THE ROLE OF THE INSTRUMENTAL PRINCIPLE IN ECONOMIC EXPLANATIONS

A thesis submitted in partial fulfilment of the requirements for the degree of

MASTER OF ARTS

of

RHODES UNIVERSITY

by

Nimi Hoffmann

February 2009

ABSTRACT

Economic explanations tend to view individuals as acting to satisfy their preferences, so that when given a choice between goods, individuals choose those goods which have greater utility for them – they choose those goods which they believe can best satisfy their preferences in the circumstances at hand. In this thesis, I investigate how utility theory works when it is used to explain behaviour.

In theory, utility is a positive concept. It is intended to describe and explain an individual's behaviour without judging or justifying it. It also seems to be regarded as non-hypothetical, for it explains an individual's behaviour in terms of preferences which need not be shared by others, but may be wholly particular to her. This implies a distinctive way of approaching people's behaviour as isolated from and immune to the judgements of a community, for utility cannot be used as a common standard by which we judge an individual's behaviour as better or worse, appropriate or inappropriate.

I argue that this theoretical treatment of utility is substantially different from the practice of using utility to explain behaviour. In the first place, when utility is used to explain behaviour as preference-guided, it treats this behaviour as rational action. An explanation of rational action is, however, necessarily governed by the instrumental principle. This principle is normative – it stipulates the correct relation between a person's means and her ends, rather than simply describing an existing relation. The principle is also non-hypothetical – our commitment to the principle does not rely on the possession of particular ends, but on having ends in general.

The instrumental principle therefore acts as a common standard for reasoning about how to act, so that when we explain an agent's behaviour as rational action, we expect that her action will conform to standards that we all share in virtue of having ends. Thus, I contend, in order to explain the rational actions of an individual, marginal utility necessarily appeals to the judgements of a community.

CONTENTS

ACKNOWLEGEMENTS

Many thanks to my supervisor, Marius Vermaak, in the Department of Philosophy at Rhodes University. His insightful and patient guidance has sharpened my thinking, and the generous way in which he has given of his time in the midst of many duties is greatly appreciated. Special thanks to my co-supervisor, David Fryer, in the Department of Economics at Rhodes University. His scepticism about orthodox approaches has never prevented him from guiding me towards a more charitable and accurate reading of neoclassical economics. I am grateful for the support from members of the Departments of Philosophy and Economics, and in particular, Nhlanhla Mbatha and Veli Mitova, who consistently challenged my thinking and inspired an interest in the intersection between practical reason and neoclassical economics. I am indebted to the National Research Foundation, the Andrew Mellon Foundation, and the Ernest and Ethel Eriksen Trust, whose funding has allowed me to pursue this research.

I am finally able to repay a debt of gratitude to Michiel Hogerzeil, who gave with a careless, gentle heart. Jutta, Simone and Mandu – your irrepressible humour has been my solace and my guide. Thank you.

INTRODUCTION

"The present intellectual scene", writes Ben Fine, "is one in which economics imperialism is on the rampage across the other social science".[1] The story of colonisation in the social sciences is, for Fine, much like the story of colonisation in the developing world. First, economists with their greater political and financial clout begin to dictate the terms of success for other social scientists. In this way, sociologists, cultural theorists and political scientists turn to the language of economics, trying to convey the economic relevance of cultural norms, institutions and networks by dressing them up as 'human capital'. Then economists, in their recognition that the new world of social sciences might contain hitherto unimagined resources, broaden economic analyses to include non-market behaviour as the rational response to market imperfections. When the market fails, it is rational to form social structures and engage in what would otherwise seem to be non-rational behaviour like customs, norms and relations of trust. In this way, denizens of other social sciences are assimilated into the language and practices of economics.

I am not sure how much truth this picture represents. Nevertheless, I understand the attraction that economics has for those situated in, and affected by substantial socio-economic inequalities. While philosophy seems to be uniquely fitted to answer the question of what the good life consists in, it would appear that economics is tailored to answer the more practical question of how to go about securing a good life for oneself and others. This is particularly the case for utility theory in economics, which is essentially used to measure the way in which we go about securing the means to our ends in a world of scarcity. It is within this context that this thesis investigates the question of how utility theory works. At times, the discussion may seem far removed from the pressing considerations of scarcity and inequality, but it is useful to bear in mind that these considerations shape the nature and direction of the investigation.

[1] Ben Fine, 'Economics Imperialism and the New Development Economics as Kuhnian Paradigm Shift?', *World*

An enquiry into how utility theory works immediately encounters the heterogeneity of utility theory. The concept of utility is not a single, unified object. Instead, it is an intricate complex of instruments, each designed to solve particular, historically-situated problems. When classical utilitarians sought a way of thinking about the good of a society such that each person's happiness or pleasure is crucial to the overall good, they fastened on the idea of *classical utility*. It is meant to provide a universal criterion for judging the moral worth of an action in terms of the overall happiness or pleasure that it results in. As such, classical utility is part of what Amartya Sen calls the 'welfarist' project – it is used to assess an action's goodness in terms of the aggregated positive effects on individuals.[2]

Set firmly on this welfarist course, neoclassical economists nevertheless wished to distance themselves from the utilitarian's moral project. In place of classical utility, they developed the concept of *marginal utility*, an instrument for measuring and explaining – but not justifying – a person's behaviour in terms of her preferences. One can think of individuals as acting to satisfy their preferences, so that when given a choice between goods, individuals choose those goods which have greater marginal utility for them – they choose those goods which can best satisfy their preferences in the circumstances at hand. Unlike classical utility, marginal utility is treated as a positive concept – it is intended to describe and explain an agent's behaviour in terms of preferences which need not have any moral content. So, by connecting preferences to choices, marginal utility can act as the foundation for an account of rational choice shorn of its moral overtones. More than this though, it is also hypothetical, in the sense that marginal utility does not depend on a universally-shared end, like pleasure or happiness, but instead depends on preferences which can be wholly particular to an individual. This implies, I think, a distinctive way of approaching agents' actions as isolated from and immune to the judgements of a community, for marginal utility cannot be used as a common standard by which we can judge an individual's actions as better or worse. It simply describes and explains those actions in terms of preferences. Marginal utility theory does

---

Development 30, 12(2002): 2057.

not, however, have much to say on the question of how one should go about identifying these preferences.

This is the particular task of *axiomatic utility*. Rather than give itself the task of investigating mental states hidden within an agent's head, contemporary economics has turned to a behaviouristic understanding of preferences as largely interchangeable with choice behaviour. Within axiomatic utility theory, choices reveal preferences and preferences just are hypothetical choices. By closing the gap between preferences and choices, axiomatic utility theory is able to ascribe or discover an agent's preferences by examining her choices.

Although each concept of utility is concerned with a distinctive problem, one can think of the instrumental principle as the thread connecting all three concepts. This is a principle which governs the relation between means and ends, or between goods and preferences (I use both sets of terminology interchangeably). Classical utility gives moral weight to this instrumental relation by defining the goodness of an action in terms of the pleasure or happiness that it produces. In contrast, marginal utility only accords this relation explanatory weight. Axiomatic utility is minimalist in the extreme – it attempts to give the instrumental relation no weight at all, for it equates preferences with choices so that means are no longer instrumental to ends and are instead the ends themselves.

The main claim of the thesis is this: in order to explain an agent's behaviour as an instance of rational action or choice, an explainer must assume that her behaviour is governed by the instrumental principle. I take this claim to be significant in several ways.

In the first place, the instrumental principle is normative: it stipulates the correct relation between an agent's means and her ends, rather than simply describing an existing relation. The principle is also non-hypothetical: our commitment to the principle does not rely on the possession of particular ends, but on having ends in general. It therefore acts as a common standard for reasoning about how to act, so that when we explain an agent's behaviour as rational action, we express an expectation that her behaviour will conform to standards that we all share in virtue of

[2] Amartya Sen, 'Utilitarianism and Welfarism', *The Journal of Philosophy* 76, 9(1979): 464.

having ends. An explanation that draws on the instrumental principle therefore gets its sense from the community of agents in which the explainer is embedded. So when marginal utility is used to explain an agent's behaviour as rational action, it implicitly invokes communal judgements about the nature of correct reasoning. In order to explain the rational actions of an individual, the behaviour of an 'I', marginal utility appeals to the judgements of a community, the judgements of 'us'.

This kind of explanation will therefore have normative and non-hypothetical properties. Yet marginal utility is characteristically set out in positive and hypothetical terms: it is supposed to describe and explain an agent's behaviour in terms of preferences which are not necessarily universal. As such, the explanatory version of the concept differs substantially from its abstract version. For any empiricist, the formulation of theory is primarily guided by successful practice, so this conclusion should persuade economists with empirical leanings to favour the explanatory version of utility over its abstract counterpart.

Nevertheless, regardless of whether the theory of marginal utility is reconciled with its practice, I think reflection on the normative and non-hypothetical properties of utility explanations brings neoclassical practice closer to one heterodox theory – the capabilities approach. As I understand it, the role of public deliberation and judgement is central to the capabilities approach, and if it turns out that utility explanations likewise rely on an element of communal judgement, then there may be more common ground between a neoclassical economist and a capabilities theorist than one might have originally supposed. In this case, we can think of the two theories as being on a continuum with each other: marginal utility makes implicit reference to public judgements, while the capabilities approach renders the role of public judgements explicit and develops it in interesting directions.

The structure of the argument is as follows. In the first chapter, I set out the theoretical treatment of marginal utility and distinguish it from classical utility. On this reading, marginal utility is positive and non-hypothetical. In the second chapter, I look at the way in which marginal

utility can be used in practice to explain agents' behaviour. I argue that if marginal utility is used to explain an agent's behaviour as an instance of rational action, then the explanation must invoke the instrumental principle, which is normative and non-hypothetical. In the third chapter, I consider whether this argument can be applied to the concept of axiomatic utility. I argue that it can, because it implicitly invokes the distinction between means and ends which marginal utility makes explicit. Moreover, this distinction between means and ends is of a special kind, for it requires an explainer to view an agent's behaviour as an instance of rational action. The instrumental principle is therefore a necessary feature of an explanation which employs marginal utility.

In the fourth, concluding chapter I place the analysis of marginal utility in a broader context, by considering whether the capabilities approach is as different from utility theory as we have been lead to think. I contend that it is not, for both utility theory and the capabilities approach contain non-hypothetical and normative properties. Finally, the coda offers a brief discussion of two issues which I purposefully bracketed from the main argument. The first potential avenue of exploration is the concept of expected utility, which looks as if it might push the analysis towards reasons internalism. The second avenue concerns the role of behavioural economics in identifying two norms distinct from the instrumental principle: a norm for prudence and a norm for valuation. I suggest that these two norms may have a bearing on our understanding of how utility explanations work.

CHAPTER 1

Theorising about our Behaviour: The Concept of Marginal Utility

The everyday meaning of 'utility' is *usefulness*. When we say, for instance, 'this is a useful hairdryer', what we mean is that it is useful for a certain purpose that we have in mind. We cannot apply the statement 'this is a useful hairdryer' to just any context; it is not, for instance, useful for someone who wants to wash the dishes.[3] Similarly, the technical conception of utility is subjective, in the sense that a state of affairs can only have utility *for* some agent and can never have utility independent of any agent. We can follow Don Ross and distinguish two major notions of utility. The first stems from the classical utilitarians and belongs to moral philosophy. I call this concept *classical utility*. The second arises in response to classical utility, and is a defining feature of neoclassical economics. This is the concept of *marginal utility*.

Using Ross's historically-situated, critical account of utility theory, this chapter discusses the distinction between classical and marginal utility in order to draw out two distinguishing features of the concept of marginal utility. The first is that marginal utility is hypothetical: the utility of a good for an agent relative to that of another good is dependent on her particular preferences or ends.[4] This differs from classical utility, which is non-hypothetical: the utility of an action is not dependent on an agent's particular ends, but is instead dependent on the universally-shared end of pleasure. So although both marginal and classical utility are subjective, the two concepts differ in terms of whether this subjectivism takes a hypothetical or non-hypothetical form. The second feature is that marginal utility is treated in a positivist manner; this means that the concept is only used to describe

[3] This observation draws on Peter Geach's distinction between predicative and attributive adjectives. Predicative adjectives are context-independent and therefore refer to intrinsic properties. Attributive adjectives are context-dependent and therefore denote extrinsic properties. Utility is one such extrinsic property. See Peter Geach, 'Good and Evil' in *Theories of Ethics* (Oxford: Oxford University Press, 1967).

[4] I use the terms 'preferences' and 'ends' interchangeably. The former is used in economics, the latter in philosophy. There is a slight difference between the two. Preferences are comparative – one never has a preference in isolation, but always a preference for one good over another. In contrast, ends tend to be understood in isolation from each other. I treat ends comparatively, however, so this difference is irrelevant to the discussion.

the relation between means and ends. This is in contrast to the normative approach to classical utility, by which the concept is also used to prescribe the relation between means and ends. I go on to suggest that, taken together, the hypothetical and positive characteristics of marginal utility imply a distinctive way of understanding agents' actions as isolated from, and immune to, the judgements of a community.

1. Classical Utility

'Utility' denotes the property of being useful, both in everyday usage and in more technical contexts. The classical utilitarians who drew their name from the term, defined utility as a special kind of usefulness: the ability to provide an agent with pleasure, rather than pain. Utilitarians, like Jeremy Bentham and John Stuart Mill, held that the moral goodness of actions should be judged by this kind of usefulness, for they believed that we should understand goodness solely in terms of a positive psychological state like pleasure or happiness, and its opposite, badness, in terms of a negative state like pain or unhappiness.[5] Once goodness is defined in this way, an action can only be good if it produces more pleasure than pain, not just for one individual but for society at large. A dictator, for instance, may derive a great deal of pleasure from building himself a palace of ice in the desert, but doing so may rob his people of scarce water and lead to an overall amount of suffering which condemns his action as morally reprehensible.

As Ross notes, Bentham developed this moral theory as a powerful conceptual tool against paternalism, the prevalent political view of Bentham's milieu.[6] Paternalism is the belief that one knows what is best for others, in virtue of which one makes decisions for them. Like a father who cares for his children, one is in an epistemically privileged position and this warrants one exercising a degree of benevolent authority over others. Yet, if the goodness of an action resides in the amount

[5] See, for instance, Mill's discussion of the meaning of utility in *Utilitarianism* (New York: Oxford University Press, 1998): Chapter 2.
[6] Don Ross, *What People Want: The Concept of Utility from Bentham to Game* Theory (Cape Town: Univeristy of Cape

of pleasure it produces for agents, then each agent is best-placed to know what is good for her, because she is uniquely positioned to know what her brings her pleasure and pain. It follows that governing bodies can never legitimately make decisions for agents on the basis that they know what is best for them. In addition, there is no basis for privileging one agent over another, for a unit of pleasure is the same regardless of who experiences the pleasure. This concept of utility therefore grounds a non-paternalistic moral theory, in which each agent's happiness is equally crucial to determining the moral worth of an action or public policy.

As Ross nevertheless points out, a narrow conception of pleasure confronts us with a pessimistic and blatantly false picture of human nature, for it then looks as if all agents are motivated to pursue their own happiness with no thought or care for others.[7] Bentham recognises these egoistic overtones of utilitarianism when he has an imaginary opponent exclaim:

> What a picture, old and gloomy-minded man are you giving us of human nature! [A]s if there were no such quality as disinterestedness – no such quality as philanthropy – no such quality as disposition to self-sacrifice . . . ![8]

Bentham's response is to acknowledge altruism, but argue that selfishness is predominant in society.[9] Unsatisfied with this response, Mill sought to widen the concept of pleasure to encompass a larger range of human aspirations and relations, so that utilitarianism would be independent of the truth of psychological egoism.[10] This allows a utilitarian to say that a mother's sacrifice for her child brings her pain, but this is less pain than she would have had if she had not aided her child. On the other end of the spectrum, a masochist seeks physical pain precisely because it brings him a great deal of pleasure. In both examples, the balance of pleasure over pain explains – and potentially

---

Town Press, 1999): 12-13.
[7] *Ross, What People Want*, 12-13.
[8] Jeremy Bentham, 'The Psychology of Economic Man', in *Jeremy Bentham's Economic Writings*(London: George Allen and Unwin: 1954)*:* XVIII.
[9] Bentham, *The Psychology of Economic Man*: XIX, XX.
[10] John Stuart Mill, 'On the Definition of Political Economy and the Method of Philosophical Investigation Proper to

justifies – agents' actions without assuming that agents are selfish. Here is the expansive, classical definition of utility:

> Classical Utility: *X* is good only if it results in more pleasure than pain for a community of agents, where an agent need not only gain pleasure from acting in her own interest

There are several characteristic features of this definition of utility. Firstly, although Mill divorces utility from psychological egoism by providing a more expansive definition of pleasure, the concept nevertheless refers to a specific psychological state. As such, utilitarianism is still entwined with psychology. The classical utility of an action depends on whether it does *in fact* provide more pleasure than pain for a group of agents. The task of a utilitarian is therefore in part that of a psychologist: she must measure or estimate the amount of pain or pleasure that an action produces or will produce in agents; she must then consider whether there are or will be more units of pleasure than pain, and on the basis of this she must assess whether the action is good or bad.

This last observation points to the second distinguishing feature of classical utility. On the utilitarian picture, an action or state of affairs has utility insofar as it leads to more pleasure than pain, and therefore to overall goodness for a community. This concept of utility is therefore a moral one, for it is not only used to describe the pleasure or pain that an agent gains from an action, but is also used as a criterion of wellbeing for agents, in virtue of which an agent has obligations to other agents.

