Degrees of Belief — III

HT2017 / Dr Teruji Thomas

Website: users.ox.ac.uk/~mert2060/2017/Degrees-of-belief

1 The Expected Utility Hypothesis (EUH)

1.1 The Basic View

Basic Example: Clara and her umbrella.

Question: What constraints does rationality place on our preferences? E.g. how are Clara's preferences about bringing her umbrella connected to her preferences about getting wet? And how are they connected to her credences, e.g. her credence that it will rain?

General Thought: A good outcome X counts in favour of an action A in proportion to both *how much you value* X and *your credence that* X *would result from* A.

Expected Utility Hypothesis. Insofar as Clara is rational, she

1. (perhaps implicitly) assigns each possible outcome X a 'utility' value $u(X)$
2. prefers actions with higher 'expected utility' [a technical notion].

The 'expected utility' of an action A: consider each possible outcome X. Multiply the utility of X by Clara's credence that X would be the outcome of A. Sum over all outcomes.

Example. For Clara there are four relevant outcomes. 'It rains but she has the umbrella', etc. She might assign utilities like this:

	Rain	No Rain
No Umbrella	0	100
Umbrella	60	60

Prior credence it will rain: $1/2$
Posterior: $1/3$

EUH: Initially Clara should prefer to bring the umbrella (expected utility $60 > 50$); later she should prefer to leave it behind ($\frac{2}{3}100 > 60$). In general, if she is *sufficiently* confident that it will rain, she should prefer to bring the umbrella.

1.2 Distinctions and Subtleties

Subjective vs Objective.

- As far as EUH goes, Clara need not be conscious of her utility function.
- EUH is a condition for preferences to be coherent rather than to be justified. Her utility function is 'subjective' in that sense.
- There might be some further norm: the utility of an outcome should match its objective [prudential or moral] value.

Evaluation vs Choice. 'Preference' as hypothetical choice/disposition to choose vs 'preference' as desire/evaluative attitude – we might be talking about either of these.

Example 1: Buridan's Ass. Buridan's ass might would do well to systematically choose the bale on the left (dispositional), even if he thinks they are equally good (evaluative).

Example 2: Deontology. You might think that in some situation it would be better to lie (evaluative) and yet be disposed not to lie (e.g. because you are a deontologist).

The Newcomb Problem. There are two boxes in front of you. In Box A, there is £1000. In Box B, there is either £0 or £1,000,000. Your options:

onebox. Take box B.
twobox. Take both boxes.

The Catch: The contents of Box B have been decided (a month ago) by an oracle, The Predictor. If the Predictor thought you would onebox, he put in £1,000,000. If the Predictor thought you would twobox, he put nothing. You are certain that the predictor will turn out to be correct.

Argument 1. You are certain that the Predictor will turn out to be correct. So you are certain that if you onebox he will have foreseen it and you will get £1,000,000; if you twobox he will have foreseen it and you will get only £1,000. You had better onebox.

Argument 2. The Predictor made up his mind a month ago. Either he put the money in the box or he didn't; no going back now. And whether he did or not, you get strictly more money – the extra £1,000 – if you also take Box A. So you should twobox.

Evidential vs Causal Decision Theory

Common Ground: The outcome X = 'Clara gets £1,000,000' counts in favour of A = onebox. But the two arguments correspond to different views on how much weight to give it:

Evidential Decision Theory (Argument 1): Cr($X|A$), which equals 1. (For Clara, oneboxing would be good evidence that the Predictor has put the money in the box).

Causal Descision Theory (Argument 2): Cr(X would be the outcome of A). This is just Clara's credence that the Predictor has already put the money in the box.

As in Argument 2b, causal decision theorists may think that the Evidential Decision Theorist gives a correct account of the (or an) 'evaluative' notion of preference. They claim that you ought to twobox, but admit that you might feel some disappointment, or think that twobox is 'worse'.

2 Ins and Outs of EUH

Schematic Representation Theorem: If you preferences meet such-and-such plausible conditions for rationality, then EUH holds (i.e. your preferences can be *represented* by a utility function).

