Lecture 03 : Philosophical Issues in Behavioural Science

Stephen A. Butterfill
< s.butterfill@warwick.ac.uk >

Monday, 16th October 2023

Contents

1. Introduction

In this lecture we consider decision theory, an attempt to provide a mathematical characterisation of rational behaviour.

This lecture depends on you having studied some sections from a previous lecture:

- *Goal-Directed and Habitual Processes* in Lecture 01

For the minimum course of study, consider only these sections:

- *Expected Utility* (section §2)
- *What Are Preferences?* (section §3)
- *Dual Process Theory Opposes Decision Theory?* (section §4)

Alternatively, if you have more time but not enough for everything, skip *Dual Process Theory Opposes Decision Theory?* (section §4) and study the other sections.

There is a bit more than usual to cover this week, which will be hard if this is your first encounter with decision theory.

2. Expected Utility

The bare minimum you need to know about how actions and rationality are represented in decision theory and in game theory for the purposes of this course.

This section is concerned with understanding the way of representing actions and rationality used in almost any variety of decision theory.

This is not very deep. But you need to understand how the representation of actions is supposed to work in order, later, to understand the theory.

This may well already be familiar ground for you. If so, take a quick look at the slides to check you understand the terminology we will use.

I am mostly following Jeffrey (1983) as this is still the introduction that best combines a deep understanding of the topic with philosophical motivations.

2.1. Alternative Text

If you prefer to read a philosopher presenting the core ideas, Bermúdez (2009, chapter 1) is one option. (Bermúdez is summarizing Jeffrey (1983), so read Jeffrey (1983) if you can.)

2.2. Terminology

The choice of terms used in this lecture mostly follows Jeffrey (1983), with a few exceptions where his choices are less familiar.

Make sure you understand the terminology and can relate it to the example choice scenario used as an illustration.

Be sure to use the terminology consistently, and with precision, in your writing.

3. What Are Preferences?

An informal presentation of Jeffrey (1983, chapter 3) on how decision theory enables us to think of subjective probabilities and preferences as simultaneously derivable from patterns of action.

We have relied on notions of belief and desire in considering both philosophical (in ** **ERRoR! MISSING xref FOR** $unit : philosophical_t heories_h abits$ **) and psychological theories (in *Goal-Directed and Habitual Processes* in Lecture 01) of instrumental action and joint action.

But what anchors our understanding, as researchers, of these notions? While some of us might use these words in everyday life, there is probably enough diversity between individuals with different cognitive styles (e.g.~Perner & Leekam 2008), different upbringings (e.g.~Morgan et al. 2014) or different cultural backgrounds (e.g.~Dixson et al. 2018) that whatever understandings you and I have in everyday life may not entirely overlap. And invoking a philosophical theory does not seem likely to help given the level of agreement that has been reached in this regard over the last 2000 or so years.[1]

An attractive alternative is suggested by Jeffrey:

> This book has 'a philosophical end: elucidation of the notions of subjective probability and subjective desirability or utility.' (Jeffrey 1983, xi)

In this section we explore how, following Jeffrey, subjective probabilities and preferences can be identified as constructs of decision theory.

Decision theory therefore promises to be an ideal anchor for a shared understanding of these notions.

Inspired by Jeffery (and Davidson 1990), we might therefore attempt to substitute the informal, poorly understood notions of belief and desire with the

[1] There is a bit more detail on this in some notes for one section of a talk called *The Myth of Mindreading*.

theoretical constructs of subjective probabilty and preference.

3.1. Required Axioms

> 'A binary relation ⪰ on a set A is *complete* if a⪰b or b⪰a for every a ∈ A and b ∈ A,
>
> *reflexive* if a⪰a for every a ∈ A, and
>
> *transitive* if a⪰c whenever a⪰b and b⪰c.
>
> A preference relation is a complete reflexive transitive binary relation' (Osborne & Rubinstein 1994, p. 7).

The *Continuity Axiom* states that if c⪰b⪰a then there is some probability p such that you are indifferent between (i) b happening with certainty and (ii) a happening with probability p and c happening with probability (1-p).

> 'Continuity implies that no outcome A is so bad that you would not be willing to take some gamble that might result in you ending up with that outcome, but might otherwise result in you ending up with an outcome (C) that you find to be a marginal improvement on your status quo (B), provided that the chance of A is small enough.' (Steele & Stefánsson 2020)

The *Independence Axiom* states that if b⪰a then for any probability p, $\{p\text{A},(1-p)\text{C}\} \succeq \{p\text{B},(1-p)\text{C}\}$. Put roughly, if you prefer a to b then you should prefer a and c to b and c.