Thirdly, although classical utility is dependent on an agent's particular psychological states of pleasure and pain, classical utilitarianism claims that all agents seek pleasure over pain.[11] This means that utility is a non-hypothetical concept, because it is not based on the particular ends of an

---

it', in *Collected Works of John Stuart Mill Vol 4* (Toronto: Toronto University Press).

[11] Bentham writes, for instance, that 'successfully, or unsuccessfully, [man] always aims at happiness and so will continue to aim as long as he continues to be man, in every thing he does' (*The Psychology of Economic Man*, Chapter

agent; instead, it is dependent on an end which all agents share. As I use it, a non-hypothetical concept is one which does not rely on an agent's particular ends; this means it can either rely on universally-shared ends, or it can be wholly independent of ends. An end is something towards which an agent aims; it is the target or goal of her action.[12] On the utilitarian view, pleasure is an end that is common to all agents, for it is something which all agents hold to be good and therefore worth aiming at.

2. Marginal Utility

On Ross' account, the birth of neoclassical economics in the nineteenth century marks a significant divergence from the concept of classical utility, for it formulates a *marginal* or relational definition of utility. The starting point of this formulation is to divorce marginal utility from any and all psychological states. William Stanley Jevons contends that the concept does not denote any intrinsic quality, but is better described as 'a circumstance of things arising out of their relation to man's requirements'.[13] Marginal utility is therefore a broader kind of usefulness than classical utility, for it signifies the ability of a state of affairs to realise an agent's ends, irrespective of whether it results in pleasure or pain, or any other kind of psychological state. In this sense, any state of affairs which is a means to satisfying an agent's ends has utility for that agent. Economists call an agent's ends *preferences*, and the means to those ends *goods*.

On the marginalist account, the locus of utility shifts from the happiness of an agent, and is instead located in the relationship between an agent's preferences and the goods which can realise those preferences. Marginal utility is therefore a thoroughly relational concept. Firstly, it captures the instrumental relationship between an agent's preference and a good.[14] Secondly, the notion of a preference is also relational: an agent cannot prefer an option A in isolation, but always prefers A to

---

2: I).

[12] This is taken from Aristotle's characterisation of ends in the *Nicomachean Ethics*, I.1-2.

[13] William Stanley Jevons, 'Theory of Utility', in *The Theory of Political Economy*(London: MacMillian Press, 1911):

some other option *B*.[15] Thirdly, there are always alternative goods which lie before an agent, and she must consider which goods can best help her satisfy her preference for *A* over *B*. So the utility of a good for an agent is always relative to that of another good for her.

Jevons takes this relational concept of utility and adds a plausible psychological assumption. He claims that if one has a preference for *A* over *B*, and a good *X* satisfies preference *A* more than another good *Y*, then the more one has of *X* the more one's preferences are satisfied, up until a certain point, where there is a levelling off of satisfaction.[16] If you are hungry, then a spoon of food will help ease that hunger more than a breath of air, and a plate of food will satiate you, but after a certain point, eating more food instead of breathing will not satisfy your preference for food to any noticeable extent.

This is the principle of diminishing marginal returns: the utility of a good approaches zero as the amount of the good increases beyond a certain point. Given this baseline at which the utility of a good approaches zero, we can compare the degree to which a certain quantity of a good realises a preference with that of an infinitesimally small increase in quantity. The difference between the two will tell us the extent to which a preference is realised by the acquisition of a certain quantity of a good, *relative* to an increase or decrease in the good. So, although we cannot get an absolute grip on the degree to which a person is satisfied by a good, we have a comparative understanding of the utility of a good for an agent relative to an increase or decrease in that good. This is the concept of marginal utility.

> Marginal Utility: *x* has greater utility than *y* for an agent only if it could satisfy her preference for *A* over *B*, where the degree to which *x* satisfies her preference for *A* over *B* is relative to a corresponding increase or decrease in the quantity of *x*.

92.
[14] *Ross, What People Want*, 4-30.
[15] John Broome, *Ethics out of Economics*(Cambridge: Cambridge University Press, 1999): 8-11

Ross argues that the principle of marginal returns allows economists to define part of the qualitative aspect of utility in quantitative terms.[17] Consider an agent who has a preference for wine over beer; yet she is not indifferent to the kind of wine she drinks, for one unit of box-wine does not satisfy her preference as much as one unit of a finer wine. In this case, the quantity of wine is the same, but the two are qualitatively different for her, and therefore not straightforwardly substitutable. It is fairly easy to measure units of a particular good, because we already just do carve up the world into units, and so we can use these pre-existing empirical units as the basis for quantitative measurements. On the face of it, however, there are no pre-existing empirical units which could be used as the basis for qualitative measurements, because the intensity or the degree to which a person is satisfied by a certain good is dependent on facts about her. Jevons sidesteps the problem of how to provide an independent criterion of varying levels of intensity by defining the intensity of a good's utility in terms of changes in the amount of that good.

This gives us one reason for favouring marginal utility over classical utility – it can be used to measure and compare the value of goods for an agent. The principle of marginal returns makes it possible to place the utility of goods for an agent in a hierarchy and thereby compare the differing levels of utility that the acquisition of different goods have for an agent. This is helpful in considering the question of how to distribute scarce resources such that the greatest number of agents can satisfy their preferences, for the provision of goods with the greatest utility for an agent will satisfy her preferences more so than the provision of goods which have relatively little utility for an agent. Nevertheless, such a ranking still leaves it open as to whether an agent has incomplete preferences. For instance, faced with the option of drinking Rooibos tea or Ceylon tea, an agent does not have a preference for Rooibos over Ceylon and neither does she have a preference for Ceylon over Rooibos.[18] Without further conditions in place, the concept of marginal utility only implies the possibility of a partial ordering of goods.

---

[16] Jevons, 'The Theory of Utility', in *The Theory of Political Economy* (London: Macmillan Press, 1911).
[17] *Ross, What People Want*, 25.
[18] This is a variation of Amartya Sen's example of Buridan's ass. In the story, a donkey is faced with two haystacks;

3. Comparing Marginal and Classical Utility

There are several features of marginal utility which distinguish it from the utilitarian concept. Firstly, a preference is an empty concept, in the sense that it lacks specific content. Unlike particular psychological states, the concept of a preference is therefore applicable to any agent in any situation. Whether a mother does in fact derive pleasure from sacrificing herself for her child is open to psychological speculation. It may be the case that she derives not a single unit of pleasure from doing so, and until a utilitarian knows how much pleasure results from her action, he cannot pronounce on the utility of her action. In this case, the explanation (and potential justification) of her action is complicated by and contingent upon psychological findings. On the neoclassical account, however, the explanation is simpler and less contingent. Regardless of how much pleasure is involved in her sacrifice, one can still say that the mother's sacrifice has the ability to satisfy her preference to save her child; the utility of this sacrifice relative to her not sacrificing herself would then explain her action. The neoclassical account therefore enhances Mill's attempt to distance classical utility from the truth-value of a particular psychological theory, by divorcing marginal utility from any particular psychological state.

Secondly, defining marginal utility in terms of preference satisfaction has the effect of exempting marginal utility from counting as an obvious moral concept. To see what I mean, consider whether the satisfaction of preferences is a morally good thing. By 'good' I simply mean our ordinary use of the word to denote admirable qualities. It might be that many cases of preference satisfaction are good; my preference to be kind rather than cruel to strangers might be good, or my preference for an education over illiteracy might be good. Yet, it does not seem like we would ordinarily condone the satisfaction of a preference for rape over respect for a woman's sexual integrity, for instance, as *good*. On the face of it, preference satisfaction might be morally good in

---

unable to choose between the two the donkey dies of starvation. See Sen, 'Maximisation and the Act of Choice', in

some cases, and in others it might be bad. This suggests that the concept does not have enough content for it to count as a clear synonym for moral goodness and some argument is necessary to show that it is a moral concept. In contrast to utilitarianism then, the marginalist definition of utility seems to be morally neutral and can be used to explain both morally good and bad phenomena.

In this regard, Lionel Robbins quotes Cannan who complains that a 'large trade has existed since history began in supplying certain satisfactions of a sensual character which are never regarded as economic goods.'[19] Robbins' response is worth quoting in full:

> Economists, equally with other human beings, may regard the services of prostitutes as conducive to no 'good' in the ultimate ethical sense. But to deny that such services are scarce in the sense in which we use the term and that there is therefore an economic *aspect* of hired love, susceptible to treatment in the same categories of general analysis as enable us to explain fluctuations in the price of hired rhetoric, does not seem to be in accordance with the facts.[20]

As John Broome argues, this ambiguous use of the term 'utility' still haunts modern economics.[21] Economists often slip into using the older concept of classical utility, which denotes the moral goodness of an action, all the while maintaining that they use the morally neutral notion of marginal utility, without providing an argument for this shift. Broome insists that consistency and clarity demand that if economists explicitly adopt marginal utility (and they do in fact do so), then they should not covertly use the utilitarian concept. This does not imply that marginal utility cannot function as a moral concept, but it is first necessary to show how it is a moral concept. Neither does it suggest that moral concerns do not or cannot play a significant or fundamental role in economic analyses, but marginal utility is not an obvious contender for grounding these concerns.

Rationality and Freedom(New Delhi: Oxford University Press, 2008): 184.

[19] Lionel Robbins, 'Ends and Means', in *On the Nature and Significance of Economic Science* (London: Macmillan Press, 1969): II.

Yet Jevons and Robbins make an even stronger distinction between marginal and classical utility. They claim that not only is marginal utility not a moral concept, but it is also not a *normative* one. Moral concepts are a subset of normative concepts. To think of the world in a normative way, as Korsgaard writes,

> [is to] think of ways that things could be better, more perfect, and so of course different than they are; and of ways that we ourselves could be better, more perfect, and so of course different, than we are . . . [Normative ideas] *outstrip* the world we experience and seem to call it into question, to render judgement on it, to say that it does not measure up, that it is not what it ought to be.[22]

Unlike positive concepts which simply describe the way the world is, normative concepts prescribe the way the world ought to be; it is for this reason that Korsgaard characterises normative concepts as those which outstrip the world. Normative concepts include rules of logic, epistemic values like truth and coherence, scientific values like elegance and simplicity, and aesthetic values like beauty and harmony. Moral concepts like classical utility, honesty and kindness, which govern the conduct of agents towards each other, are therefore just one kind of normative concept amongst many.

Jevons, in particular, characterises marginal utility as a fundamental, purely descriptive tool for analysing agents 'not as they ought to be, but as they are.'[23] His argument proceeds largely from elimination: if marginal utility is not a moral concept, it must be a positive one. Similarly, Robbins develops this line of thought in his argument that human ends are not amenable to rational deliberation or discussion.[24] This means that ends are purely conative or arational. Consequently,

---

[20] Robbins, 'Ends and Means': II.
[21] Broome, 'Utility', in *Ethics out of Economics*.
[22]Christine Korsgaard, *The Sources of Normativity,* (Cambridge: Cambridge University Press, 1996): 1. Emphasis added.
[23] Jevons, 'Theory of Utility': 2.
[24] Robbins, 'Interpersonal Comparisons of Utility', in *Economic Journal* 48, 192(1938):635-641.

Robbins contends:

> If we disagree about ends it is a case of thy blood or mine – or live or let live according to the importance of the difference, or the relative strength of our opponents. But if we disagree about means, then scientific analysis can often help us resolve our difficulties. If we disagree about the morality of the taking of interest . . . then there is no room for argument.[25]

On this view, if the object of analysis is an agent's ends, then the analysis is a moral one. If the object of analysis is the means to an agent's ends, however, then the analysis is a scientific one. Robbins understands a scientific analysis as a positive or descriptive enterprise.[26] Like Jevons, he too contrasts a moral analysis with a positive one; he thinks that if the concept of utility is not used in moral analyses, it must be used in positive ones. Hence, the concept of marginal utility does not outstrip the way agents are by pointing to the way they ought to be or ought to act, but instead expresses their existing natures, by denoting the way they are and the way they act.

I will put aside the question of whether moral and descriptive properties are exhaustive categories and consider this question in the following chapters. For the moment though, we can see that the positive treatment of marginal utility does not imply that the whole of economics is a positive or descriptive discipline, limited only to the explanation and prediction of economic phenomena (although Jevons and Robbins seem to think so). Ross, for instance, who endorses a positivist interpretation of economics, can still make room for normative considerations in the discipline, because he thinks that what interests and guides economists in their investigations are typically moral concerns like welfare.[27] He maintains that the concepts with which economists

[25] Robbins, 'Means and Ends': 132.
[26] Robbins, 'The Subject Matter of Economics', in *On the Nature and Significance of Economic Science:* 139.
[27] More particularly, Ross is concerned to show that economics is guided by a classical utilitarian outlook.

work, in particular the notion of marginal utility, are nevertheless descriptive.[28] Similarly, Daniel Hausman and Michael McPherson argue that positive economic analyses are often steered by the moral values of individual economists and can be refined by sustained reflection on these moral values.[29]

Thirdly, the concept of marginal utility is arguably a hypothetical one. Although I am not aware of any economist who explicitly claims that the concept is hypothetical, I nevertheless think that it is plausible and useful to interpret marginal utility as dependent upon an agent's particular preferences. To begin with, a preference does not have much content, for it just denotes the end or aim of an agent's actions. As such, one could have a preference for anything, and the realisation of a preference could be attached to an array of affective states. There is therefore not enough information with which to make claims about whether agents share certain preferences. If this is correct, then at a minimum, the empty nature of a preference leads to agnosticism about whether specific preferences are shared by agents. Moreover, as Robbins argues, economics confines itself to an analysis of the way in which agents act to satisfy their preferences, given the scarcity of goods to satisfy those preferences. This means that economic analyses are neutral about the nature of preferences.[30] In both cases then, any positive views about the universal nature of preferences are absent. In light of this agnosticism, it is reasonable to interpret marginal utility as dependent upon preferences which are particular to agents, until it is proven otherwise, for at the very least, we only know that some preferences are not shared by all agents. I, for instance, can attest to having *never* had a preference for an embroidered punjabi over a plain one, although several friends of mine have clear preferences regarding embroidered punjabi over plain ones. By default then, it seems as if the concept of marginal utility is hypothetical.

Yet there are stronger grounds for holding that marginal utility is hypothetical. Recall that the marginal utility of a good for an agent is dependent on its relation to the utility of another good

[28] Ross, *What People Want*, 371-373.
[29] Hausman and McPherson, *Economic Analysis, Moral Philosophy and Public Policy*(Cambridge: Cambridge University Press, 2006).

for that agent, and an increase or decrease of that good relative to her preferences. This means that utility is not just dependent on a preference in isolation, but is dependent on an agent's preference structure. It might be plausible to claim that all agents share a particular preference, but it strikes me as implausible to claim that all agents share the same complex combinations of preferences. To see why, we can imagine a situation in which there are two agents: Xoli and Castro. Xoli is completely unaware of the existence of butternut ice-cream but knows about chocolate ice-cream, while Castro knows of and has tasted both ice-creams. It does not seem as if we could sensibly ascribe preferences regarding chocolate and butternut ice-cream to Xoli, whereas it is easy to imagine that Castro has preferences regarding chocolate and butternut ice-cream. So it seems like Xoli and Castro will have different preference structures simply because Castro knows about butternut ice-cream, whereas Xoli does not. If this is correct, then the concept of marginal utility is dependent upon an agent's particular combination of preferences, a combination which is not universally shared.

It is worth pointing out that both marginal and classical utility are subjective: a state of affairs only has marginal or classical utility *for* an agent. The distinction between hypotheticality and non-hypotheticality is useful in picking out two significantly different kinds of subjectivism. As discussed, the concept of classical utility depends on ends which all agents share and is therefore non-hypothetical. Coupled with the concept's normative dimension, this allows classical utility to function as a criterion for judging actions. If acting on a preference for an ice palace in the middle of a desert results in more pain than pleasure, then the action has negative utility and can be criticised as morally bad. So although classical utility is a subjective concept, it allows us to say that preferences which might be acted upon are open to rational deliberation and moral criticism from all. While a utilitarian judgement takes the form of a moral evaluation, there are many other kinds of judgements which non-hypotheticality could allow, like judgements about rationality, epistemic judgements, aesthetic ones and so on.

[30] Robbins, 'Ends and Means': II.

Marginal utility, however, seems to be limited to the particular preferences of an agent, and does not make reference to universally-shared preferences. In conjunction with positivism, this hypotheticality means that marginal utility does not provide a common standard by which an agent can deliberate on his preferences and judge the actions of others. If a dictator prefers to build an ice-palace in a desert over other options, then doing so has utility for him and this provides him with a reason to begin construction. The question of whether it has low utility for others is a separate matter, for the analysis of what has utility for him is confined to his particular preference set. If this is correct, then the hypothetical nature of marginal utility leads to a view of action which is substantially different from that of classical utilitarianism, for the only guide as to what has marginal utility for an agent is the agent herself, rather than the community in which she is embedded. On this view then, an agent's actions are understood in isolation from her community, where this isolation precludes others' judgement of her actions in terms of marginal utility.

Conclusion

Although marginal utility is rooted in classical utility, there are significant differences between the two concepts. Below is a table which summarises the differences:

	Classical Utility	Marginal Utility
1.	Normative: moral concept	Positive: descriptive concept
2.	Non-hypothetical: depends on universally-shared ends	Hypothetical: depends on a particular combination of ends
3.	Implication: Acts as universal criterion for deliberating about preferences that we act on, and for criticising the actions of others	Implication: Does not act as a universal criterion for deliberating about preferences that we act on, and cannot be used to criticise the actions of others

Thus far, the concept of marginal utility is a purely abstract entity, in the sense that it has yet to be used in explanations or predictions of an agent's behaviour. The next chapter looks at marginal utility when it is used to explain agent's behaviour. I focus on the way in which economic explanations of human behaviour seem to view this behaviour as an instance of rational action, and I then investigate which conditions are necessary for this kind of explanation to work. I think that with these conditions in hand, we can compare the theoretical treatment of marginal utility with its application in explanations of behaviour.

