16: Should You Be Instrumentally Rational?

s.butterfill@warwick.ac.uk

To exhibit *instrumental rationality* is to select those actions which you expect to best satisfy your preferences (textbook: Jeffrey 1983).

'the laws of decision theory (or any other theory of rationality) are not empirical generalisations about all agents. What they do is define what is meant … by being rational' (Davidson 1987, p. 43)

'the revealed preference revolution of the 1930s (Samuelson, 1938) … replaced the supposition that people are attempting to optimize any externally given criterion (e.g., some psychologically interpretable motion of utility, perhaps to be quantified in units of pleasure and pain). Rather, if economic agents are typically assumed to be subject to relatively mild consistency conditions (e.g., such as transitivity …), it can be shown that there will exist a set of probabilities and utilities such that each agent's choices will be just "as if" that agent were maximizing expected utility' (Chater 2014).

What being instrumentally rational require? 'As ordinarily understood, the prescription to maximize your expected utility presupposes that there is some measure of expected utility that applies to you and that your preferences are therefore obliged to maximize. But in the context of decision theory, the utility and probability functions that apply to you are constructed out of your preferences, and so your expected utility is not an independent measure that your preferences can be obliged to maximize; rather, your expected utility is whatever your preferences do maximize, if they obey the axioms. Hence, the injunction to maximize your expected utility can at most mean that you should have preferences that can be represented as maximizing some measure (or measures) of expected utility, which will then apply to you by virtue of being maximized by your preferences' (Velleman 2000, p. 149)

1. Motivational States

'The pattern of results accords […] with a role for an incentive learning process in the reinforcer devaluation effect; not only must consumption of the reinforcer be paired with toxicosis, the animals must also have an opportunity to contact the reinforcer after aversion conditioning if there is to be a change in instrumental performance' (Balleine & Dickinson 1991, p. 293)

'The pattern of results accords […] with a role for an incentive learning process in the reinforcer devaluation effect; not only must consumption of the reinforcer be paired with toxicosis, the animals must also have an opportunity to contact the reinforcer after aversion conditioning if there is to be a change in instrumental performance' (Balleine & Dickinson 1991, p. 293)

'we should search in vain among the literature for a consensus about the psychological processes by which primary motivational states, such as hunger and thirst, regulate simple goal-directed [i.e. instrumental] acts' (Dickinson & Balleine 1994, p. 1)

1.1. Complication: Discrepant Actions

'The dissociation between lever pressing and magazine entries produced by re-exposure is […] problematic for the incentive learning account. To recapitulate, this explanation assumes that instrumental performance is mediated by some "representation" of the relationship between the instrumental action and reinforcer that also encodes the current incentive value of the reinforcer. The represented incentive value can only be changed, however, after aversion conditioning by exposure to the reinforcer. Given this account, the question immediately arises as to why re-exposure is necessary for a change in lever pressing but not magazine entries' (Balleine & Dickinson 1991, p. 293)

'A possible resolution to this discrepancy lies with the differing contingencies controlling lever pressing and magazine entry. There is evidence that simple anticipatory approach to a food source, such as magazine entry, is primarily under the control of Pavlovian as opposed to instrumental contingencies (e.g. Hol-

land, 1979),thus raising the possibility that incentive learning is necessary for instrumental but not Pavlovian reinforcer revaluation effects. There is, in fact, independent evidence that accords with this analysis' (Balleine & Dickinson 1991, p. 294)

2. Dilemma

Horn 1. Prioritise one kind of motivational state over all others. Define instrumental rationality in terms of optimally satisfying motivational states of this kind.

Horn 2. Assume that despite multiple kinds of motivational state at the level of representations and algorithms, the system as a whole will satisfy the axioms governing preferences (e.g. transitivity).

3. Appendix: Consequences for Mindreading?

If we have multiple, somewhat independent systems of motivational states, how can we justify using decision theory to charaterise behaviour?

'once we accept that there are complex and subtle non-intentional processes, such as those mediating basic goal-approach and the adjustment to changes in motivational state, that can mimic true intentional control in many situations, we can understand why the propensity to perceive actions as intentional may have developed. Given that either there is nothing in the stimulus input per se to distinguish intentional from non-intentional behaviour or that such a discrimination yields little of consequence in most situations, it may well pay the perceiver to treat both classes of behaviour as intentional in predicting the subsequent course of events' (Heyes & Dickinson 1990, p. 102).

References

Balleine, B. & Dickinson, A. (1991). Instrumental performance following reinforcer devaluation depends upon incentive learning. *The Quarterly Journal of Experimental Psychology Section B*, *43*(3), 279–296.

Chater, N. (2014). Cognitive Science as an Interface Between Rational and Mechanistic Explanation. *Topics in Cognitive Science*, *6*(2), 331–337.

Davidson, D. (1987). Problems in the explanation of action. In P. Pettit, R. Sylvan, & J. Norman (Eds.), *Metaphysics and Morality: Essays in Honour of J. J. C. Smart* (pp. 35–49). Oxford: Blackwell.

Dickinson, A. & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18.

Heyes, C. & Dickinson, A. (1990). The Intentionality of Animal Action. *Mind & Language*, *5*(1), 87–103.

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

Velleman, D. (2000). *The Possibility of Practical Reason.* Oxford: Oxford University Press.