> 'Intuitively, this means that preferences between lotteries should be governed only by the features of the lotteries that differ; the commonalities between the lotteries should be effectively ignored.' (Steele & Stefánsson 2020)

A preference relation is *independent of irrelevant alternatives* exactly if 'no change in the set of candidates (addition to or subtraction from) [can] change the rankings of the unaffected candidates' (Dixit et al. 2014, p. 600).

4. Dual Process Theory Opposes Decision Theory?

Do any of the findings that support the dual-process theory of instrumental action enable us to construct a good objection to decision theory as an elucidation of subjective probabilities and preferences?

4.1. Background

The dual-process theory of instrumental action was introduced in *Goal-Directed and Habitual Processes* in Lecture 01.

We considered decision theory as an elucidation of subjective probabilities and preferences in *What Are Preferences?* (section §3).

4.2. Argument

The following claims cannot all be true:

1. Decision theory provides an 'elucidation of the notions of subjective probability and subjective desirability or utility' (Jeffrey 1983, xi).
2. The notions elucidated are those of belief and desire, which also feature in goal-directed processs.
3. Some instrumental actions are dominated by habitual processs (see *Goal-Directed and Habitual: Some Evidence* in Lecture 02).
4. Habitual and goal-directed processes can pull in opposing directions (see *The Minor Puzzle about Habitual Processes* in Lecture 02).

What think that these are jointly inconsistent? Because their truth would indicate that we have no grounds to expect agents to act in accordance with the axioms required to make (1) true.[2]

The joint inconsistency of these claims is significant: it suggests that we cannot use decision theory as an anchor for thinking about notions of belief and desire. But perhaps there is a way to avoid this conclusion?

Are the claims actually inconsistent? Or is there some way to use decision theory as an anchor for thinking about notions of belief and desire despite the inconsistency of the above claims?

4.3. Note on Sources

One possibile response to the joint inconsistency of the claims above discussed in the lecture involves a distinction between between computational description and implementation details. This is a rough-and-ready approximation to a famous three-fold distinction from Marr (1982); in terms of that theory my 'implementation details' are what Marr calls representations and algorithms.

[2] On what the axioms are, see *What Are Preferences?* (section §3).

5. An Objection to Decision Theory?

This section introduces the Ellsberg Paradox (Ellsberg 1961) and considers how it might be used as an objection to decision theory.

This is an optional section that was not covered in all versions of the lecture this year.

5.1. The Objection

You can hardly pick up a recent work on decision theory without finding an objection to its axioms.

This section introduces on objection linked to the Ellsberg Paradox (Ellsberg 1961; see Hargreaves-Heap & Varoufakis 2004 for an concise and easy to read presentation if you prefer not to watch the recording).

This is just one of many potential objections. I chose it arbitrarily. It gives me an excuse for sharing a fun fact about Ellsberg himself, which illustrates how research in decision making has had life-or-death consequences.

It would be useful to become familiar with other potential objections if you have time. See, for example, Steele & Stefánsson (2020, §2.3) who present the Allais Paradox; or the various objections in Hargreaves-Heap & Varoufakis (2004, Chapter 1); or almost any recent text on decision theory.[3]

It is perhaps tempting, initially, to think that the objections are simple. They show that decision theory is wrong, misguided or at least too limited to characterise the full richness of human behaviour. But, as we will eventually see, things are much more interesting than that. For it turns out that whether something is an objection depends on what you are using decision theory for.

5.2. How to Object

0. State the construal of decision theory you are considering.
1. State the finding.
    - (In this case, the finding is Ellsberg's discovery of cases where people prefer A over B but also prefer B or C over A or C.[4])

[3] There are some interesting and influential considerations in Sugden (1991), but this is not the place to start so I recommend considering it only if you already have a good understanding of decision theory and comparatively straightforward objections.

[4] You can also mention Jia et al. (2020)'s findings if you are being especially thorough.

2. State the axiom it contradicts.
3. Explain how the finding contradicts the axiom.
4. (If possible, explain why it is significant.)
5. Consider responses.

5.3. Independence Axiom

The *Independence Axiom* states that if b⊠a then for any probability *p*, {*p*A,(1−*p*)C}⊠{*p*B,(1−*p*)C}. Put roughly, if you prefer a to b then you should prefer a and c to b and c.