Two most important of the 'plausible conditions':

(1) **Transitivity.** If you prefer X to Y and you prefer Y to Z, then you prefer X to Z.
   * Very plausible if your preferences reflect judgments of 'value'.
   * Very plausible but less compelling if we are just talking about choice.

Mere Addition Paradox. God can:

(A) Create very happy Adam (welfare 100)
(B) Create super happy Adam (110) and sufficiently happy Eve (50)
(C) Create happy Adam (90) and happy Eve (90)

(2) **Sure Thing Principle (STP).**[1] Suppose Clara is choosing between A and B. (E.g. bring umbrella or not.) Suppose also that she doesn't know whether E is true (e.g. it will rain). But if she learnt that E was true, she would prefer A to B, and if she learnt E was false, she would be indifferent between A and B. Then she now prefers A to B.

Example. Clara is thinking of (A) applying for German citizenship or (B) not. She is not sure whether (E) hard Brexit will happen; hard Brexit makes her more inclined to apply. However, she reasons: "If I knew that hard Brexit will happen, I'd prefer to apply; if I knew that it wouldn't, I'd be indifferent. So I'd prefer to apply."

Putative Counterexamples to STP. STP is *very* widely violated in practice.

Allais Paradox, Part I. Clara can buy a ticket in either of two raffles.

Ticket #	Raffle A	Raffle B
1–10:	£25 mil	£5 mil
11:	0	£5 mil
12–100:	£5 mil	£5 mil

Allais Paradox, Part II. Clara can buy a ticket in either of two raffles.

Ticket #	Raffle A	Raffle B
1–10:	£25 mil	£5 mil
11:	0	£5 mil
12–100:	0	0

Fairness. Clara has two children, Bob and Belinda. Grandpa Joe has given Clara £1,000 to pass onto them. For obscure reasons she can only give it to one of them. Clara would be just as happy for Bob to get it as for Belinda to get it. But she would prefer to decide by flipping a coin (even if it costs her £1).

If we understand the outcomes as 'Bob gets £1,000' and 'Belinda gets £1,000', then flipping a coin is just as good as given the money to Bob.

Pascal's Mugging. A mugger approaches you. He has no weapon, but he says, 'Hand over your wallet! In return, I will give you any finite amount of utility that you ask for. I'm able to do this because I have secret powers.'

3 Applications in Ethics

3.1 Subjective Consequentialism

Jackson's Pill Case. Clara is has a mild chronic illness. She has one of two illnessess, but Dr Smith does not know which. Dr Smith has three options to prescribe: (1) Pill A will cure her if she has *first* illness, but kill her otherwise. (2) Pill B will cure her if she has the *second* illness, but kill her otherwise. (3) Pill C will almost but not entirely cure her. Which should he prescribe? [In fact Clara has the first disease.]

'Objective' consequentialism. He ought to prescribe the pill that will have the best results. Problem: he doesn't know which one that is; this ought is not 'action-guiding'.

Railton (?). OK, but he should *aim* at getting the best results. (Nope!)

[1] The exact principle one needs depends on the framework; cognate principles are called 'independence' and 'separability'. STP as I give it here fits best with CDT rather than EDT.

'Subjective' or 'prospective' consequentialism. He ought to prescribe the pill with the best *expected* value. (Depending on the details, this could be Pill C.)

Claim: The 'ought' of subjective consequentialism is not only recognisable but important: e.g. Dr Smith is in no way blameworthy for prescribing pill C, and he would be morally blameworthy for prescribing pill A.

A Challenge for Non-Consequentialist. What the above tell us is that consequentialism works fairly well in cases of uncertainty. Why: EUH tells us that *in some sense* it's rational to choose actions based on the value of their outcomes. It's unclear what non-consequentialists can say about uncertainty. E.g. if *killing is wrong* then what about a 1% chance of killing?

3.2 Moral Uncertainty

Trickier cases are ones in which you are uncertain about the moral facts – 'moral' or 'normative' uncertainty vs 'empirical' uncertainty.

Example. Clara is in a position to kill Davros. She knows that doing so will have great consequences. But she is not sure whether consequentialism is true. She has some credence that consequentialism is true, but also some credence that killing is wrong, full stop. [In fact consequentialism is true.] Is there a sense in which Clara ought not to kill Davros? Is she morally blameworthy if she does kill Davros?

