Instrumental Rationality

Ralph Wedgwood
ralph.wedgwood@merton.ox.ac.uk

0. Is there any distinctive requirement of rationality (or aspect of rationality) that deserves the label "*instrumental rationality*"? If so, what is it exactly? (The goal here is just to give a correct *specification* of this requirement of rationality – not to seek its *ultimate explanation*.)

1. At first glance, instrumental reasoning seems typically to involve the following steps:

- **(a)** You intend to achieve a *end*.
- **(b)** You form beliefs about what means are *available* for achieving that end.
- **(c)** You forms beliefs about which of these means are *better*, and which are *worse*, than the others.
- **(d)** Finally you *choose* one of these available means that you believe to be *optimal* (i.e. no worse than any other) – or at least, means that you do not believe to be worse than any other.

What is it to believe that a certain course of action is "available"?

Proposal: believing that a certain course of action is available in this way has two components:

- **i.** One must regard it as at least *epistemically possible* that one will intend that course of action
- **ii.** One must also have a confident conditional belief that one will in fact take that course of action if one intends to do so.

What is it to believe that one of these possible means for accomplishing the end is "*better*" than another?

It need not be just to believe that the first means is more effective at achieving the end. (The most effective way of accomplishing a goal will frequently be too costly or disagreeable or boring or painful … to count as the best, or as even one of the best, ways of accomplishing the goal.)

Proposal: the notion of what is "better" or "best" that is being employed here is simply the general notion of *choiceworthiness.*

To take a certain course of action *A* as a "means" to a certain end *E* is to carry out an *intention* to take that course of action *A in order to* achieve that end *E*. When you intend a course of action *A* "in order to" achieve an end *E*, your intention to take *A* is in a way *subordinate* to your intention to achieve *E*. Roughly, your intention to achieve end *E* "controls" or "guides" the way in which you take the course of action *A*.

Instrumental *rationality* consists, presumably, in doing such instrumental reasoning *rationally*. So, roughly, someone is instrumentally rational when they respond to the information that they have with (i) rational beliefs about what means of achieving that end are *available*, and (ii) rational beliefs about which of these means are *better* and which are *worse*, and (iii) intentions that are *in line with* these rational beliefs.

Roughly, for your intentions to "be line" with your beliefs, you must never simultaneously (i) intend an end, (ii) believe, of a certain set of alternative means, that each of them is an available and optimal means to the end,and yet (iii) intend none of these means – at least so long as (iv) you also believe that you will not achieve this end in an optimal way unless you now decide on one of these means.

PLEASE TURN OVER …

This description of instrumental rationality is still rough and imprecise in several respects. E.g.:

- **(a)** Some courses of action are *parts* or *components* of larger courses of action. How does this description accommodate this point?
- **(b)** What about occasions when we are *uncertain* about crucial features of our situation?

2a. Joseph Raz (2005) focuses on the "facilitative principle": If one has a sufficient reason to pursue an end, one also has a reason to take any course of action that facilitates that end.

However, our intuition is not just that it is rationally *permissible* to intend the means to our ends, but that it is rationally *impermissible* to intend an end, while believing that certain means are the optimal means to that end, without ever intending any of those means (at least so long as you also believe that you will not achieve the end in an optimal way unless you now decide on one of these means).

Even if one believes that one has "a reason" (or even a "sufficient reason") to take a course of action, there need be nothing rationally impermissible about not intending that course of action. So it seems impossible for Raz's principle to capture the fact that this bad combination of attitudes is rationally forbidden or impermissible in this way.

(The same point also shows that there is little prospect of the problem's being solved by the neo-Humean idea that our *desires* generate reasons for action.)

2b. In their discussions of instrumental rationality, John Broome (1999) and Kieran Setiya (2007) both focus only on the very special case of *necessary* means – that is, means that are such that (you believe that) you *will not* achieve the end without taking those means.

But in fact only a tiny part of our instrumental reasoning is concerned with reasoning our way from an intention to achieve an end to an intention to take the *necessary* means to the end.

Moreover, the requirement to take what one believes to be necessary means to one's end can be derived from the requirement to intend some member of the set of means each of which one believes to be optimal. So the latter requirement seems more fundamental than the former.

Still, Setiya (2007a) is right that it is rational to intend a course of action only if it is simultaneously rational to *believe* that one will *successfully* carry out the intention.

Indeed, this point follows from the fact that an intention can be rational only if it rational to believe the intended course of action to be *available* (given my interpretation of "availability").

In fact, however, both sensitivity to "reasons" and "value" (which Raz focuses on) and judgments of "availability" (which underlies what Broome and Setiya focus on) are features of *all* practical reasoning, not just instrumental reasoning.

This suggests that if there is anything distinctive of *instrumental* rationality, we will have to look elsewhere to find it.

3. What about *causal decision theory* (CDT) – the sort of decision theory that has been developed by Allan Gibbard and William Harper (1978), and by David Lewis (1981)?