> 'Intuitively, this means that preferences between lotteries should be governed only by the features of the lotteries that differ; the commonalities between the lotteries should be effectively ignored.' (Steele & Stefánsson 2020)

5.4. The Paradox of Decision Theory

On the one hand, it has become a commonplace that there are plenty of objections to the idea that decision theory characterises how people choose.

On the other hand, there is a growing range of cases in which decision theory (or something based on it, like game theory) has been fruitfully applied. Motor control is a prominent example (see Trommershäuser et al. 2009; Wolpert & Landy 2012).

If the objects are as decisive as usually assumed, why have applications of decision theory proved so fruitful?

Perhaps the answer is that decision theory is a model. Like any model, it can be given different construals. The objections are not objections to decision theory as such, which is simply a model. Instead each objection is an objection to one or more construals of decision theory.

If this is right, it will be important to be clear about which construals your objections concern.

6. Conclusion

After this lecture you should understand what decision theory is, why we need something to anchor a shared understanding among us, as researchers, of the notions of belief and desire, why it is at least theoretically coherent to construe decision theory as providing this, and why construing decision

theory in this way is difficult or impossible to combine with accepting the dual-process theory of instrumental action.

The overall aim of the course: to discover why people act, individually and jointly.

To have any chance of achieving this, we need a synthesis of:

- the kind of theoretical framework provided by philosophical thinking;
- a body of evidence provided by experimental psychology; and
- a formal model.

At this point, we have considered all three items.

This lecture was about the formal models. The best studied, most influential of these is decision theory.

Why do we need decision theory, and how does it fit with the philosophical and psychological theories considered so far?

One possibility is that decision theory provides an elucidation of the notions of belief and desire that we need to characterise goal-directed processs (Jeffrey 1983); see *What Are Preferences?* (section §3).

But, as we saw in *Dual Process Theory Opposes Decision Theory?* (section §4), it is not straightforward to combine this idea with the dual-process theory of instrumental action.

Glossary

anchor A theory, fact or other thing that is used by a group of researchers to ensure that they have a shared understanding of a phenomenon. An anchor is needed when it is unclear whether different researchers are offering incompatible claims about a single phenomenon or compatible claims about distinct phenomena. For example, we might take decision theory to anchor a shared understanding of belief and desire. 3, 5

computational description A computational description of a system or ability specifies what the thing is for and how it achieves this. Marr (1982) distinguishes the computational description of a system from representations and algorithms and its hardware implementation. 5, 10

decision theory I use 'decision theory' for the theory elaborated by Jeffrey (1983). Variants are variously called 'expected utility theory'

(Hargreaves-Heap & Varoufakis 2004), 'revealed preference theory' (Sen 1973) and 'the theory of rational choice' (Sugden 1991). As the differences between variants are not important for our purposes, the term can be used for any of core formal parts of the standard approaches based on Ramsey (1931) and Savage (1972). 2–8

dual-process theory of instrumental action Instrumental action 'is controlled by two dissociable processes: a goal-directed and an habitual process' (Dickinson 2016, p. 177). (See instrumental action.) 4, 5, 8

game theory This term is used for any version of the theory based on the ideas of von Neumann et al. (1953) and presented in any of the standard textbooks including. Hargreaves-Heap & Varoufakis (2004); Osborne & Rubinstein (1994); Tadelis (2013); Rasmusen (2007). 7

goal-directed process A process which involves 'a representation of the causal relationship between the action and outcome and a representation of the current incentive value, or utility, of the outcome' and which influences an action 'in a way that rationalizes the action as instrumental for attaining the goal' (Dickinson 2016, p. 177). 5, 8

habitual process A process underpinning some instrumental actions which obeys *Thorndyke's Law of Effect*: 'The presentation of an effective [=rewarding] outcome following an action [...] reinforces a connection between the stimuli present when the action is performed and the action itself so that subsequent presentations of these stimuli elicit the [...] action as a response' (Dickinson 1994, p.48). (Interesting complication which you can safely ignore: there is probably much more to say about under what conditions the stimulus–action connection is strengthened; e.g. Thrailkill et al. 2018.) 5

instrumental action An action is *instrumental* if it happens in order to bring about an outcome, as when you press a lever in order to obtain food. (In this case, obtaining food is the outcome, lever pressing is the action, and the action is instrumental because it occurs in order to bring it about that you obtain food.) You may encounter variations on this definition of *instrumental* in the literature. For instance, Dickinson (2016, p. 177) characterises instrumental actions differently: in place of the teleological 'in order to bring about an outcome', he stipulates that an instrumental action is one that is 'controlled by the contingency between' the action and an outcome. And de Wit & Dickinson (2009, p. 464) stipulate that 'instrumental actions are *learned*'. 3, 9

model A model is a way some part or aspect of the world could be. 7

representations and algorithms To specify the representations and algorithms involved in a system is to specify how the inputs and outputs are represented and how the transformation from input to output is accomplished. Marr (1982) distinguishes the representations and algorithms from the computational description of a system and its hardware implementation. 5, 8