CHAPTER 2

Explaining our Behaviour: The Practice of Marginal Utility

Here is a highly simplified explanation of an agent's actions in terms of marginal utility. When faced with a choice of coffee or tea each morning, Xoli always drinks coffee. The repeated selection of coffee over tea reveals her preference for coffee over tea. It follows that coffee has greater marginal utility than tea for Xoli, and this is why she chose coffee instead of tea this morning.

In this example, the explainer seems to view Xoli's behaviour as an instance of rational action. Firstly, her behaviour is not treated as aimless or random: Xoli drinks coffee in order to satisfy her preference for coffee over tea, or in philosophical terms, Xoli drinks coffee as a means to realising her end of drinking coffee over tea. In virtue of its intentional nature then, her behaviour counts as an action, and because this action is selective in nature, it counts as a choice. Secondly, the explainer interprets her action as an appropriate means to satisfying a preference for coffee, rather than a preference for something else like tea, or beer, or any other good. As such, we are invited to see Xoli's preference for coffee as giving her a reason to drink coffee, although the explainer is silent on whether it is a good reason. In this explanation then, Xoli's behaviour seems to count as something which is rational, in the sense that she acts on a reason; it may be more or less rational depending on whether she acts on a good or a bad reason.

Furthermore, the explanation is articulated in instrumental terms. We understand Xoli as rationally choosing the means to satisfy her ends, rather than acting on non-instrumental deliberations concerning the nature of her ends, for instance. This suggests that the instrumental principle plays a role in explanations which employ marginal utility. It is a rule which governs agents' reasons for acting, in cases where agents act to take the means to their ends. One way of framing the principle is this: the possession of an end gives you a reason to take the means to that end.

This chapter assesses the role of the instrumental principle in explanations of agents' behaviour that view this behaviour in terms of rational action. I begin with Peter Railton's account of the instrumental principle, which aims to show that the principle is indispensable to our understanding of rational action. I argue that Railton's account works, because the principle governs the shape of practical reasons. In virtue of its logical structure, the principle necessarily plays a normative and non-hypothetical role in explanations of rational action. If this is correct, then an explainer is faced with only two options: if he explains an agent's behaviour in terms of rational action, then his explanation must invoke the instrumental principle, which is normative and non-hypothetical. If he denies that the explanation invokes a normative instrumental principle, then he cannot explain an agent's behaviour in terms of rational action. Insofar as marginal utility is used in an explanation of rational action, it is governed by the principle and has two properties: it is normative and non-hypothetical.

1. Rationality as Consistency

Decision theory in economics sets out a number of conditions which define rational choice. These conditions are limited to an agent's preference structure, or the pattern of her ends. The constraints are largely requirements of consistency: an agent is irrational, for instance, if she has circular or non-transitive preferences.[31] One constraint, in particular, is relevant to this chapter. It stipulates that an agent's preferences about her means should be consistent with her preferences about outcomes. Suppose you prefer outcome A to outcome B, then on this condition, it is irrational for you to prefer the means to B over the means to A.[32]

Although not explicitly articulated as such, this condition is a version of the instrumental principle, for the principle is concerned with the relation between means and ends. This formulation

[31] See, for instance, Paul A Samuelson, 'The Empirical Implications of Utility Analysis' in *Econometrica 6*(Oxford: Blackwell Publishers, 1938).
[32] This formulation is drawn from John Broome, 'Can a Humean be Moderate?', in *Ethics out of Economics*: 68.

of the principle is framed solely in terms of the pattern of preferences, and does not include reference to concepts like reasons or beliefs. It does not assert that a preference for sobriety gives you a reason to prefer the means to satisfying this preference by abstaining from alcohol, nor does it stipulate that the chosen means should ordinarily be effective in securing your preference. Instead of these conditions for correspondence between means and ends, it only declares that if you prefer one goal over another, like sobriety over intoxication, then it is irrational to prefer copious drinking over more abstemious behaviour. As such then, it is a fairly minimal formulation of the principle.

I believe this minimal version of the instrumental principle conceals within it a stronger condition of correspondence between means and ends. Let us return to Xoli and her chronic pursuit of coffee. Suppose we do not presume that drinking coffee corresponds with a preference for coffee; that is, suppose we do not imagine that her action is intended as an efficacious way of realising a preference for coffee over anything else. What are we to make of her action then? Without a criterion for correspondence between means and ends, we can as much infer that her action is a means to satisfying a preference for tea or roast beef as for coffee, because the connection between means and ends is a matter of whim, rather than an empirically fixed pattern of correspondence. So, even though Xoli prefers coffee to tea, it is not inconsistent of her to prefer drinking tea to coffee, because drinking tea might be her way of satisfying a preference for coffee. This suggests that a rule for consistency about preferences for means and ends is a corollary of a more fundamental rule for correspondence between means and ends. I will therefore take the instrumental principle to encapsulate a correspondence rule between means and ends.

2. The Structure of the Instrumental Principle

What does a principle that stipulates a correspondence between means and ends look like? Here is one formulation: the possession of a relevant end provides an agent with a reason to take the acknowledged means to that end. There are other formulations, but for reasons which I discuss later,

this one has the merit of being fairly uncontroversial.[33] If Xoli desires coffee, and believes that drinking coffee will satisfy this desire, then she has a reason to drink coffee. This reason has two elements: (i) an end and (ii) a belief connecting the means to that end. It is a practical reason, rather than a theoretical one, because it is concerned with an agent's actions and not her beliefs.[34] Further, this reason applies exclusively to actions; it does not apply to unintentional behaviour, like tickles and sneezes.

Notice that the stipulated connection between means and ends in terms of reasons amounts to the requirement that an agent give her ends deliberative weight in her reasoning. For an agent to recognise that an end gives her a reason to pursue the corresponding means is for her to accord deliberative weight to her end in deliberations about how to act, such that she considers pursuing the corresponding means to that end. I will occasionally characterise the principle in terms of deliberative weight for the sake of concision, but this should be taken as interchangeable with a principle framed in terms of reasons. This does not mean that an agent represents her deliberations about how to act as a reason. She may not always be aware of her deliberations, and she may not think of these deliberations as 'reasons'. Nevertheless, when an agent's beliefs and ends combine in such a way as to weigh in favour of her pursuing an action, then this combination allows us to understand the action by showing that it is, from her point of view, appropriate. A basic and undefended assumption in this chapter is that this explanatory role is taken up by a practical reason: it is a favouring relation which can render an agent's actions intelligible to others.[35] A practical reason may have other functions – it might be used to guide another's behaviour or express disapproval at her conduct – but I will not discuss these.

---

[33] Other formulations include substantive views on the nature of an end. Korsgaard, for instance, formulates the principle like this: if you have a reason to choose an end, then you have a reason to take the means to that end. This means that you only ought to take the means to your end if, upon reflective deliberation, you ought to have the end in the first place. Such a formulation presupposes that ends *must* be cognitive, and only works from the perspective of a Kantian position on practical reason. See Korsgaard's 'The Normativity of the Instrumental Principle' for further discussion.

[34] More precisely, a practical reason is concerned with what an agent ought to do, while a theoretical reason is concerned with what an agent ought to believe.

[35] The idea of a practical reason as a favouring relation is taken from Jonathon Dancy, 'Enticing Reasons' in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (Oxford: Clarendon Press, 2006). The idea of a practical reason as rendering actions intelligible is drawn from Bernard Williams, 'Internal and External Reasons' in *Moral Luck*

We can now distinguish between motivating and normative reasons.[36] Xoli believes that a cup contains coffee and desires coffee, so she has a motivating reason to drink the contents of the cup. In actual fact, however, the cup contains tea. This means she does not have a normative reason to drink the contents of the cup, because her deliberation is based on a false belief. At a minimum, the distinction between the two kinds of reasons lies in whether we take the truth-value of her belief into account. The truth-value of a belief is not an element of a motivating reason, but it is an element of a normative reason. One would expect an explanation of an agent's rational action to draw primarily on an agent's motivating reasons. Xoli's false belief that the cup contains coffee coupled with the desire for coffee provides a motivating reason which explains her drinking tea, but she has no normative reason to drink the tea. In this case, the explanation cannot plausibly draw on a normative reason. On the other hand, if her belief was true and the cup contained coffee, then she would have both a normative and a motivating reason to drink the contents of the cup, so that the normative reason would coincide with her motivating reason and could be used to explain her action.[37]

It is plain that the instrumental principle governs an agent's normative reasons for acting, by stipulating that an agent ought to give deliberative weight to his ends.[38] This means that his reason for acting should connect the action to an end of his, so that the end and the means to his end have the right kind of fit. Bernard Williams characterises this right kind of fit in terms of the possession of true beliefs.[39] If Castro drinks paraffin on the mistaken belief that it is water and will quench his thirst, then his action has missed its target. Similarly, if he drinks paraffin on the mistaken belief that paraffin will quench his thirst, he has likewise failed to fit the means and the end together. The stipulation that one ought to have true beliefs is a theoretical rule, but the stipulation that one's

---

(Cambridge: Cambridge University Press, 1981).

[36] This distinction is taken principally from Michael Smith, *The Moral Problem*(Oxford: Blackwell, 1995): 131-132.

[37] This invokes Williams' constraint on normative reasons as being *capable* of explaining an agent's actions once acted on. See Williams, 'Internal and External Reasons', in *Moral Luck* (Cambridge: CUP 1981).

[38] This characterisation of the instrumental principle in terms of deliberative weight is taken from Peter Railton, 'On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Actions', in *Ethics and Practical Reason,* edited by Garret Cullity and Berys Gaut (Oxford: Clarendon Press, 1997).

[39] This is a variation on an example taken from Bernard Williams, 'Internal and External Reasons' in *Moral Luck*.

means and ends ought to fit together by virtue of true beliefs is a practical rule, because it is concerned with action. This rule is the instrumental principle, and in both examples, it plays a clear normative role, for it lays down a standard which outstrips the agent's actual, motivating reason and points to ways in which this reason can improve its fit between means and ends by the possession of true beliefs. There may be other ways in which the instrumental principle stipulates the right kind of fit, or correspondence, between means and ends, but the possession of true beliefs is a clear and compelling instance of this correspondence. It therefore seems fairly obvious that the principle governs normative reasons, but we need to enquire further as to whether it governs motivating reasons as well.

Although this conception of the instrumental principle is stronger than one framed solely in terms of consistency of preferences, it is nevertheless fairly undemanding. Firstly, characterising the principle in terms of deliberative weight is weaker than characterising it as stipulating that one ought to take the means to one's ends. This is because it is sometimes ill-advised to take the means to one's ends. It might, for instance, be imprudent and even irrational to indulge one's preference for sobriety, because it conflicts with other ends like intoxication or revelry. Yet although one might not pursue sobriety, one still accords this end deliberative weight in reasoning how to act. So this formulation of the instrumental principle can accommodate instances in which one ought not to take the means to one's ends, or instances in which one does not take the means to one's ends, where doing so is rational.[40]

Secondly, this characterisation of the principle does not specify what the nature of an end is. An end might be purely conative, like a desire. It might be purely cognitive, like a belief. Or it might have both aspects, like a desire that an agent chooses to pursue based on reasons. It is therefore neutral between competing accounts of practical reason concerning whether a reason to act is always relative to the conative motivations of an agent. This means it does not hang on any

[40] The problematic formulation I have in mind belongs particularly to James Dreier: "If you desire to ψ, and believe that by φ-ing you will ψ, then you ought to φ". See Dreier, 'Humean Doubts about the Practical Justification of Morality', in *Ethics and Practical Reason*: 38.

particular outcome of the reasons internalism/externalism debate.[41] I think that this renders the formulation of the principle fairly uncontroversial, and in particular, those who are committed to the explanatory potential of marginal utility can accept this formulation, because it does not conflict with the neoclassical interpretation of preferences as purely conative.[42]

Finally, if one views explanation as primarily a descriptive enterprise, then one can accept the normativity of the instrumental principle, for it is not obvious that it applies to the motivating reasons which are used in explanations. The treatment of marginal utility as a positive concept is consistent with this, for explanations which draw on marginal utility are primarily composed of motivating reasons, and it still remains to be shown whether motivating reasons are governed by a normative instrumental principle.

At first blush then, the instrumental principle only plays a peripheral role in explanations of agents' actions, by governing the fit between means and ends in terms of the possession of true beliefs. As the truth-value of a belief only enters into normative reasons, it appears that the principle does not apply directly to the motivating reasons which are fundamental to explanations. Although the principle is normative, this normativity does not seem to be in conflict with the positive treatment of marginal utility.

3. The Indispensability of the Instrumental Principle

In the last section, I framed the instrumental principle in terms of a stipulation about the right kind of fit, or correspondence, between means and ends in the form of a reason to act. In this section, I discuss Peter Railton's account of the instrumental principle, which aims to demonstrate that the principle has an indispensable role in practical reason. This is a familiar view, and otherwise competing accounts of practical reason agree on the bedrock status of the principle. The task of this

[41] Roughly speaking, reasons internalism claims that a normative reason for an agent to act makes necessary reference to her motivations, which are conative. Reasons externalism denies this. I discuss the debate in a little more depth in the Coda.

section is to consider how the indispensability of the principle can function as common ground between rival theories of practical reason. I suggest that the feature which explains the success of Railton's argument is the principle's logical structure: it does not govern the content of practical reasons, but instead governs the structure of practical reasons. I go on to argue that the logical structure of the principle does not support instrumentalism about practical reason, and therefore has no bearing on the debate between theorists as to whether all rules in practical reason, like moral rules, are ultimately instrumental in nature.

Railton asks: imagine you object to following the instrumental principle.[43] You agree that you want to quench your thirst and you believe that drinking a glass of water will do so, but you are not motivated to drink the water; the end of quenching your thirst has no deliberative weight in your reasoning. You ask why you ought to be motivated in the first place. Say I cite your preference for being rational, and the principle as a way of meeting the criterion for rationality. Yet you object that you see no reason to follow the instrumental principle. You ask for a reason to give deliberative weight to your ends.

Here is the practical syllogism, the conclusion of which you reject:

i. *E* is an end of mine

ii. Means *M* would secure *E*

Conclusion: E has deliberative weight in my reasoning; that is, it moves me to consider pursuing *M*[44]

You deny the conclusion, but you accept the truth of the premises. So, you must think there is a problem with the structure of the argument – it is missing a premise. On the face of it, the missing premise is that you have some further aim, call it *F*, of choosing so as to bring about the realisation

[42] I discussed Robbins' conativism about ends in the first chapter. See his 'Means and Ends', in *On the Nature and Significance of Economic Science,* 132.
[43] Peter Railton, 'On the Hypothetical and the Non-Hypothetical' in *Ethics and Practical Reason.*

of your ends:

iii. *F* [choosing so as to bring about the realisation of my ends] is an end of mine

Yet, because you do not feel the weight of your ends in your deliberation, this further end will not bring you to accept the conclusion as a means to realising *F*. So you still grant the first three premises without accepting the conclusion. You are forced to add increasingly complex premises, but none of them can compel you to accept the conclusion unless you already follow the instrumental principle.

To understand the significance of this regress, consider that Railton's argument is a variation of Lewis Carroll's Tortoise Argument for the indispensability of theoretical rules of inference, like *modus ponens*.[45] Achilles entertains an argument which has the form of a *modus ponens* inference:[46]

i. If *p* then *q*

ii. *p*

Conclusion: *q*

Tortoise asks whether the structure of the argument is flawed. He thinks that one could grant the first two premises, but not accept the conclusion, unless one also granted:

iii. If [(if *p* then *q*) & *p*] then *q*

Achilles consents, but now Tortoise merrily argues that in order to accept (iii) we need an additional premise:

---

[44] This is a version of Railton's formulation in 'On the Hypothetical and the Non-Hypothetical': 77.
45 Lewis Carroll, 'What the Tortoise Said to Achilles', *Mind* 4, 14(1895): 278-280.
[46] This is a simplification of the argument. Carroll frames it in terms of Euclid's first proposition: 'things which are

iv. If {([(if *p* then *q*) & *p*] then *q*) & (if *p* then *q*) & *p*} then *q*

It has slowly dawned on Tortoise that he has launched a regress, because he is forced to add increasingly complex premises, but none of them can get Achilles to accept the conclusion, unless Achilles consents to make a *modus ponens* inference. The lesson to be learnt from Tortoise's story is that rules of inference cannot function as premises in an argument, on pain of regress. Railton thinks this implies that just as one cannot ask for a reason to follow *modus ponens* without already invoking *modus ponens*, one cannot ask for a reason to follow the instrumental principle without already following the principle. If this is correct, then the principle is indispensable to practical reason, for it is so bound up in our reasoning about how to act that the very idea of a practical reason only makes sense if one follows the instrumental principle.

At first glance, it seems like this conclusion must be right. Notice, however, that the principle can only play an indispensable role in practical *reason* if it governs the reasons one might have to take the means to one's ends. Yet the principle might not be at all concerned with reasons, but simply with the relation between means and ends.[47] This relation takes the form of what Broome calls a *requirement*.[48] It stipulates: you ought to see that if you have an end, then a certain action is the means to that end. For instance, you ought to see that if you want to get to Johannesburg today, then catching the train at 16:00 is the means to doing so. The principle does not give you a reason to catch the train. You might, after all, have no reason to be in Johannesburg; in this case, you would not have a reason to catch the train unless you also had a reason to go to Johannesburg. Nevertheless, your intending to go to Johannesburg normatively requires you to intend to catch the train at 16:00.