In fact, there are idealizing assumptions built into CDT, which guarantee that CDT has nothing to say about instrumental rationality.

For various reasons, the "acts" that CDT focuses on must be *extraordinarily specific* acts – acts that include everything that is within the agent's control that is relevant to determining how good or valuable these acts are.

In the overwhelming majority of cases, the factors that are relevant to determining the value of an act include *both* the end that is achieved *and* the means that are used in order to achieve that end.

In this way, then, CDT cannot model the kind of decision process that proceeds piecemeal, by first deciding on what end to pursue, and then deciding on what means to use to achieve the end. Instead, it can at best serve only as a way of identifying the *complete packages of means and ends* that could be the *total upshot* of an ideally rational process of decision-making.

(James M. Joyce (1999) is aware of this limitation of CDT, but his solution is not in my view responsive to the problem.)

4. We saw in Sections 1–2 that rational practical reasoning typically involves some kind of estimate of (i) the *availability*, and (ii) the *value* or *desirability*, of various options.

We have now seen, in Section 3, that it also requires (iii) some kind of *integration* of one's various piecemeal decisions, so that they collectively lead to broadly the same course of action as a rational "grand-world" decision.

Presumably, this sort of "integration" will have in some way to concern both (i) the availability and (ii) the desirability of the overall upshot of this series of piecemeal decisions.

My proposal was: Believing that an option is "available" is having a high conditional credence – in the circumstances amounting for all practical purposes to conditional certainty – that if one intends the option, one will execute one's intention and act accordingly; and for an intention to be rational, it must also be rational for the agent to believe the intended option to be available in this way.

In addition, I propose the following constraint on *sets* of intentions: If one is rational, one's intentions should *not* be such that one has a high conditional probability that if one has precisely those intentions, one will *not* carry out all of one's intentions.

With respect to the dimension of value or desirability, I propose that a rational agent will have a set of intentions that in some way *maximizes expected value*.

However, is it (i) the set of *intentions* or (ii) the set of intended *courses of action* that must maximize expected value?

- **i.** It might seem that one's intentions must maximize expected value: even if something is a good thing to *do*, it may not be a good thing to *intend* to do (having the intention may not help).

- **ii.** It might seem that if our *intentions* have to maximize expected value, then the rational agent must intend to drink the toxin in Gregory Kavka's (1983) "toxin puzzle" case.

I propose that to assess a given set of intentions, the relevant probabilities are *conditional* probabilities – the probabilities of the various relevant propositions conditional on one's having that set of intentions. However, the propositions whose probability is in question are propositions about the value of the *course of action* that one would take if one were to execute those intentions.

This solves both (i) the problem of things that are good things to do but not good things to intend, and (ii) the toxin puzzle.

CDT may be correct about what determines the degree to which a course of action approximates to being (as we might put it) what one *objectively ought* to do: this may indeed be determined by a comparison of the most specific and detailed available acts, in all the practically available possible worlds.

But even if CDT is right about what one objectively ought to do, it seems wrong about *rationality* (or at least about the rationality of the sort of reasoning of which we are capable).

Agents can entertain propositions of the form, 'Out of all the currently available courses of action, the course of action that I will actually take will be good or choiceworthy to degree *d*' ("value-specifying propositions"). I propose that we should use propositions of this sort to define a notion of the "expected value" of a set of intentions, together with the probabilities of the various relevant propositions conditional on one's having this set of intentions.

Very roughly, a rational agent must have a set of intentions that collectively makes it rational for the agent to believe, not just that she will carry out those intentions, but also that in so doing, she will be doing something that is a suitably good thing to do.

This proposal can capture the intuitive features of instrumental reasoning that I identified earlier, while also clarifying the doubtful points that I mentioned.

5. This then is what instrumental rationality is. It is the rationality of the process of i*ntegrating* the various different intentions that one forms, in the course of piecemeal "small-worlds" practical reasoning, into a coherent overall set of intentions. For this process to be rational, it must be sensitive in the appropriate way to evidence of the *availability* and *desirability* of the course of action that will result from carrying out one's intentions.

References

Broome, John (1999). "Normative requirements", *Ratio* 12 (4): 348-419. DOI: 10.1111/1467-9329.00101

Gibbard, Allan, and Harper, William (1978). "Counterfactuals and Two Kinds of Expected Utility", in C. A. Hooker et al., eds., *Foundations and Applications of Decision Theory* (Dordrecht: Reidel), vol. 1: 125–62.

Joyce, J. M. (1999). *Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press).

Kavka, Gregory (1983). "The Toxin Puzzle", Analysis 43, 33–36.

Lewis, David (1981) "Causal Decision Theory", *Australasian Journal of Philosophy* 59: 5–30.

Raz, Joseph (2005). "The myth of instrumental of rationality", *Journal of Ethics and Social Philosophy* 1 (1): 2-28. http://www.jesp.org/

Setiya, Kieran (2007). "Cognitivism about instrumental reason", *Ethics* 117 (4): 649-73.