.

2.2 Subjective Probability

So much for the objects of uncertainty. Now for its measurement. You may be more certain of one thing than another. I am highly confident that next July the

[18] Notwithstanding my notation, a partition may be infinite, though none of the applications here require this.

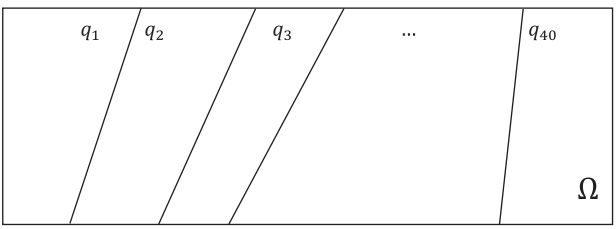

Figure 2.5 Partition of the set of possible worlds

average temperature (in the UK) will exceed what it was last December. I am less confident that it will snow next January; less confident still that it will snow next August. These comparisons suggest a scale for measuring confidence. It runs from zero to one and is called **credence** or **subjective probability**.

Intuitively we can visualize confidence in a proposition as the *area* of the corresponding region. See again Figure 2.2. Suppose you know that exactly one point in Figure 2.2 is the actual world, and you are equally confident, for any region of a given area, that *it* contains the actual world (as if God chooses which world to actualize by randomly sticking a pin in Ω). Then for any given region, your confidence that @ lies in *that* region is proportional to its area. So your confidence that a proposition is *true* (i.e. true at @) corresponds to the associated area.

For instance, your confidence that it rains in Tokyo on New Year's Day 2020 is the area of the shaded region in Figure 2.2. We can express this area as a proportion of the whole rectangle. The closer it is to 1, the greater your confidence that the proposition is true. The closer it is to 0, the greater your confidence that it is false.

More formally, credence is a **probability function**: it assigns to each set of worlds (proposition) p a number $Cr(p)$ between 0 and 1, such that the following rules hold for any propositions p and q:

$$Cr(p) \geq 0 \tag{2.1}$$

$$Cr(\Omega) = 1 \tag{2.2}$$

$$p \cap q = \emptyset \rightarrow Cr(p \cup q) = Cr(p) + Cr(q). \tag{2.3}$$

These rules make intuitive sense when we interpret probability as area. (2.1) says that every region has some non-negative area (maybe zero). (2.2) says that the area is measured in units of the area of the whole rectangle Ω . (2.3) says that