This is because the principle governs your *attitudes*: you have an attitude to making it true

---

equal to the same are equal to each other'. In the argument, the proposition is applied to the sides of a triangle, and the consequent holds that the sides of a triangle are equal.

[47] Thank you to John Brunero for pointing out this objection to me.

that you are in Johannesburg, and coupled with the belief that catching the train is necessary to do so, you are required to have an attitude to making it true that you are on the train to Johannesburg at 16:00. As Broome thinks of it, practical rationality "does not consist entirely in acting for good reasons, as is commonly supposed. To a large extent it consists in following normative requirements. Consequently, rationality may bring you to do things you have no reason to do."[49] On this view then, the instrumental principle does not provide one with reasons to act, so it is not indispensable to practical *reason*, although it might turn out to be indispensable to practical rationality. All that the Tortoise argument can demonstrate is that you cannot have the principle as an end in your deliberations about how to act and simultaneously expect it to guide you.

This objection derives its appeal by allowing for scenarios in which it is rational to take the means to one's end, so that instrumental reasoning is not paralysed if one's ends happen to be those one should not have. This is an attractive position, and it works because the principle is drawn in relational terms: it governs the relation between means and ends, rather than stipulating which means one ought to take. I will discuss the divergence between correct instrumental reasoning and ends one should not have in Section 4. For the moment, I want to focus on whether viewing the principle as a requirement grounds an objection against the indispensable status of the instrumental principle in practical reason. I do not think that it can do so. One can affirm that (i) the principle governs the relation between means and ends and (ii) the principle does not *provide* reasons to act and nevertheless maintain that (iii) the principle can be framed in terms of reasons.

As a requirement, the principle insists that an agent's ends have some kind of force in her deliberations about how to act, such that an end coupled with an appropriate belief moves her to consider taking the corresponding means. This combination of a belief and an end weighs in favour of pursuing a particular action. In cases where an agent acts on her ends, this favouring relation would help explain why she acted the way she did. Yet, a favouring relation that can explain an agent's actions is precisely what practical reasons are. If this is correct, then interpreting the

[48] Broome, 'Normative Requirements', in *Normativity* (Oxford: Blackwell, 2000).

instrumental principle as a requirement suggests that the principle governs the shape of a practical reason so that it fits means and ends together. Consequently, it is not inconsistent to claim that the principle does not provide reasons to act, and simultaneously to claim that it governs the shape of reasons to act. On the contrary, viewing the principle as a requirement involves viewing it as a shaper of practical reasons.

Moreover, it looks as if the Tortoise argument actually implies that the instrumental principle is a requirement or a shaper of practical reasons. Railton explicitly draws our attention to the way in which Tortoise's story shows us that rules of inference cannot function as premises in an argument, on pain of regress. Yet this story is also supposed to show us that rules of inference still have a role to play concerning the argument. To my mind, the most obvious way of explaining this role is to interpret rules as governing the form of the argument. While modus ponens governs the shape of a theoretical reason by determining the right kind of fit between an antecedent and a consequent, the instrumental principle stipulates the shape of a practical reason so that it fits means and ends together.

Although Railton does not explicitly state this, the logical structure of the principle is an important feature in his argument for the indispensable status of the principle. If the principle governed the contents of a reason to act, then it seems as if we could have reasons to act independently of following the principle. This is because we often encounter ourselves and others as acting on reasons which have poor content, like false beliefs. If a reason is shaped by the principle, however, then practical reasoning cannot take place without a basic commitment to following the principle. I think we can understand this claim better by looking at a simple analogy with language. The instrumental principle is like the syntax of a sentence, while the beliefs and desires of practical reasons are like the semantic content of a sentence. Although rules which govern the meaning of words can sometimes be dispensed with (one can invent new meanings, for instance, or twist old meanings), it seems like an individual cannot communicate without following some

[49] Broome, 'Normative Requirements': 97.

kind of syntax. Even if she does so badly, she will nevertheless use words or gestures that refer to an object or express an emotion, and in doing so, obey elementary rules which structure our communication with each other. Similarly, an individual cannot have a reason to act without following the instrumental principle. Any action is undertaken to fulfil an end, and where there is a connection between an action and an end, there is a commitment to following the principle, even if this is only done imperfectly. So although one could have an immoral or imprudent end or one could have false beliefs, one might nevertheless still have a reason to act because this reason conforms to some extent with the instrumental principle.

The logical structure of the instrumental principle therefore explains why the Tortoise argument can successfully ground the indispensable status of the principle in practical reason. The Tortoise argument works because it shows that a practical reason is only intelligible if shaped by the principle, and that as a consequence, practical reasoning cannot take place without a basic commitment to following the principle. If this is correct, then interpreting the principle as a requirement does not undercut Railton's argument; instead it supports the view that the principle is indispensable to practical reason.

Yet, the scope of Railton's argument is more ambitious than this. He thinks that the indispensability of the principle grounds *instrumentalism* about practical reason, or the view that all rules and reasons concerned with actions are instrumental. There are traditionally three basic principles of practical reason – the instrumental principle governing the relation between means and ends, moral principles governing our conduct towards others, and the principle of prudence which governs the harmonious relation between ends. On Railton's view, moral and prudential considerations are wholly relativised to an agent's ends, which are conative. If, for instance, you desire to cause suffering regardless of the consequences for yourself or others, then no considerations about human dignity or self-preservation can get a rational grip on you.

This is certainly an option, but it only works if one views ends as necessarily conative. An

alternative view is presented by Korsgaard, who argues for a distinction between desires and ends.[50] A desire is a conative inclination, whereas an end is a desire which one is committed to pursuing based on reasons. She thinks that ends must have this cognitive aspect, because one can only have a reason to take the means to one's end if one has a reason to fulfil this end. She goes on to argue that the idea of a reason as wholly private is incoherent, and only makes sense if we view it as a public phenomenon. This means that an agent cannot have an end unless he is able to explain the reason for having this end to others. Notice that in virtue of its other-directed nature, a rule for universalisability has moral overtones. So, having an end necessarily invokes a rule for universalisability; at the same time, however, having an end invokes the instrumental principle. On Korsgaard's account then, the instrumental principle and a rule for universalisability are equally indispensable. Invoking the one invokes the other.

Hence, the indispensability of the instrumental principle does not ground instrumentalism about practical reason, unless one views ends as wholly conative. This is because the logical structure of the principle, which is responsible for its indispensable status, does not obviously entail conativism about ends. Without an argument for conativism about ends, the principle's indispensable status does not rule out the possibility that there are other rules of practical reason which are as fundamental and indispensable. This account of the principle is therefore neutral between competing views of practical reason as to whether there is scope for the rational criticism of ends, and what form this criticism should take – instrumental, moral or prudential.[51] The idea that there is at least one structural requirement on our reasons for acting can therefore function as common ground amongst significantly different views about the nature and scope of practical reason. As I consider Railton's argument to be persuasive as it stands. I will not examine his argument further. Instead, I want to assess the implications of this indispensability for explanations of behaviour as rational action.

[50]Christine Korsgaard, 'The Normativity of Instrumental Reason', in *Ethics and Practical Reason.*

[51] Roughly speaking, the instrumental view is neo-Humean; the moral view is Kantian; the prudential view is aretaic. I do not discuss an aretaic conception of practical reason. For an example of this, one can look to TH Irwin, who has an

4. Indispensability and Normativity

This section examines the relation between the principle's indispensability and its normativity. I argue that in virtue of its indispensable status, rejection of the normativity of the principle leads to scepticism about practical reason. The normativity of the instrumental principle is therefore a necessary condition for belief in practical reason. Hence, in order to explain an agent's behaviour as rational action, the explanation must invoke a normative instrumental principle. I go on to argue that this normativity does not require the attitudes of praise and blame characteristic to moral judgements.

A good place to start is by getting clearer on the difference between a normative conception of the instrumental principle, and a positive one. The normative conception sees the principle as stipulating a good fit between means and ends, such that an agent's reason for acting should take a particular shape – it should have an end which is connected to a means by the possession of true belief. As such, it is always possible that an agent's behaviour could fail to conform wholly to the principle, and her reason for acting would be worse rather than better. She could fall short of according her end sufficient deliberative weight in two ways. At a basic level, she could fail to consider taking the means to her end, and at a more sophisticated level, she could have mistaken beliefs.

A positive treatment of the principle, on the other hand, does not approach the principle as a concept which outstrips the way agents behave; it only sees the principle as capturing the way that agents invariably do behave. On this view, it is simply a fact that agents are reliably motivated to take the means to their ends, and the principle denotes this constant conjunction of events, rather than stipulating a good fit between means and ends. Consequently, it cannot be used to distinguish good reasons for acting from bad ones. This is because an agent who fails to act on a good reason

---

aretaic account of principles of practical reason as unified by the ultimate human end of flourishing. See Irwin,

fails to meet a criterion for what counts as a good reason. If the principle is to cleave the good reasons from the bad, it must be possible for it to outstrip an agent's existing behaviour. The positive treatment of the principle makes this impossible.

This has a bearing on whether we count behaviour as rational. As I understand it, in order for an agent's behaviour to be capable of counting as rational, we must be able to see an agent as acting on reasons, because these reasons make the action intelligible to us. A reason does so in virtue of working as a favouring relation – it shows us how a certain combination of beliefs and ends can favour a particular action and thereby helps us to understand why the action was performed. This means that it can weigh in favour of an action more or less, depending on the truth-value of the beliefs involved. If the reason for action is based on a false belief, then although the reason favours the action, it does not favour it as much as a corresponding reason based on true beliefs. So the favouring relation is a normative one, in the sense that it can be better or worse with respect to true beliefs. Consequently, a reason must be capable of being better or worse in order for it to act as a favouring relation, and therefore, in order for it to explain action.

So the distinction between good and bad reasons is necessary for us to see an action as rational. There may be additional requirements on rational behaviour, but I think the notion of a practical reason, as something that renders behaviour intelligible by showing that a consideration weighs in favour of an agent pursuing an action, is central to our understanding of rationality. I do not know how we can understand behaviour as rational if there is no way in which can explain it by referring to this favouring relation.

Could some other rule of practical reason function as a criterion for distinguishing good from bad reasons, rational from irrational action? I do not think so. The instrumental principle is indispensable to all other rules of practical reason, such that no other rule works unless the normativity of the principle is in place. Consider Korsgaard's rule for universalisability. This rule stipulates that one ought to have a reason for one's end. As Korsgaard recognises, however, one can

---

'Practical Reason Divided', in *Ethics and Practical Reason*.

only have an end if one aims to realise it (I discuss this constraint on ends in further detail later on). This means that one can only have a reason for one's end – a rule for universalisability – if one has a reason to take the means to one's end – the instrumental principle. So the rule for universalisability can only get a grip on us if the principle has normative force for us.

On a more prosaic note, a rule which stipulates that good citizens ought to pay their taxes could only be justified by appealing to some end that individuals had in mind. One could say, 'You ought to pay your taxes because if you don't, you will be thrown in jail", or "If nobody pays her taxes, then society will fall apart and you don't want that." It is only on the assumption that individuals prefer to stay out of jail or that they prefer to have a well-ordered society, and on the assumption that individuals ought to give deliberative weight to their ends, that a tax-paying rule has normative force. The normativity of the instrumental principle therefore plays a crucial and irreplaceable role in the normativity of other rules of practical rationality. If the principle is not normative, then no other practical rule is.

This implies that no rule of practical reason can be used to distinguish good reasons for acting from bad ones, rational from irrational behaviour. One could use a theoretical rule, like the rule for true beliefs, but this would only apply to the epistemic aspect of a reason, not its action-guiding function. As such, it could make sense to talk about reasons for believing, which could be better or worse, and one's beliefs could therefore be more or less rational. Nevertheless, it would not make sense to talk about reasons for acting, because one could never act in a rational or irrational manner, although one could certainly believe rationally or irrationally. On this view then, practical reason is something of a misnomer. The appearance of genuinely rational action is just that – an appearance; scratch the surface and one sees theoretical reasons at work, but no practical ones.[52]

Hence, rejection of the normativity of the instrumental principle implies scepticism about practical reason. Conversely, the normativity of the principle is a necessary condition for holding

[52]The sceptical position I have in mind here is sometimes attributed to Hume. See, for instance, Thomas Nagel's description of Humean scepticism about practical reason in *The Last Word* (Oxford: Oxford University Press, 1997): 102.

that practical reason exists.[53] This means that an explanation of an agent's behaviour as rational must invoke a normative instrumental principle.

A corollary of this conclusion is that the instrumental principle governs the shape of both normative reasons, and the motivating reasons which explain agents' actions. This sounds strange, given that motivating reasons were earlier contrasted with normative ones, so that it was natural to infer that motivating reasons were descriptive. I think this is the wrong way of comparing the two kinds of reasons, and that a better way of comparing them is in terms of the strength of normativity. For both normative and motivating reasons, the principle stipulates that one ought to give deliberative weight to one's ends and it thereby governs the shape of any reason to act. In the case of normative reasons, however, we take the requirement to be stronger, because it is important to us that an agent give correct or *good* deliberative weight to her ends by possessing true beliefs.

In order to understand the minimal normativity of a motivating reason, we can consider James Skidmore's account of an agent who suffers from a failure of instrumental reasoning.[54] When asked what he plans to do this weekend, Castro replies that he wants to go swimming:

> *Questioner*: What a lovely idea. Are you going for a dip in the dam then?
>
> *Castro*: I don't know, but what difference does that make?
>
> *Questioner*: Well, if you want to go swimming, you'll need to get wet.
>
> *Castro*: Yes, I know. But I don't care to get into the water.
>
> *Questioner*: Oh, so you'll go to the beach then I suppose?
>
> *Castro*: No, as I said, I don't care to get into water, whether it's the sea, the dam or the pool.

[53] It is not obvious that the normativity of the instrumental principle is sufficient for holding that practical reason exists. If one believes that reasons for acting and certain kinds of ought statements are distinct, then it is possible that the instrumental principle is normative in the sense that it stipulates 'one ought to accord deliberative weight to one's end'. This would not imply that doing so, however, was rational, in the sense that one acted on a good reason. Following the instrumental principle may therefore bring you to do things which you have no good reason to do. For an example of how ought statements and practical reasons can come apart, see John Broome, 'Reasons', in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by Jay Wallace, Michael Smith, Samuel Scheffler and Philip Pettit(Oxford: Oxford University Press, 2004): 28–55.

[54] James Skidmore, 'Skepticism about Practical Reason: Transcendental Arguments and Their Limits', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 109, 2 (2002): 121-141. The following example is a variation on Skidmore's example of a person who wants to go to the beach.

Questioner: But you said you'd like to go swimming.

Castro: I do. It's perfect weather for it.

Questioner: And the only way you can do so is by getting into water?

Castro: Yes, I know, but what's the point. I'm not getting wet. Ah, but how glorious it would be to spend the weekend swimming...

What are we to make of such a person? As Skidmore points out, if Castro is utterly indifferent to the means he knows are necessary to go swimming, we cannot interpret Castro as genuinely intending to swim. We might understand him as wishing he could go swimming, or reflecting fondly on the prospect of swimming, but we cannot understand him as having the end of swimming, unless this end plays a role in his deliberation about how to act. If the case was different, however, and Castro suffered from a global or total failure of instrumental reasoning, then he could not be said to have any ends whatsoever. This kind of person recognises that an action will secure the fulfilment of an outcome that he desires and yet can never understand why he should pursue it. His behaviour is torn asunder from the loves, loyalties and desires he harbours, and floats free, with no goals to guide it. This means that his behaviour cannot be classified as intentional, and consequently, cannot be classified as an action.

This is what would happen if our explanatory attitude towards agents' actions did not assume that their motivating reasons were governed by the instrumental principle – we would not be able to understand them as acting to fulfil their ends. So the motivating reasons we use to explain agents' behaviour already contain a minimal normativity to them, by articulating our expectation that agents give deliberative weight to their ends in reasoning about how to act. The normativity lies in the way in which we expect their reasons for acting to conform to the shape stipulated by the principle.

The qualification 'minimal' is important, because an explainer who draws on a motivating

.

reason to explain an agent's actions need not endorse his actions. Castro drinks paraffin because he is thirsty and believes that the glass contains water. The explainer need not agree that Castro ought to have acted as he did, as his reason for acting was based on a false belief. Nevertheless, the explanation works by showing that Castro's action was, from his perspective, proper, because he understood his action as corresponding with his end of relieving thirst. Alternatively, Castro drinks paraffin in order to turn himself into a human fire-bomb. The explainer may condemn this action as morally reprehensible and imprudent, but nevertheless affirm that, given his intention, it is appropriate for him to drink paraffin.[55]

We can return to the notion of a requirement, as this offers us an insightful way of thinking about minimal normativity. Broome distinguishes between two kinds of normativity: a 'requirement' and an 'ought'.[56] An ought governs a proposition. For instance, if *p* is true then you ought to see to it that *q*: if a child is about to walk into oncoming traffic and die, then you ought to warn the child not to do so. So the normativity in this statement attaches to the contents of the consequent. If we put it in more formal terms, the normativity would look like this: p→Oq, where the 'O' stands for 'you ought to see to it that'. In contrast, a requirement governs the relation between propositions. For instance, you ought to see to it that if *p* is true then *q*: you ought to see to it that if you have a desire to turn into a human fire-bomb, then drinking paraffin is the means to doing so. The normativity in this statement attaches to the relation between means and ends, and the requirement here is a version of the instrumental principle. In formal terms, it looks like this: O(p→q). A requirement is therefore different in scope and logical structure to an ought. A requirement is wide-scope and applies to the relation between means and ends, while an ought is narrow scope and applies to the content of a statement.