Extreme Case. Suppose Clara is *certain* that murder is morally required. Surely (a) she would be irrational if she did not go around murdering; but (b) she is in no way morally excused.

A Jackson Case. Utilitarianism is true, and everyone knows it. But the Philosopher King is uncertain about the correct theory of wellbeing. What is it that makes a life go well – is it *pleasure* or *autonomy*? [In fact it is autonomy.] Should he set up society so that people have (A) lots of pleasure but no autonomy; (B) lots of autonomy but no pleasure; (C) lots (but not quite as much) of each?

??? The Philosopher King is blameless for choosing (C) even though he knows it is objectively morally wrong?

4 Summing up

1. The orthodox account is the EUH that, in cases of uncertainty, rationality requires that we maximize expected utility.
2. There are some subtleties: are we talking about evaluative attitudes or choice? Are outcomes picked out by evidence or by counterfactuals?
3. The core content of the EUH is arguably the sure thing principle. But apparent violations of STP are widespread, and it's not clear what to make of this.
4. EUH gives a nice story about how consequentialists can handle empirical uncertainty. This is a challenge for non-consequentialists.
5. It's unclear what we should say about cases where the agent is uncertain about moral facts.

Further Reading

Surveys

Peterson, M. (2009) *An Introduction to Decision Theory*. CUP.
[Start here: a textbook introduction.]

Resnik, M. (1987) *Choices: An Introduction to Decision Theory*. U of Minnesota Press.
[An older but still useful textbook.]

Newcomb, EDT, and CDT

Nozick, R. (1969) 'Newcomb's problem and two principles of choice'. In Nicholas Rescher (ed.), Essays in Honor of Carl G. Hempel. Reidel. pp. 114–146 (1969)
[The original philosophical discussion of Newcomb's problem. A huge literature has grown up around it, arguing both sides – you can consult Philpapers.]

Jeffrey, R. C. (1983) 'Deliberation: A Bayesian Framework', chapter 1 in his *The Logic of Decision*, second edition. University of Chicago Press.
[Jeffrey's exposition of evidential decision theory, with plenty of interesting examples/exercises. The second edition discusses Necomb, although he later changed his mind what to say about it – see Joyce below.]

Gibbard, A. and Harper, W (1978). 'Counterfactuals and Two Kinds of Expected Utility.' In A. Hooker, J. J. Leach & E. F. McClennen (eds.), *Foundations and Applications of Decision Theory*. D. Reidel. pp. 125-162.
[An early discussion of CDT compatible with what I've done here. Note there are several different frameworks for developing CDT – probably the most widely cited is 'Savage Decision Theory'.]

Joyce, J. M. (2007) 'Are Newcomb problems really decisions?' *Synthese* 156:537-562.
[Considers Jeffrey's treatment of the Newcomb problem, discussing in particular Jeffrey's late view that the Newcomb case isn't a genuine choice situation.]

Pascal's Mugging

Bostrom, N. (2009) 'Pascal's mugging'. *Analysis* 69 (3): 443–44

The Sure Thing Principle

Broome, J. (1991) *Weighing Goods: Equality, Uncertainty, and Time*. Wiley-Blackwell.
[Chapter 5 argues for EUH, including STP; discusses Diamond's 'fairness' example.]

Buchak, L. (2013) *Risk and Rationality*. Oxford University Press.
[Develops an alternative to EUH. Chapter 5 argues against the Sure Thing Principle.]

Subjective Consequentialism and Moral Uncertainty

Jackson, F. (1991) 'Decision-Theoretic Consequentialism and the Nearest and Dearest Objection.' *Ethics* 101 461–482.
[Explains 'subjective' consequentialism and argues that it helps with some standard objections to consequentialism.]

Weatherson, B. (2014). Running risks morally. *Philosophical Studies*, 167(1), 141–163.
[Argues that there is no 'subjective ought' in cases of moral uncertainty.]

Sepielli, Andrew (2016). Moral uncertainty and fetishistic motivation. *Philosophical Studies* 173 (11):2951–2968.
[Response to Weatherson.]