References

Bermúdez, J. L. (2009). *Decision Theory and Rationality*. Oxford: Oxford University Press.

Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy*, *87*(6), 279–328.

de Wit, S. & Dickinson, A. (2009). Associative theories of goal-directed behaviour: A case for animal–human translational models. *Psychological Research PRPF*, *73*(4), 463–476.

Dickinson, A. (1994). Instrumental conditioning. In N. Mackintosh (Ed.), *Animal Learning and Cognition*. London: Academic Press.

Dickinson, A. (2016). Instrumental conditioning revisited: Updating dual-process theory. In J. B. Trobalon & V. D. Chamizo (Eds.), *Associative learning and cognition*, volume 51 (pp. 177–195). Edicions Universitat Barcelona.

Dixit, A., Skeath, S., & Reiley, D. (2014). *Games of Strategy*. New York: W. W. Norton and Company.

Dixson, H. G. W., Komugabe-Dixson, A. F., Dixson, B. J., & Low, J. (2018). Scaling Theory of Mind in a Small-Scale Society: A Case Study From Vanuatu. *Child Development*, *89*(6), 2157–2175.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669.

Hargreaves-Heap, S. & Varoufakis, Y. (2004). *Game theory: a critical introduction*. London: Routledge.

Jeffrey, R. C. (1983). *The Logic of Decision, second edition*. Chicago: University of Chicago Press.

Jia, R., Furlong, E., Gao, S., Santos, L. R., & Levy, I. (2020). Learning about the Ellsberg Paradox reduces, but does not abolish, ambiguity aversion. *PLOS ONE*, *15*(3), e0228782.

Marr, D. (1982). *Vision : a computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

Morgan, G., Meristo, M., Mann, W., Hjelmquist, E., Surian, L., & Siegal, M. (2014). Mental state language and quality of conversational experience in deaf and hearing children. *Cognitive Development*, *29*, 41–49.

Osborne, M. J. & Rubinstein, A. (1994). *A course in game theory*. MIT press.

Perner, J. & Leekam, S. (2008). The Curious Incident of the Photo that was Accused of Being False: Issues of Domain Specificity in Development, Autism, and Brain Imaging. *Quarterly Journal of Experimental Psychology*, *61*(1), 76–89.

Ramsey, F. (1931). Truth and probability. In R. Braithwaite (Ed.), *The Foundations of Mathematics and Other Logical Essays*. London: Routledge.

Rasmusen, E. (2007). *Games and Information: An Introduction to Game Theory* (4th ed ed.). Malden, MA ; Oxford: Blackwell Pub.

Savage, L. J. (1972). *The Foundations of Statistics* (2nd rev. ed ed.). New York: Dover Publications.

Sen, A. (1973). Behaviour and the Concept of Preference. *Economica*, *40*(159), 241–259.

Steele, K. & Stefánsson, H. O. (2020). Decision Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University.

Sugden, R. (1991). Rational Choice: A Survey of Contributions from Economics and Philosophy. *The Economic Journal*, *101*(407), 751–785.

Tadelis, S. (2013). *Game Theory: An Introduction*. Princeton: Princeton University Press.

Thrailkill, E. A., Trask, S., Vidal, P., Alcalá, J. A., & Bouton, M. E. (2018). Stimulus control of actions and habits: A role for reinforcer predictability and attention in the development of habitual behavior. *Journal of Experimental Psychology: Animal Learning and Cognition*, *44*, 370–384.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2009). Chapter 8 - The Expected Utility of Movement. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics* (pp. 95–111). London: Academic Press.

von Neumann, J., Morgenstern, O., Rubinstein, A., & Kuhn, H. W. (1953). *Theory of Games and Economic Behavior*. Princeton, N.J. ; Woodstock: Princeton University Press.

Wolpert, D. M. & Landy, M. S. (2012). Motor control is decision-making. *Current Opinion in Neurobiology*, *22*(6), 996–1003.