This means that you could satisfy a requirement, like the instrumental principle, in one of two ways: either you could consider taking the means to your end or you could relinquish the end.

[55] Depending on one's theory of practical reason, this might also be the case for normative reasons. If, like Korsgaard, one thinks normative reasons must also fulfill moral and/or prudential conditions, then one cannot have a normative reason which contravenes moral and prudential principles.

An ought, on the other hand, does not allow for this kind of revision, because it does not govern the relation between premises. There is only one way you can satisfy an ought, and that is to ensure compliance with the content of the statement. The logical structure of a requirement therefore renders it less stringent than an ought. This is why an explainer can consistently affirm that Castro is required to drink paraffin if he wants to turn himself into a fire-bomb, and yet reject Castro's action as morally repugnant. She affirms that the relation is appropriate relative to his end, but denies that Castro ought to have the end he does.[57]

As such, although the shape of a motivating reason is governed by a normative instrumental principle, this normativity does not necessarily consist in the kind of endorsement and censure characteristic to moral judgements. In the first chapter I pointed to several kinds of normative concepts which did not fit in the dichotomy between moral and descriptive properties that Jevons and Robbins seem to propose. I think that the instrumental principle is a plain, persuasive example of the way in which normative properties are not synonymous with moral ones. I have argued in this section that commitment to the view that an agent's behaviour counts as rational action requires commitment to the view that it is governed by a normative instrumental principle, but this normativity is not necessarily moral, nor is it prudential.

5. Indispensability and Non-Hypotheticality

In the last section I argued that normativity is an ineliminable component of rationality, and that in the case of practical reason this normativity must take the form of the instrumental principle, because the principle is indispensable to practical reason. In this section I argue that we should take this indispensable status to be a kind of non-hypotheticality.

Recall the conditions for consistency in decision theory. An agent should not, for instance,

---

[56] John Broome, 'Normative Requirements': 79-83.

[57] This also offers another way of explaining the indispensable role of the principle in the normativity of other rules of practical reason. The normativity of a moral ought, for instance, requires the normativity which the principle attaches to

have circular preferences. Notice that an agent has preferences prior to, and independent of, this criterion. So the conditions for consistency are only applied *after* preferences have been formed as a test of their rationality, and are therefore external to the practice of reasoning. In contrast, the instrumental principle has a substantially different function to the conditions for consistency in decision theory. The Tortoise argument demonstrates that the principle governs the shape of practical reasons, and is therefore indispensable to practical reason, by showing that a reason to act must aim to conform to the principle in order for it to count as a reason. This means that we cannot act on reasons and then subsequently apply the principle as a test of the rationality of our actions. Unlike the conditions for consistency, the principle is not an abstract standard which exists independent of practical reasons, and is then manifest in our actions. Instead, the principle is internal to the practice of reasoning, as it governs the form of reasons themselves. As Korsgaard puts it, the instrumental principle is 'essentially the principle of application itself', for a practical reason which does not already aspire to conform to the principle is no reason at all.[58]

It follows that the principle is not dependent upon the particular ends of an agent, but is instead dependent on the possession of ends in general, for part of what is involved in having an end is aiming to give that end deliberative weight in one's reasoning. As Railton points out, this cannot be a linguistic point about the meaning of having an end, for if it was, the principle would be trivial and not rationally compelling.[59] Rather, it is a point about the structure of our reasoning. When an agent reasons about how to act, he can sometimes fail to give an end sufficient deliberative weight, but he cannot fail to *aim* at giving his ends deliberative weight. As such, the principle does not depend on having a particular end, like the end of following the principle; likewise, the absence of any particular end is irrelevant to whether one follows the principle. Rather, the possession of an end in general comes with a simultaneous and inextricable commitment to the instrumental principle. Importantly, it does not follow from this that it is impossible to flout the principle, but if

---

the relationship between means and ends in order for it to make sense.

[58] Korsgaard, 'The Normativity of the Instrumental Principle': 241.

[59] Railton actually makes this point about constitutive arguments, and not the nature of an end, but the principle is the

an individual does stop following it entirely, he has ceased to reason practically and therefore ceased to act intentionally. In this case, indispensability is a kind of non-hypotheticality.

Approaching the instrumental principle as a non-hypothetical rule helps us understand the explanatory process a little better. In explaining an agent's behaviour as an action, as conduct directed towards her end, an explainer implicitly appeals to a shared structure of reasoning about how to act. This invocation of the principle, however, is an inherent part of the explanation, and does not depend for its application on the nature of the ends being examined. Unlike conditions for consistency, the principle must be in place for one to view an agent's ends as ends. We can therefore only grasp an agent's behaviour as an instance of rational action by appealing to a feature which we take to be common to all agents. We can say that such an explanation only gets its sense from the community of agents in which the explainer is embedded.

Conclusion

I have argued that the instrumental principle governs the shape of reasons for action, and this explains why it is indispensable to practical reason. In virtue of its indispensable status, commitment to the existence of practical reason implies commitment to the view that the instrumental principle is normative. Conversely, rejection of the normativity of the principle implies scepticism about practical reason. In order to explain an agent's behaviour in terms of rational action, an explainer must therefore assume that the behaviour is governed by a normative instrumental principle. Denial of the normativity of the principle is inconsistent with an explanation of an agent's behaviour in terms of rational action. Furthermore, I have argued that the indispensable status of the principle is a kind of non-hypotheticality.

It follows that when marginal utility is employed in explanations of human behaviour as rational action the concept is governed by the instrumental principle, and therefore has a normative

---

same. See his 'On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action': 69-72.

and non-hypothetical component. It is normative, because it is used to articulate the explainer's expectation that an agent acts to realise her ends; this means that she has a reason for acting, where this reason aims to conform to the structure stipulated by the instrumental principle. It is non-hypothetical, because the structure to which this reason ought to conform is independent of any particular end she might have.

It is not yet clear whether marginal utility is, or must be used to explain agents' behaviour as *rational action*. I address this question in the next chapter. For now though, we can ask what happens to marginal utility when it is used in this way. Although it shares normative and non-hypothetical characteristics with classical utility, it is distinct from classical utility, for it is neither a moral concept, nor tied to a specific universal end, like pleasure or flourishing. So, an agent can have a preference which she knows to detract from her wellbeing or that of others; she can, to use Hume's memorable phrase, prefer the destruction of the world to the scratching of her finger. The good which satisfies this preference for destruction nevertheless has greater marginal utility for her than finger-scratching. This does not mean, however, that any of us should endorse this action as moral or prudent, or that we should seek to mimic her behaviour. So, even if marginal utility were normative and non-hypothetical, it could not be used as classical utility is; by itself, it could not legitimately function as a decision procedure for how to act in terms of one's own wellbeing or that of others.

Instead, this use of marginal utility would be expressive of a shared understanding of each other as intentional, reasoning beings. In explaining that Xoli chooses coffee over tea because it has greater utility for her, we are invited to make a judgement about how we expect any agent to act in that situation, given this preference ordering. We expect that any agent will act to fulfil her ends, because this is what it means for any of us to act, and we expect that drinking coffee corresponds with some preference, because our actions ordinarily aim to correspond with our ends. So, when marginal utility is used in an explanation of an agent's rational action, the explanation already

contains a positive evaluation within it, an evaluation which makes this kind of explanation possible in the first place. As such, the purpose of this chapter has not been to argue for a different use of marginal utility, but to reflect on the existing use of the concept as inextricably interwoven with communal judgements about what counts as rational action.

CHAPTER 3

The Axiomatic Challenge: A Closer Look at Economic Explanations

Thus far, I have argued that one rule of practical reason, the instrumental principle, is indispensable to practical reason, and is both normative and non-hypothetical. This means that an explanation of an agent's behaviour in terms of marginal utility can only work in one of two ways: (i) either it explains an agent's behaviour as rational action and invokes the instrumental principle, which is normative and non-hypothetical, or (ii) it rejects the instrumental principle and cannot explain an agent's behaviour as rational action. Call the first kind an intentional explanation, and the second kind a non-intentional explanation.[60]

This characterisation of explanations does not focus on the way in which agents do in fact behave; instead, it draws out the differences between the explanatory strategies available to an explainer. The distinction works by showing that if an explainer rejects the normativity of the instrumental principle, then she is forced to construct a non-intentional explanation of an agent's behaviour. This does not mean that intentional and non-intentional explanations are in competition with each other, or that one cannot have an explanation which includes both non-intentional and intentional elements. I use Nancy Cartwright's account of what she calls *socio-economic machines* as an example of how an explanation can have a mix of intentional and non-intentional elements. I argue that it is nevertheless useful to classify this kind of mixed explanation as intentional, because the instrumental principle has a crucial function in the explanation.

I then go on to defend the relevance of this distinction for axiomatic utility theory. The first

[60] I have borrowed this terminology from Daniel Dennett, but it is not meant to be identical to Dennett's sense. As I use it, an intentional explanation does not refer to all intentional states, like consciousness, and nor does it imply any view of the nature of consciousness (it does not matter whether consciousness is behaviouristic, hidden inside an agent's head or simply invented). It only refers to behaviour that is treated as an instance of rational action. I have not used the more traditional term 'rationalising explanation', because a rationalising explanation provides an agent's primary reason for acting, where this reason has played a causal role in her coming to act. My account of explanation is neutral as to whether an agent actually acts on reasons, so it would misrepresent the account to frame it in terms of rationalising explanations. For an authoritative presentation of rationalising explanations, see Donald Davidson, 'Actions, Reasons

two chapters focused on the concept of marginal utility, which was framed by its original authors in terms of means and ends, and was therefore useful in exploring the role of the instrumental principle. The concept of marginal utility, however, has evolved to denote preferences or ends which satisfy certain axioms, rather than the relation between means and ends. For the sake of clarity, I call this *axiomatic utility*. As it collapses the distinction between means and ends upon which the instrumental principle relies, axiomatic utility poses a substantial challenge to classifying explanations in intentional/non-intentional terms. The purpose of this chapter is to consider whether the instrumental principle is relevant to understanding explanations which employ axiomatic utility. I argue that it is, because axiomatic utility is elliptical for marginal utility. I go on to make a case for holding that the instrumental principle is a necessary feature of an explanation which employs marginal utility.

1. Socio-Economic Machines and Mixed Explanations

There are two poles in theorising about the way in which economic explanations work. At the one extreme is what Cartwright calls the 'regularity theorist', who thinks a scientific law is a statement about a regular association, and that a scientific explanation therefore involves showing how a particular event falls under a law, how it instantiates some regular association.[61] As Cartwright correctly points out, this poses a substantial challenge to economics' claim to being a legitimate science. This is because economic accounts only hold *ceteris paribus*. They only obtain if there are no interferences or disturbing causes. For example, models based on utility theory characteristically assume that commodities are infinitely divisible, that everyone has near-perfect information, that there are no transaction costs, or that environmental conditions are static. Such claims are often exaggerations or false. So economists need to keep working on their accounts until they express

---

and Causes', in *Essays on Actions and Events*(Oxford: Clarendon Press, 1980).

[61] The following discussion is drawn primarily from Cartwright's '*Ceteris Paribus* Laws and Socio-Economic Machines' in *The Dappled World: A Study of the Boundaries of Science* (Cambridge: Cambridge University Press,

true regularities. Without these true regularities, economics is at best poor science; at worst, it is pseudo-science.

Cartwright argues that there is a more useful way of thinking about economic accounts as a kind of socio-economic machine, in which no mechanism can work alone. She asks us to consider a very simple economic machine – the demand-supply curve basic to an equilibrium analysis. It acts like a seesaw: if you increase demand and keep supply constant, then the price of a good will rise. If you increase the supply and keep the demand constant, then the price of a good will fall. If supply and demand are perfectly balanced, then we have a situation of partial equilibrium.[62]

We can conceive of the demand mechanism in terms of an individual or an institution. In either case though, we think of the attendant behaviour as preference-guided and constrained by certain requirements on the structure of a preference, because this is essentially what demand is – a preference for a particular outcome in virtue of which an agent chooses one good over another. The demand mechanism is intentional and invokes the instrumental principle, but it cannot generate any regular behaviour on its own. It must work together with a supply mechanism, which need not have any intentional properties whatsoever. The supply of wood in a forest, for instance, can be explained in terms of rainfall patterns, the number of seed-dispersing birds and the rate of growth of new trees to the loss of old ones. For the demand-supply seesaw to work, it must be affixed to a fulcrum which connects both the supply and the demand mechanisms, and it must also be set in a stable environment in which it is not jiggled about. Rapid and unpredictable changes in the weather, monopolies on wood production, soil erosion from rapid deforestation[63] and the presence of persistent false beliefs are all factors which would knock the seesaw off balance. On Cartwright's account, economic mechanisms are like machine parts: they must be assembled and set running in a stable environment before any regular behaviour results.

1999).

[62] A partial equilibrium analysis only looks at a part of an economy at one time and assumes that the rest stays constant. A general equilibrium analysis attempts a more holistic picture, but still assumes *ceteris paribus* clauses, like psychological egoism, consistency of preferences and so on. See, for instance, Amartya Sen, 'Rational Fools: A Critique of the Behavioural Foundations of Economic Theory', *Philosophy and Public Affairs* 6, 4(1977): 317-344.

[63] Soil erosion is an example of 'externalities' – the unintended consequences of a particular action which can interfere

A regularity theorist would not be able to make much sense of the law of supply and demand. The law is too stylised – it does not represent pixel for pixel each regularity occurring in a socio-economic setting. Yet if we follow Mill, this abstraction may be a strength, not a weakness. He argues that there are an extraordinary number of potential factors involved in any socio-economic event, so it is difficult to observe regularities directly.[64] In his view, one first needs to posit a set of causal laws operating in a restricted domain, and then investigate their combined consequences; by doing so, one can test the validity of one's prior assumptions. Similar to the way in which a Kandinsky draws our attention to the complex interplay of emotional colour and movement in a symphony by abstracting away the instruments from the painting, an economic explanation can draw our attention to fundamental economic forces by abstracting away those factors that detract from our understanding of a situation by complicating it beyond our comprehension. In this respect, Cartwright is clearly in agreement with Mill, for she holds that an economic explanation is able to generate the generalisations with which we make social events intelligible in virtue of *ceteris paribus* clauses, not in spite of them.[65]

While I am sympathetic to Cartwright's account, because it is helpful in making sense of economic explanations, the purpose of this section is not to adjudicate between the two competing views. Instead, her account serves as an example of the way in which an explanation can plausibly have a mix of intentional and non-intentional elements. Nevertheless, in a socio-economic machine, like a demand-supply curve, all of the mechanisms need to be in place for the machine to work. So we can still characterise an explanation that draws on the instrumental principle as an intentional explanation, because the principle makes the explanation possible. If you want a door to work, there must be a hinge; it you want to use marginal utility to generate an explanation with intentional elements, the instrumental principle must be in place. As such, although the distinction between

---

with the successful running of the socio-economic machine.

[64] John Stuart Mill, 'On the Definition of Political Economy and the Method of Investigation Proper to It', in *Collected Works of John Stuart Mill*, vol. 4(Toronto: University of Toronto Press, 1967).

[65] This defence of *ceteris paribus* clauses is a methodological justification, not a justification of their content. So we can still interrogate the content of these clauses; we can, for instance, question whether it is legitimate to presume that the environment is static or that agents always have true beliefs.

intentional and non-intentional explanations is perhaps too strongly drawn, it serves to draw out the role of the instrumental principle in explanations that invoke marginal utility, even if these explanations also rely on mechanisms that have nothing to with agents or rational action.

2. An Axiomatic Challenge

Marginal utility was originally framed in terms of the relationship between means and ends, so that one could infer the ends towards which an agent acted from the means that she chose. This is no longer the case, however, for axiomatic utility theory has developed the concept into what one can call *axiomatic utility*.[66] Axiomatic utility differs from marginal utility in one important respect: provided that an agent's preferences conform to certain axioms, utility represents an agent's preferences, rather than the marginal capacity of a good to satisfy an agent's preferences.[67] Given a choice between two options, *A* and *B*, *A* has greater axiomatic utility than *B* for an agent if she chooses *A* over *B*. In cases where an agent is not in a situation to choose between alternatives – for instance, between becoming president and winning the lottery – we can ask an agent which options she would choose if given the opportunity. The choice for *A* is said to be a revealed preference for *A*, while a preference is considered to be a hypothetical choice. So, while a preference is not strictly identical to a choice, it is roughly equitable, for the only difference between the two is whether one is faced with present options or hypothetical ones. Axiomatic utility theory therefore equates choices with preferences, means with ends.

The explanation of unemployment as a voluntary choice is one such use of axiomatic utility. In an economic slump, the demand for labour decreases, because companies receive less revenue. By standard microeconomic theory, wages will go down until the excess supply of labour –

[66] Characterising this change as an explicit, conscious development may be erring on the side of generosity. Theorists like Sen first had to argue that there was an ambiguous use of the term 'preference' to denote both an agent's choices and the motives underlying that choice before the two senses of utility were explicitly separated. See his `Behavior and the Concept of Preference', *Economica*, 41(1973): 241-259.

[67] This characterisation of axiomatic utility theory is taken from John Broome, 'Utility', in *Ethics out of Economics.* A fuller formulation of the theory can be found in John Hicks and R.G.D. Allen, 'A Reconsideration of the Theory of

unemployed people – is eliminated. If this is correct, then there should never be excess labour or unemployment – there should always be an equilibrium between the demand and the supply of labour. Yet the phenomenon of unemployment is real. Theorists in the 'rational expectation' school argue that the excess supply of labour is due to the preference of unemployed people to remain unemployed rather than accept lower wages. This is because they conceive of individuals as faced with two options: remain unemployed or work for the wage that will bring the labour market into equilibrium. If individuals do not choose to work for a lower wage, then they prefer unemployment to low-paying employment. Robert Lucas explains unemployment as a voluntary phenomenon in this way:

> the unemployed worker at any time can always find *some* job at once, and a firm can always fill a vacancy instantaneously. That neither typically does so *by choice* is not difficult to understand given the quality of the jobs and the employees that are easiest to find. Thus there is an involuntary element in *all* unemployment, in the sense that no one chooses bad lack over good; there is also a voluntary element in all unemployment, in the sense that however miserable one's current work options, one can always choose to accept them. [68]

This explanation of unemployment as voluntary suggests that government intervention (in the form of tighter labour control, public works projects and social grants) will not work at all, because it will not change the fact that some individuals prefer not working to the low wages they will receive. Irrespective of whether one agrees with this diagnosis (and there are many who do not),[69] it is clear from this example that the concept of axiomatic utility has a definite explanatory role with potentially significant implications for how we understand and respond to each other.

Value. Part 1', in *Economics* 1(1934): 52-76.

[68] Robert Lucas, 'Unemployment Policy', *American Economic Review* 68(1978): 354.

[69] One of the cornerstones of Marxism, for instance, is the coercive nature of labour relations in a free market. For a less

The equation of choices with preferences also has theoretical significance for understanding the nature of explanations. By eroding the distinction between means and ends, axiomatic utility removes the instrumental principle from the picture. The principle stipulates that agents ought to give deliberative weight to their ends so that they are moved to consider taking the means to their ends; it therefore relies on a distinction between means and ends. Yet economic explanations typically feature axiomatic utility, and not marginal utility. Categorising economic explanations in intentional/non-intentional terms is therefore inappropriate, because doing so draws out features of explanations which are irrelevant to an economic explanation – namely, the relation between means and ends in an agent's deliberation about how to act, a relation which the instrumental principle governs.

Instead of contrasting intentional with non-intentional explanations, the axiomatic challenge contends, one should differentiate between economic and non-economic explanations. An economic explanation applies to any features of a system which can be explained using economic concepts, like marginal and axiomatic utility. Any features which cannot be explained using these concepts will have a non-economic explanation. As Ross argues, this distinction is important to understand the scope and aims of economics. This is especially the case for utility theory, because it is neutral with regards to psychology and ethics, and can therefore be applied to any aspect of behaviour that is preference-guided: love, sex, war, education – the list is as long as there are activities. On this wide and expansive definition of utility, he writes:

> What [the definition] says is that the whole of human life has an economic *aspect*; but it does not imply a *reduction* of *all* aspects of human life to the economic dimension.[70]

---

ideological account, one can look to Arthur Okun's *Equality and Efficiency: The Big Trade Off* (Washington: Brookings, 1975).

[70] *Ross, What People Want*: 30-31.

If Ross is right, then an understanding of the limits of economic enquiry gives one an understanding of how economics works. This understanding requires one to draw a distinction in kind between economic and non-economic explanations. We can say the distinction has a fundamental methodological role in economics.

In this respect, consider Gary Becker's application of utility theory to the phenomenon of marriage.[71] He begins by approaching marriage behaviour as an individual's decision to enter the marriage market preceded by the search for an appropriate mate. Once in the marriage market, a person searches for a mate until the marginal utility of any expected improvement in the mate she can find is not greater than the cost of her time and opportunity costs to procreate. This explains why women tend to marry younger than men, because women have a shorter time-span in which to produce what Becker charmingly calls 'units of output' or children, so their opportunity costs are greater. Similarly, individuals will marry at a younger age when life expectancy is lower, because later marriages increase the probability that they will not bear children.

His explanation therefore conceives of marriage behaviour as a choice between various options which maps onto a utility function. It does not make reference to the biological features of attraction, or the social norms governing marriage, the class politics informing marriage choices, or the more shaky convictions of some that our love-life is determined by the alignment of the stars and planets. These are features of biological, sociological, political and astrological explanations, not economic explanations. Yet whatever it is that distinguishes these non-economic explanations from Becker's, it is *not* the instrumental principle, for the principle can belong to sociological and political explanations, without appearing in astrological or biological explanations. For instance, one can explain an individual's choice to marry as intended towards complying with a norm for marriage; in this sociological explanation, her choice is instrumental to an end of hers, and it is therefore governed by the instrumental principle. On the other hand, one can explain an individual's attraction to another as an outcome of biological features like genetic compatibility and physical

[71] Gary Becker, 'A Theory of Marriage: Part 2', *Journal of Political Economy* 82, 2(1974): 11-26.

indices of fertility. In this explanation, it is irrelevant whether the individual in question is an agent acting to fulfil his ends, or a collection of biological forces. So the principle does not play a role in differentiating between economic and non-economic explanations. Moreover, even within the domain of economics, its role varies: it might be a feature of marginal explanations, but it does not seem to be a feature of axiomatic explanations. Consequently, the instrumental principle may be indispensable to practical reason, but this is irrelevant to economists who are more concerned with how economic explanations work.

The axiomatic challenge therefore presents two objections against applying the categories of intentional and non-intentional explanations to axiomatic explanations. Both are structured in terms of the instrumental principle. The first objection is that axiomatic utility does not discriminate between means and ends. As the instrumental principle relies on the difference between means and ends, it is irrelevant. Call this the spectre of behaviourism. The second objection is that the distinction between economic and non-economic explanations is crucial to understanding how economic explanations work, but this difference cannot be picked out with the aid of the instrumental principle. Call this the republican's objection. The thrust of each objection is that the categories of intentional versus non-intentional explanations are inapplicable. I consider each objection in turn.

3. The Spectre of Behaviourism

The equation of preferences with choices has its roots in behaviourism, a theory that views mental states as a pattern of behaviour, rather than properties hidden inside an agent's head.[72] It responds to the lacuna in early neoclassical work on marginal utility, which left it open as to how one should go about identifying agents' preferences. By defining choice as revealed preference, and preference as

[72] This view of axiomatic utility is very close to *revealed preference theory*. Paul Samuelson, the pioneer of revealed preference theory, initially sought to remove any reference to utility, because his theory collapses the distinction between choice and preference, means and end. See Samuelson, 'A Note on the Pure Theory of Consumer Behaviour',

hypothetical choice, it gives contemporary economists a powerful tool with which to measure agents' preference structures, because economists can use actual or hypothetical choice behaviour to evaluate preferences. The behaviourist's objection is this: in an explanation that employs axiomatic utility, the distinction between means and ends falls by the wayside, and with it, the role of the instrumental principle. Call this an *axiomatic explanation*. It differs from an explanation that employs marginal utility, which distinguishes between means and ends. Call this a *marginal explanation*.

In this section, I argue that an explainer is only able to equate choices with preferences, means with ends, by first making a distinction in kind between means and ends. So, an axiomatic explanation is a convenient simplification of the explanatory process. This does not undercut behaviourism, but rather quells one objection to classifying axiomatic explanations in intentional/non-intentional terms, by showing that these explanations rely on a distinction between means and ends.

We can begin by asking how a behaviouristic equation of preferences with choices functions in an explanation, by considering the application of axiomatic utility to the problem of unemployment. Suppose there is an agent, Castro, who was employed as a construction worker, and earned a wage of R2000 per month. When the recession hit, however, he lost his job and has since been unemployed. In line with the rational expectation school, we can conceive of Castro as faced with two options: remain unemployed or work for the wage that will bring the labour market into equilibrium, at say R1500 per month.

We can begin by asking for the criterion by which unemployment is identified with a preference for unemployment in particular. There are, after all, a myriad number of outcomes that a choice might aim towards. Say Castro has a preference for working over being unemployed, and a preference for caring for his children over working. In this case, his choice to stay unemployed would arise out of a preference for looking after his children, not a preference for being

---

in *Economica New Series*, 5, 17(1938): 61-71.

unemployed. On the face of it, I think axiomatic utility theory tries to bar this option by appealing to a linguistic criterion. Just as a 'bachelor' means 'unmarried man', choices are defined in such a way that they are equated with preferences. So the same term 'unemployment' occurs in the statements 'Castro chooses unemployment' and 'Castro prefers unemployment', and this means that Castro's choice reveals a preference for unemployment rather than anything else.

It is not clear, though, that this parallelism acts as a criterion for interpreting Castro's choice with a preference for unemployment in particular. If we follow Kevin Lancaster and Gary Becker,[73] this linguistic criterion is at once too strong and too weak to secure a definite preference. For Becker and Lancaster, the choice of unemployment reveals a preference for some of the attributes of unemployment, rather than unemployment itself. Castro might choose unemployment in order to have more leisure time, or to avoid harassment from his xenophobic co-workers in virtue of his foreign nationality, or to get unemployment benefits. If Castro is unemployed in order to stay in bed longer in the mornings, however, then finding a job as a night watchman could satisfy his preference just as well. The point is that a variety of actions can satisfy a single preference, while a single action can have a variety of preferences that it satisfies. Consequently, the linguistic criterion is too strong, because it conceals the preferences that could be at work (that is, preferences for the attributes of a good, rather than the good itself). At the same time, the criterion is too weak to help an explainer decide that choosing unemployment reveals a preference for unemployment in particular, rather than something else like leisure time or social grants.

Nevertheless, a friend of axiomatic utility could respond that choosing unemployment reveals something about goods associated with unemployment. Moreover, the above objections misconstrue axiomatic explanations. These explanations do not assess the once-off consumption of goods; rather, they look at bundles of goods, so that the appropriate unit of analysis is not an individual good, but a bundle of goods. This means that if Castro consistently chooses unemployment over employment, then even if he claims to prefer working to unemployment, his

[73] Kevin Lancaster, 'A New Approach to Consumer Theory', *Journal of Political Economy*.74, 2 (1966): 132-157. Gary

behaviour reveals a preference for a good related to unemployment. Analysing Castro's behaviour as a bundle then, allows an explainer to narrow down which attributes are preferred.

I think this response does not quite answer the request for a criterion. We can still ask why long periods of unemployment reveal a preference for unemployment in particular, because there are two kinds of possibility which a bundle analysis does not rule out. In the first instance, Castro's chronic unemployment may not be a choice to work, but a situation in which he is forced not to work. If construction companies are prohibited by law from offering jobs under R2000 per month (which they currently are), then even if they are willing to offer jobs at wages that will bring the labour market into equilibrium, they cannot do so. Castro is not faced with the choice of two options – work at a lower wage or not work at all – but is forced into chronic unemployment. In this scenario, his unemployment has nothing to do with his acting on his preferences, and everything to do with the force of external constraints.

Alternatively, Castro's unemployment could represent repeated, bad attempts to satisfy a preference for work based on false beliefs. Suppose Castro believes that the only way to get employment is to go through a labour brokerage, but the broker has no intention of hiring out Castro's labour because she dislikes foreign nationals. Castro is nevertheless convinced that the problem lies with the lack of jobs and not the broker, so he stays with the broker in the hope that he will eventually find some kind of employment. The problem of false belief is not so much that we cannot always be certain of the intended outcome of a given action. Rather, it is that this uncertainty generalises out to the possibility of extravagantly false beliefs, which completely break any correspondence between means and ends. When an agent acts on a conviction that unemployment is the means to flying to the moon, we are at a loss as to how one can explain this action – if she is acting on a reason, it is not a reason that the rest of us can understand.

In general, these two gremlins interfere with the efficacy of the linguistic criterion, because they impede the equation of a choice for *x* with a preference for *x* or one of its attributes. Appealing

---

Becker, 'A Theory of the Allocation of Time', *The Economic Journal*, 75, 299(1965): 493-517.

to the stated preferences of an individual does nothing to remove these possibilities, because this is just a verbal form of behaviour and is therefore subject to the same problems of coercion and recurring false beliefs.[74]

In order to eliminate them, two conditions must be in place:

iii. an explainer interprets Castro's unemployment as goal-directed behaviour – it is undertaken as a means towards some end

iv. and she assumes that Castro's means correspond with his ends; that is, she assumes that Castro is unemployed because he believes it will satisfy a corresponding preference

The first condition excludes the possibility that Castro's unemployment is akin to imprisonment; it ensures that unemployment counts as an action, rather than a state. The second condition secures against the risk that Castro's action is a function of persistent and extravagant false beliefs. Hence, it is only once Castro's behaviour has been characterised as a means aiming towards a corresponding end that the linguistic criterion can pick out a term which overlaps in statements about his unemployment and his preference for unemployment. A behaviouristic equation of preferences with choices therefore relies on a logically prior distinction between means and corresponding ends. This does not mean that preferences cannot be understood in terms of choices, but it does mean that there is an interpretational element to this approach – preferences are not discernible from just any kind of explanatory stance, but can only be distinguished if one treats behaviour as a means towards some corresponding end.

It follows that the distinction between means and ends plays a significant role in axiomatic explanations. Although an explainer abstracts away the distinction, she only does so in order to simplify an explanation which is complicated by the separation of means and ends, and thereby the

[74] Stated preferences typically come in two forms: statements about an individual's willingness to pay for a certain good, and statements about an individual's willingness to accept compensation for the loss of a certain good. Stated preferences therefore involve a further complication: probabilistic beliefs.

many ways in which an agent's behaviour can be forced upon him by circumstances, and the ways in which he can fail to fit his means and ends together by entertaining false beliefs. So, while an explainer must consider these complicating factors in constructing her explanation, she may not include these in the final explanation for the sake of elegance, simplicity or concision.

Notice that a marginal explanation only differs from an axiomatic one in its distinction between means and ends. Once this difference falls away, an axiomatic explanation looks to be elliptical for a marginal one: it implicitly invokes the distinction between means and ends, while a marginal explanation makes this distinction explicit. A surprising corollary of this conclusion is that axiomatic utility does not differ in kind from marginal utility, but only differs in terms of the degree of simplification. Marginal utility is more complex and logically fundamental to axiomatic utility, because it harbours the distinction between means and ends upon which axiomatic utility relies. Thus, although the spectre of behaviourism fails as an objection, it is nevertheless interesting, because it reveals the logical relationship between axiomatic and marginal utility.

4. The Republican's Objection

The republican's objection essentially defends the sovereignty of economics from interference on the part of practical reason. I think it is more subtle than one might initially take it to be, and it is helpful to lay out the bare bones of the argument:

i. The distinction between economic and non-economic explanations is fundamental to understanding how economics works
ii. This distinction cannot be picked out with the aid of the instrumental principle, because both economic and non-economic explanations can invoke the principle
iii. The principle is therefore irrelevant to understanding how economic explanations work.

iv. Hence, it is pointless to categorise economic explanations in intentional/non-intentional terms, for these categories depend upon the instrumental principle, which is methodologically irrelevant.

Notice that this objection does not claim that the instrumental principle plays *no* role in economic explanations. Instead, it contends that it really does not matter whether the principle plays a role or not, because our understanding of how economic explanations work arises out of comparison between economic and non-economic explanations. So, this objection can grant that the instrumental principle plays an indispensable role in some economic explanations, and nevertheless maintain that the principle is not particularly interesting or useful in understanding how these explanations work.

Consider a parallel case. In our everyday life, we observe that free-falling oranges drop to the ground. When we next encounter a freefalling orange, we expect that it too will fall to the ground. Our expectation is explained by the way in which we understand a particular orange to instantiate a general pattern of behaviour. The orangeness of the orange is present for all cases, but contributes no explanatory power as to why our expectation is legitimate. On the republican's objection, pointing to the instrumental principle is like pointing to the orangeness of the orange: it gives us no further insight into how economic explanations work, regardless of whether it is indispensable.

In this section, I argue against the third premise. Building on the last section, I contend that the instrumental principle is a necessary component of axiomatic explanations. Since an axiomatic explanation is simply shorthand for a marginal one, the principle also applies to marginal explanations. This means that it plays a role in understanding the intentional nature of these two explanations. As such, it is both appropriate and helpful to frame marginal and axiomatic explanations in intentional versus non-intentional terms. So although the principle might not distinguish all economic explanations from non-economic ones, it is relevant to understanding how

a sub-class of economic explanations work. This is enough, I think, to rescue the applicability of the intentional/non-intentional categorisation to the domain of economics.

Recall the two conditions needed for an axiomatic explanation to work:

i. an explainer interprets Castro's unemployment as goal-directed behaviour – it is undertaken as a means towards some end

ii. and she assumes that Castro's means correspond with his ends; that is, she assumes that Castro is unemployed because he believes it will satisfy a corresponding preference

The first is a condition for viewing an agent's behaviour as intentional, and therefore as an action. The second condition is ambiguous between a weak and a strong interpretation of correspondence. On the weak reading, an explainer views an agent's end as giving him a reason to take the corresponding means. On the strong reading, an explainer views an agent's end as giving him a *good* reason to take the corresponding means, because he has true beliefs. This strong reading is typically found in neoclassical assumptions about an agent having perfect information; the weak reading is characteristic of behavioural economics, which takes imperfect information in the form of bounded rationality as its starting point.[75] On either reading though, Castro's behaviour is rational in the minimal sense that his end gives him a reason to pursue a particular action, even if the reason is an imperfect one.

An axiomatic explanation of an agent's behaviour must therefore treat this behaviour as *rational action*. Yet, as I argued in Chapter 2, an explanation of an agent's behaviour as rational action necessarily invokes the instrumental principle, which is normative and non-hypothetical. So the instrumental principle must be a feature of an axiomatic explanation, and as this explanation is elliptical for a marginal one, the principle must also be a feature of a marginal explanation. The instrumental principle therefore plays a necessary role in explanations which employ either

[75] I discuss behavioural economics more fully in the Coda at the end of this thesis.

axiomatic or marginal utility.

Simply put, marginal and axiomatic utility approach behaviour as preference-guided. Understanding behaviour as preference-guided is the same as understanding behaviour as rational action, so one cannot use marginal or axiomatic utility in an explanation of behaviour without invoking the instrumental principle.

It is not clear, however, that this answers the republican's objection. Remember, the objection is not that the principle plays no role in some economic explanations; it is that the principle does not help us understand how these explanations work. This is because it cannot be used to distinguish economic from non-economic explanations. This latter claim must be true, for the principle only characterises intentional explanations in opposition to non-intentional ones. An explanation of an agent's behaviour in terms of a universally-held fear of castration in men, for instance, may not count as an economic explanation, because the latter kind excludes reference to specific psychological content. It might nevertheless rely on the instrumental principle, for it could explain his action as a means aiming towards an (unconsciously-held) end. Still, I think the principle does not have to set the entire domain of economic explanations apart from all other kinds of explanations in order for it to play a useful role in helping us understand how a sub-class of explanations – marginal and axiomatic explanations – function.

Recall the axiomatic explanation of Castro's preference for unemployment. The explanation works by showing that Castro's preference for unemployment instantiates a pattern of choice. On the surface at least, we are not invited to view Castro as an agent acting to fulfil his ends, but as a bundle of tendencies to behave in particular ways. As such, this explanation seems to be non-intentional. Yet, an analysis of the conditions necessary for this explanation to work shows that its apparent non-intentional nature is a ruse. The explanation can only work if the explainer views Castro's behaviour as rational action, and therefore as guided by the instrumental principle. As such, an axiomatic explanation is substantially different from true non-intentional explanations of human behaviour. Compare an axiomatic explanation of an agent's behaviour with a medical one. Suppose

Castro has an increased heart rate and is sweating profusely. We could explain this by pointing to raised levels of adrenalin. In this explanation, it is utterly irrelevant whether Castro is intentional and rational, or catatonic. In an axiomatic explanation, however, the outcome of the explanation is very much dependent upon whether we treat Castro as a reasoning, intentional being, because an agent's preferences can only be identified with his choices on condition that we regard his behaviour in the light of rational action. So, although the intentional nature of an axiomatic explanation is an implicit, rather than explicit feature, it is nevertheless important to understanding how the explanation works.

The instrumental principle therefore contributes to our understanding of the way in which a sub-class of economic explanations – axiomatic and marginal explanations – work. Although it does not distinguish economic from non-economic explanations, it nevertheless gives us insight into how this sub-class of explanations functions. In this regard, we can return to the analogy with oranges. When we predict how a particular free-falling orange will behave, we do so by pointing to the way in which it instantiates a known pattern of behaviour in all free-falling oranges; even though orangeness is present in all cases, it does not play a role in the prediction, because we consider it to be separate from the objects' behaviour. When we explain an agent's behaviour by pointing to the way in which it instantiates a known pattern of choice, however, the instrumental principle is entwined in this explanation, so that if we try to take it out of the picture completely, the explanation unravels. Consequently, there is good reason to stand by the classification of intentional and non-intentional explanations, for axiomatic and marginal accounts are fundamentally characterised by their intentional nature.

Conclusion

I have defended the relevance of intentional/non-intentional explanations to understanding axiomatic explanations. In the first place, I argued that axiomatic explanations rely on an implicit

distinction between means and ends, and are therefore elliptical for marginal explanations. Moreover, this distinction between means and ends is of a special kind, for it requires an explainer to view an agent's behaviour as an instance of rational action. In order to do so, however, an explainer must invoke the instrumental principle, which is non-hypothetical and normative. The principle is therefore a necessary feature of both axiomatic and marginal explanations. As such, it is appropriate and helpful to approach these explanations as intentional rather than non-intentional.

At this point, it might be useful to consider Daniel Dennett's version of behaviourism.[76] He claims that (i) belief is a perfectly objective phenomenon, in the sense that a belief is simply a set of behaviouristic traits which we can observe in an object. Yet he also thinks that (ii) belief is only discernible from a certain predictive strategy – the intentional stance – and its existence can only be confirmed by considering whether it is useful to adopt the intentional stance with respect to a particular object. I have argued for a version of the second claim: a preference can only be discerned if one assumes that an agent's behaviour counts as rational action and is therefore subject to the instrumental principle. I am agnostic about the first claim – I do not think it matters much to my argument whether one believes that preferences are behaviouristic, hidden in an agent's head, or invented. This agnosticism is deliberate, because it allows for the application of utility theory to a wide variety of systems, like individual agents, households, corporations, states and animals, irrespective of whether one believes that a particular system actually possesses preferences and acts on reasons.

More importantly, it seems as if marginal and axiomatic utility can only be used in intentional explanations. In practice then, they necessarily exhibit normative and non-hypothetical properties. In theory, however, they are treated as positive and hypothetical. Consistency demands that we bring the abstract and applied versions of utility in line with each other, but which version should dominate? In this, I am inclined to agree with Cartwright, who writes: "I am an empiricist. I

[76] Daniel Dennett, 'True Believers' in *The Intentional Stance*(Cambridge, Mass: MIT Press, 1995): 15.

know no guide to principle except successful practice."[77] With this understanding of empiricism in mind, I think theorists with empirical leanings should be persuaded to favour the explanatory version of utility over its abstract counterpart.

Nevertheless, even if one is not persuaded to give up the more abstract version in favour of its explanatory counterpart, this conclusion should have some bearing on the gap between neoclassical practice and the capabilities approach. Given that axiomatic utility is elliptical for marginal utility, I take marginal utility to be paradigmatic of utility theory as a whole. In the next and final chapter, I argue that the capabilities approach makes explicit, and develops, the non-hypothetical and normative aspects of utility when it is used to explain human behaviour as an instance of rational action.

[77] Cartwright, *The Dappled World*: 2.

CHAPTER 4

Utility in Perspective: Just How Heterodox is the Capabilities Approach?

The preceding chapters have slowly built up a case for understanding marginal utility as normative and non-hypothetical when it is used to explain agents' behaviour. Although we can take marginal utility to be a paradigmatic case of utility theory in neoclassical economics,[78] this reflection on the methodology of utility theory does not yield any pressing recommendations for doing economics differently. It does, however, bring neoclassical practice closer to at least one heterodox theory. The dissenting theory I have in mind is the capabilities approach.

The capabilities approach is considered a significant departure from neoclassical economics and, in particular, utility theory. In place of the concept of utility and its emphasis on preference-guided behaviour, it offers 'capabilities' and 'functionings'. A capability is the opportunity to do things, or achieve functionings, which an agent has reason to value. While this approach comes in a variety of strains, this chapter focuses on the pioneering work of Amartya Sen. I argue that the gap between the capabilities approach and utility theory is not as great as Sen and others understand it to be. The capabilities approach makes explicit, and develops, the non-hypothetical and normative aspects of utility when it is used to explain human behaviour.

1. The Double-Edged Nature of Preferences

Over the course of four decades, Sen has developed a trenchant and multi-faceted critique of utility theory in general. This section concentrates on his argument against building an account of social choice in terms of individual preferences. One way of framing his argument is in terms of the paucity of information: the lack of content in a preference blocks a successful aggregation of

[78] I take marginal utility to be paradigmatic for utility theory in neoclassical economics, as axiomatic utility is elliptical

individual preferences into a social preference, and therefore cannot ground a criterion for social choice. So, even if preferences can be used to understand individual choices regarding private goods, they are not helpful in approaching social choices regarding public goods. This epistemic critique of preferences is, I think, an important component of the capabilities approach, so it is worth considering his argument in some detail.

As discussed in the first and second chapters, the concept of a preference upon which utility depends has no psychological or moral content. In one respect this is advantageous, for this renders utility independent of the truth of any particular psychological or moral theory, so that it can be applied to a variety of contexts – regardless of whether an agent acts altruistically or egotistically, cruelly or kindly, we can use utility theory to explain the action. On the other hand, this also means that utility cannot function as a criterion for judging the actions of ourselves and others. Something more must be added if it is to help us deliberate about how to act; that is, if it is to work as a standard for social choice.

The Pareto criterion is typically added to utility theory with the intention of rendering it applicable to questions of public goods. The principle states that if someone prefers *B* to *A*, and no one prefers *A* to *B*, then if that person's preference for *B* is satisfied, the system has increased her utility without decreasing anybody else's. It is Pareto superior to the system preceding it. If there is no state of affairs that is Pareto superior to this, then the system is Pareto optimal. A Pareto optimal market is a perfectly competitive or efficient market, because all available resources are used to improve some agents' utility, while no agents' utility is lowered.

The Pareto criterion is intended to encapsulate a commitment to what Sen calls ‘minimal liberalism’ – the view that each agent should be sovereign with respect to her private domain. It is minimal in the sense that it defines liberty in negative terms, rather than positive ones, for an individual’s liberty is a function of what others do not do to her. So a Pareto inferior system is meant to indicate an unjust state of affairs, in the sense that some individual’s sovereignty over her

---

for marginal utility.

choices has been violated. If it works, it can act as a criterion for guiding social choice towards more just states of affairs. Sen, along with many others, has argued that Pareto liberalism generates contradictory results, because an individual's 'meddlesome' preferences concerning others are treated on a par with more innocuous, private preferences.[79]

Sen asks us to consider a situation in which two individuals have meddlesome preferences. His original argument is framed in terms of two agents – Prude and Lewd – and their preferences regarding who should read *Lady Chatterley's Lover*.[80] We can imagine a more contemporary situation, in which a tree-hugging vegetarian and a self-avowed carnivore are about to order a meal. Tree-hugger prefers that both of them should eat vegetarian food; if only one of them eats vegetarian, however, he wishes it to be his meat-eating friend, as this would show her the value of vegetarian food. Carnivore, on the other hand, would prefer it if both of them ate meat, but would rather eat vegetarian herself than indulge her friend's descent into tree-hugging. Here are their preferences:

Tree-hugger	Carnivore
(*a*) Both eat vegetarian	(*d*) Neither eats vegetarian
(*b*) Only Carnivore eats vegetarian	(*b*) Only Carnivore eats vegetarian
(*c*) Only Tree-hugger eats vegetarian	(*c*) Only Tree-hugger eats vegetarian
(*d*) Neither eat vegetarian	(*a*) Both eat vegetarian

Notice that both have certain preference-orderings which interfere with the other dinner partner's preferences. Tree-hugger, for instance, prefers it if Carnivore eats vegetarian rather than only him eating vegetarian, while Carnivore would rather eat vegetarian than indulge her friend's unnatural appetite for vegetables. These meddlesome preferences are ruled out by minimal liberalism, because

[79] This is Julian Blau's term for preferences which are not solely private, but have implications for others. See his 'Liberal Values and Independence', *Review of Economic Studies* 43(1975): 395-401.
[80] Sen, 'The Impossibility of the Paretian Liberal', *Journal of Political Economy* 78(1970): 152-157.

they undermine the sovereignty of the other dinner partner over his/her private domain. On minimal liberalism then, Tree-hugger's preference for (*a* over *b*) or (*c* over *d*) may form part of a social choice, but his preference for (*b* over *c*) may not. Similarly, Carnivore's preferences for (*d* over *b*) or (*c* over *a*) are up for consideration, but her preference for (*b* over *c*) is not.

Suppose Tree-hugger prefers *a* over *b* and Carnivore prefers *c* over *a*. Since Tree-hugger prefers *a* to *b*, *a* should be socially preferred, and since Carnivore prefers *c* to *a*, *c* should be socially preferred to *a*. On a liberal account then, the socially optimal situation is one in which only Tree-hugger eats vegetarian food. Yet, both prefer *b* to *c*, so on the Pareto criterion, a socially optimal situation is one in which only Carnivore eats vegetarian food. This means that the Pareto criterion cannot guarantee that if everyone in a situation prefers a social state *b* to a social state *c*, then *b* will be chosen over *c*.

Sen thinks this shows that one cannot have a commitment to both the Pareto criterion and minimal liberalism, because this generates a contradiction. While the Pareto criterion does not add any content to the empty preferences which constitute utility, minimal liberalism demands some kind of restriction on preferences, such that they do not interfere with another's preferences – it requires a concept which has some content to it. For this reason, the moment an individual has a meddlesome preference the Pareto criterion runs up against the minimal liberalism it is supposed to espouse. Since meddlesome preferences are precisely those concerning public goods, it looks as if utility theory will only be able to accommodate the nominal demands of minimal liberalism when these demands are irrelevant to social choice – when they apply to private goods rather than public ones. Sen suggests that an 'informational enrichment' is needed to correct the conflict which the empty nature of a preference generates when it is applied to questions concerning public goods.[81]

2. Differentiating Commitment from Sympathy

[81] See, for instance, Sen 'The Possibility of Social Choice' in *Rationality and Freedom*: 92-94.

For Sen, part of this informational enrichment involves taking seriously people's account of their reasons for acting.[82] Some reasons for acting are based on what he calls *sympathy*, where a person feels that her wellbeing is affected by the state of others. In acting to improve their conditions then, she improves her own wellbeing. Other reasons for acting are based on *commitment*, where a person comes to have a preference on the basis of a reasoned analysis. One can be both sympathetic and committed to the same cause or person; the difference between the two is that sympathy is affective or conative, while commitment has a cognitive element.[83]

In an early paper, Sen frames commitment in explicitly moral terms,[84] but he later comes to recognise that there are many instances in which this need not be the case. Say you prefer to get drunk tonight, and you also prefer to finish writing a paper tonight, but you believe that the two are incompatible.[85] Furthermore, you believe that intellectual work is more enduring than intoxication. Were you to have a preference for the permanent over the fleeting, then you would have a reason to commit to working. This commitment arises through reflection on the nature of intellectual work and what Sen calls "meta-rankings" – the reasoned coordination of preferences.[86] Suppose you subsequently come to see that intellectual work is not as enduring as you had previously thought, or that your desire for permanence ranks below your desire for alcohol; it is then possible for you to relinquish your commitment to work in light of these considerations. So although commitment and sympathy both belong to the class of preferences, commitment is distinct in that it is formed by, and is amenable to, rational deliberation.

This cognitive aspect of commitment can be drawn out in three different ways. The first approach conceives of commitment as receptive to considerations which are wholly independent of

---

[82] Sen uses both fiction and thought experiments to draw out our everyday understanding of ourselves as acting on reasons which are based on commitments. See, for instance, 'Rational Fools: A Critique of the Behavioral Foundations of Economic Theory', *Philosophy and Public Affairs* 6, 4(1977): 326-329.

[83] As Sen recognises, the distinction between commitment and sympathy has a long pedigree. Adam Smith invokes it, for instance, in his discussion of the differing roles and make-up of prudence and affective states like sympathy and generosity. See Sen, 'Why Exactly is Commitment Important to Rationality?', *Economics and Philosophy* 21(2005): 10.

[84] In 'Rational Fools' Sen claims that "commitment is of course closely connected to one's morals" (1977, 329).

[85] This is a variation on an example taken from 'Rationality and Other People', in *The Idea of Justice*, 192. In Sen's example, commitment is other-directed; my example takes out the interpersonal dimension to show more clearly how non-moral commitment is possible.

[86] Sen, 'Liberty and Social Choice', in *Rationality and Freedom*: 402.

an agent's preferences. A Kantian, for instance, would maintain that rationality requires one to follow certain rules and these rules give one reason to form certain commitments.[87] If you commit to working on the grounds that you have a duty to do so, then your deliberation has Kantian overtones. In this instance, we can interpret your reason for commitment as independent of any preferences you might have. Instead, we can understand it as a reflection of your identity as an agent and of the principles which you allow yourself to be guided by in virtue of being an agent. Through a process of rational deliberation, you then come to form a new preference for working, or commit to a pre-existing preference for working.[88]

A second, less demanding approach is suggested by Elizabeth Anderson. Drawing on a more aretaic conception of reasoning, she argues that commitment is based on reasons that we can all have in virtue of shared preferences.[89] A preference for survival in one form or another might in this sense give one a reason to commit to producing what one takes to be enduring intellectual work. On the grounds that intellectual work is a means towards the end of survival, one has a reason to commit to work; insofar as others share a preference for survival, it is possible for them to have a similar reason to commit to intellectual work.

A third approach views commitment as an outcome of public debate, in which one takes on the preferences of another agent that are relevant to one's commitment and reasons through them to see whether one would arrive at a similar commitment. The thought here is that one's commitment must be able to sustain reflection, and part of that involves opening it up to public scrutiny. It is not obvious that doing so requires one's particular ends to be a rational function of duty or shared ends.

---

[87] For an incisive and subtle Kantian account, one can look to Korsgaard's *The Sources of Normativity*.

[88] Preference-independent reasons for acting can also be construed along the lines of ethical realism, although Sen does not pursue this avenue. I have in mind here a version of Cornell Realism, which views moral properties as straightforward natural properties, where recognition of these properties gives one a reason to form a commitment. When I ask why you are working instead of drinking, and you respond by pointing to some quality of intellectual endeavour that, in and of itself, weighs in favour of commitment, you are thinking like a Cornell realist. See, for instance, Nicholas Sturgeon, 'Ethical Intuitionism and Ethical Naturalism,' in *Ethical Intuitonism: Re-evaluations* (Oxford: Clarendon Press 2002).

[89] Anderson frames shared preferences in terms of group agency, but as a universal preference for individual survival is meant to show, group agency is not obviously necessary to shared preferences. I have therefore omitted the group agency aspect as an unnecessary complication. Elizabeth Anderson, 'Unstrapping the Straightjacket of 'Preference': A Comment on Amartya Sen's Contributions to Philosophy and Economics', *Economics and Philosophy* 17 (2001): 21-38.

Instead, one's ends can be informed by a kind of identification with others. Thomas Scanlon describes the process thus:

> thinking about right and wrong is, at the most basic level, about what could be justified to others on grounds that they, if appropriately motivated, could not reasonably reject.[90]

The submission of one's commitments to public scrutiny is not a uniquely moral phenomenon. When citizens come together to debate the merits of nominating a particular musician for a lifetime achievement award, they allow their reasons for commitment to be tested by public deliberation, by admitting the reasons that others might have for or against a particular musician into their deliberations. Describing the process as a kind of replication of another's decision process, Jane Heal writes:

> I can imagine how my tastes, aims and opinions might change and work out what would be sensible to do or believe in the circumstances. My ability to do these things makes possible a certain sort of understanding of other people. I can harness all my complex theoretical knowledge about the world and my ability to imagine to yield an insight into other people *without any further elaborate theorizing about them*. Only one simple assumption is needed: that they are like me in being thinkers, that they possess the same fundamental cognitive capacities and propensities that I do. The method works like this … I place myself in what I take to be [an agent's] initial state by imagining the world as it would appear from his point of view and I then deliberate, reason and reflect to see what decision emerges.[91]

[90] Thomas Scanlon, *What we Owe to Each Other*(Cambridge, Mass: Harvard University Press, 1995): 5.
[91] Jane Heal, 'Replication and Functionalism', in *Language, Mind, and Logic*(Cambridge: Cambridge University Press, 1986): 137.

Unlike a Kantian or aretaic conception of reasoning, this kind of reasoned identification with others does not rely on universal principles or ends, but simply on the recognition that others are sufficiently similar to oneself. It relies on a sense that other human beings can think and feel, and that they too have needs and desires.

Sen has at one time or another suggested all three interpretations of commitment, and tends to see them as linked. For Sen, there are mutual benefits to be reaped by considering whether others could rationally share our commitments and allowing our commitments to be guided by social norms.[92] Reasoned reflection on our commitments can therefore take different forms. "If rationality were a church", he writes, "it would be a rather broad church."[93]

Common to all three approaches, however, is the view that commitment is non-hypothetical: it does not depend on preferences which are particular or unique to an agent, but is instead dependent on reasons and/or preferences that can be shared by others.[94] Central to Sen's positive story is therefore the claim that a sub-class of preferences – commitment – is open to public deliberation. Moreover, as a commitment is based on reasons, it can be better or worse with respect to those reasons. If one's commitment to intellectual work is based on some false belief, like the belief that duty to one's society demands it, or that it is necessary to achieve posterity, or that anyone in the same position could be reasonably expected to give up everything for the sake of work, then this reason is not as good as one which is based on true belief. Taken together, the normative and non-hypothetical properties of commitment provide a way of understanding commitment as bound up in, and responsive to, the judgements of a community.

In separating commitment out from sympathy, Sen intends to drive a wedge between an

[92] In 'Rationality and Other People', Sen argues that a commitment can arise out of respect for social norms – this is a kind of deontological position. In 'Plurality of Impartial Reasons', Sen points to the way in which we come to recognise the needs of others by considering our own needs.
[93] Sen, 'Plurality and Impartial Reasons', *The Idea of Justice*: 195.
[94] The notion of commitment is agnostic with respect to the reasons internalism/externalism debate. Internalism claims that all reasons for acting are relative to an agent's conative motives, while externalism denies this. For instance, one can understand a commitment as a reasoned affirmation of a particular desire, so that one's reason for acting depends on a conative motive that one *could* commit to if perfectly rational. This deontological position satisfies the internalist

agent's choice behaviour and her wellbeing, because a commitment may lead her to act in ways which are not obviously beneficial to her, but which are still rational.[95] Individuals' sacrifice under apartheid in virtue of their commitment to a common humanity is a compelling example of the way in which actions can detract directly from one's wellbeing. In this way, the concept of a commitment forms part of a broader critique of egoism in economics, because it aims to show that actions can be rationally guided by considerations which have nothing to do with one's own wellbeing. As in philosophy, egoism takes three forms in economics.[96] Psychological egoism claims that agents always act out of self-interest; normative egoism maintains that acting out of self-interest is the only morally good kind of action; rational egoism claims that it is irrational not to act of out self-interest. The construct of *homo economicus*, as an agent who always acts out of self-interest, is only rational when he does so and ultimately ensures the wellbeing of society at large by acting out of self-interest, best encompasses these three facets of egoism. Given the empty nature of a preference, however, utility theory does not imply egoism; we can use utility and reject any and all forms of egoism with perfect consistency.[97]

Sen's critique of egoism in economics is therefore less important than his attempt to turn economic analyses towards a more complex understanding of normativity. By separating individual wellbeing from rational choice, Sen draws economists' attention to the way in which the domain of the normative is more nuanced and multi-faceted than they may have taken it to be.[98] On his account, commitment is based on reasons which need not be orientated towards personal or general wellbeing. Actions which are based on commitment, he argues, are not necessarily "a corollary of

---

requirement. See Korsgaard, 'Skepticism about Practical Reason', *The Journal of Philosophy* 83, 1(1986): 5-25.

[95] Sen, 'Rationality and Freedom' in *Rationality and Freedom*: 33-42.

[96] The appeal of egoism may spring from a tendency to misinterpret Mill and Adams as advocating a kind of psychological and normative egoism. I have not pursued this issue, but one can look to Don Ross' *The Concept of Utility* for a cogent and sustained critique of this misinterpretation in economics.

[97] For a fascinating example of utility theory shorn of egoism, see Samuel Bowles and Herbet Gintis, 'Homo Reciprocans: Altruistic Punishment of Free Riders', *Nature* 415(2002): 125-128. Bowles and Gintis argue that there is evolutionary evidence for the altruistic *punishment* of free-riders, and that one can construct a utility function for this behaviour.

[98] This is not, I think, an unfair characterisation. As I suggested in the first chapter, significant thinkers like Jevons and Robbins took the normative domain to be entirely moral.

any general pursuit of well-being", either with regards to one's own wellbeing or that of others.[99] Yet, as the conflict between drinking and working is meant to illustrate, there is nothing peculiar or mysterious in allowing our actions to be guided by commitments that are not obviously or directly related to wellbeing. Nevertheless, in these cases, one's reasons for commitment can be better or worse; if they detract from a more strongly-held preference, or depart from one's duty, or cannot survive public scrutiny then one has possibly based a commitment on bad reasons, and laying them open to public debate should reveal this. The concept therefore has a normative dimension distinct from moral or prudential concerns, for it is primarily concerned with the demands of reason and the way in which these demands can be tested in the public arena. As with instrumental reasoning, we can make sense of this kind of practical reasoning without endorsing it as moral or prudential. In and of itself then, commitment only implies the kind of minimal normativity characteristic of the instrumental principle: it does not necessarily invite the attitudes of praise and blame associated with judgements about wellbeing.

3. The Capabilities Approach: An Epistemic Turn

The concept of commitment is fundamental to the capabilities approach. Sen argues that wellbeing should be assessed in terms of the capabilities to realise functionings which agents have reason to value, or be committed to. On the face of it, the minimal normativity of a commitment seems to be at odds with Sen's focus on wellbeing, because Sen explicitly maintains that commitments are not necessarily an index of individual or general wellbeing.

The quick response to this is that commitment is a necessary, but not sufficient condition for assessing wellbeing. Some commitments may have no bearing on anyone's wellbeing (like the commitment to work), but others will have a bearing on what we ordinarily take to be our wellbeing – being able to nourish oneself or one's children, go to school, have decent shelter and so on. As

[99] Sen, 'Rationality and Other People', in *The Idea of Justice*: 192.

commitment is based on reasons that are non-hypothetical, these reasons can be shared and compared by a community of agents. On the assumption that we have reason to value wellbeing, this communal aspect of commitment gives us a way of assessing and refining our reasons for valuing wellbeing.

Moreover, by focusing on the sub-class of preferences that is open to rational deliberation, the capabilities approach is able to take into account the way in which severely disadvantaged, vulnerable individuals adapt their preferences to their limited opportunities or have their preferences shaped by them through the force of social constraints. In this regard, Sen writes:

> The extent of a person's deprivation . . . may not at all show up in the metric of desire-fulfilment, even though he or she may be quite unable to be adequately nourished, decently clothed, minimally educated, and properly sheltered"[100]

A woman who has suffered multiple rapes, for instance, may be indifferent to the prospect of being raped again, because a fiercely-held contrary preference is more psychologically damaging than indifference or acceptance. If we take the satisfaction of preferences in general as a fundamental indicator of wellbeing, then we cannot think of her as substantially worse off than a woman who has not been raped. Yet if we ask whether a *commitment* to being raped is warranted, then we can assess the reasons underlying the commitment and find it severely wanting. Regardless of her indifference or acceptance, we might have good reason to consider her as having been deeply injured by multiple rapes, and thereby arrive at a commitment to protecting women from being raped.

Nevertheless, as Sen conceives it, the capabilities approach is not intended to function as a decision procedure for how to act; nor does it provide a single, overarching criterion for judging wellbeing. Instead, it has the more modest aim of acting as an "informational focus".[101] Writing on the kind of informational focus that he thinks appeals to Sen, Scanlon characterises it thus:

[100] Sen, *Inequality Reexamined* (Cambridge, Mass: Harvard University Press 1992): 55.

> [it] does not require a single standard of overall evaluation. What it requires is, rather, an account of what various individuals, in virtue of their diverse positions, have reason to want, and a way of comparing these reasons – not by deciding what is best overall but by comparing the importance of a particular benefit from one position with the importance of a burden from some other position.[102]

Sen sometimes calls this act of comparison a kind of "positional objectivity", in which we open up our reasons for commitment to critique from other agents, and in turn try to see their reasons for commitment from their perspective.[103] The capabilities approach therefore has a strong epistemic orientation to it: it tries to answer the question, 'how do we come to identify the level of wellbeing for an agent or group of agents, given that we already have a commonsense understanding of what it means to live well?' The answer is that we should look to the capabilities that agents have to realise functionings they can commit to.

Notice that a capability gets its sense and value primarily from the functioning that it is related to, and it is therefore assessed in instrumental terms. Although Sen makes room for more deontological concepts – like duties and rights – and thereby allows that some of these concepts can also have a non-instrumental role, the focus remains on the instrumental relation between capabilities and functionings. The grounds for this are partially pragmatic. It is sometimes simpler and more effective to approach another's wellbeing in terms of the capabilities to achieve functionings that one also has reason to value, rather than considering abstract arguments about the non-instrumental value of rights. The difference between the two methods largely lies in whether one takes one's own position to be important in coming to understand how another's wellbeing is undermined. In the process of reasoning about which capabilities are necessary for a given

---

[101] Sen, 'Lives, Freedoms and Capabilities', in *The Idea of Justice*: 232.
[102] Thomas Scanlon, 'Symposium on Amartya Sen's Philosophy: 3 Sen and Consequentialism', *Economics and Philosophy* 17(2001): 49-50.

functioning, one either tries to take on the appropriate preferences of another agent, or see whether one's own preferences can lead one to see how a certain capability is necessary for wellbeing. Divorcing capabilities from functionings, on the other hand, has the effect of removing ourselves from the picture, in the sense that we do not take our own preferences and perspectives into account. This makes it difficult to understand the force of certain constraints on others, because we are not required to take into account how we would feel and respond in their situation. Sen points to the way in which the "positional relevance of parenthood" can allow one to see how one has a particular duty to care for one's children, and this in turn can lead one to recognise that other children have similar needs.[104] So, restricting ourselves to the non-instrumental value of a capability also blocks off avenues of reasoning that rely on one's own positionality; it checks us from reasoning from the perspective of shared preferences or from the perspective of another's preferences. This, Sen thinks, is an arbitrary homogenisation of the plurality of reasoning processes available to us.[105]

A deeper reason for this emphasis on the instrumental relation between capabilities and functionings can be found in Scanlon's distinction between "representational consequentialism" and "foundational consequentialism".[106] Foundational consequentialism is a familiar view in debates about what makes an action morally good or right. It begins with a notion of value and then tries to show that the best state of affairs should be measured by this standard, and no other. Classical utilitarianism is an example of this approach. Representational consequentialism, on the other hand, starts with widely shared intuitions about what makes an act right or wrong – its consequences, the intention behind the act and the person performing the act. It then tries to build an account of value which can accommodate these intuitions, such that an act is morally right only if it results in the best consequences. So, while foundational consequentialism aims to provide a criterion for judging actions as morally right or wrong, representational consequentialism tries to show how it is that we

[103] See, for instance, Sen's 'Positional Objectivity', *Philosophy and Public Affairs* 22 (1993): 126-145.
[104] Sen, 'Position, Relevance and Illusion': 160-161.
[105] Sen, *The Idea of Justice*: ix- xi.

come to make these judgements in the first place.

As a kind of representational consequentialism, the capabilities approach can therefore accommodate deontological and aretaic considerations. It can, for instance, claim that the intrinsic value of a right contributes to the overall goodness of a state of affairs. Contrasting this representational approach to more foundational theories like utilitarianism, Sen writes:

> Some ethical theories (like utilitarian ethics) insist … that nonutility features, such as actions, must not be considered to be of any value or disvalue on their own (rather, only for the utilities or disutilities that they generate). But since there are good reasons to be concerned about some of these features, such as actions, motives, and the like (for reasons that are close enough to those which move deontologists, among others, to take note of them, in their own special way) the utilitarian exclusion ends up being an arbitrary exclusion of a class of reasoned demands.[107]

The emphasis on the instrumental relation between capabilities and functionings is therefore intended to help us make sense of how we come to value rights and motives in non-instrumental terms; it is not intended to ground a theory about what makes an action right or good. As such, the instrumental relation plays an epistemic role, rather than a value-conferring one.

4. The Capabilities Approach and Utility Theory

Sen, along with others, has characterised the capabilities approach as a substantial alternative to neoclassical economics. In this section, I focus on the difference between the capabilities approach and utility theory, as a significant aspect of neoclassical economics. I argue that, while there may be a great deal of distance between the capabilities approach and neoclassical economics as a whole,

[106] Scanlon, 'Symposium on Amartya Sen's Philosophy: 3 Sen and Consequentialism': 39-40.

the gap between the capabilities approach and utility theory is not so great as Sen and others take it to be.

In his comparison between utility theory and the capabilities approach, Sen writes:

> In contrast with the utility-based or resource-based lines of thinking, individual advantage is judged in the capability approach by a person's capability to do things he or she has reason to value. A person's advantage in terms of opportunities is judged to be lower than that of another if she has less capability – less real opportunity – to achieve those things she has reason to value. The focus here is on the freedom that a person actually has to this or be that – things that she may value doing or being … Indeed, it proposes a serious departure from concentrating on the *means* of living to the actual *opportunities* of living.[108]

On Sen's account then, the difference between the capabilities approach and utility theory lies in the distinction between means and opportunities, or capabilities.

At first blush, it is difficult to see precisely how means and capabilities differ. Although capabilities may have non-instrumental value, they are primarily understood as instrumental to some functioning. Capabilities, like means, are therefore predominantly instrumental in nature. Nor can the difference lie in the thought that capabilities are not restricted to resources, for an agent's means to her ends can range from an action to a state of affairs or object, and need not be confined to material goods, but can also include skills and dispositions. If an active disposition to be loyal to a group can achieve some end that an agent has in mind, then it may count as a means just as much as money or food does.[109] Similarly, the ability to read can function as a means to attending high school or gaining white-collar employment. Moreover, means and capabilities are not necessarily

---

[107] Sen, 'Consequential Evaluation and Practical Reason', *The Journal of Philosophy* 96(200): 487-488.
[108] Sen, *The Idea of Justice*: 231-233.
[109] See, for instance, George Akerlof and Rachel Kranton, 'Identity and the Economics of Organizations', in *Journal or Economic Perspectives* 19, 1(2005): 9-32.Akerlof and Kranton argue that employee's loyalty to a firm should be cultivated by managers, on the grounds that this incentivises them to work for less wages. As such, the cultivation of loyalty in employees is a means to the end of profit-maximisation for managers.

distinguishable in terms of whether one has actually taken the means to one's end. The technique of asking an individual what her preferences are, and then reasoning as to what the means are to realise these preferences, is neither alien nor particularly controversial to neoclassical economics. As discussed in Chapter 3, there are often scenarios in which an agent cannot physically choose between two options; one can then use contingent valuation techniques, like willingness to pay studies, in order to discover what an agent would choose if given the opportunity – doing so shows her hypothetical choices rather than her actual choices.[110]

Instead, following Sen, one could characterise the difference in terms of the role of reason. The thought is that a person's preference for *x* is satisfied because (i) it is her preference and (ii) she has reason to commit to *x*. In his discussion of the first condition as a kind of condition for agency, Phillip Pettit writes, "[i]t must be that my preference is in my control, so that what I get is robustly connected, not just connected by chance, with what I prefer."[111] Yet, as I argued in the preceding chapters, this robust connection between one's preference and the satisfaction of those preferences is not alien to utility theory, for it makes a similar link between one's choice of a good and its ability to satisfy one's preference in terms of the instrumental principle. Insofar as one explains a person's choice of one good over another, one assumes that she aims to connect that choice with an end of hers, for the possession of this end gives her reason to pursue the means to it. The agency of a person is therefore a crucial feature of the explanation. As in the discussion of 'voluntary' unemployment, if it turns out that a person is not actually acting on her preferences, then the explanation of unemployment as a choice cannot work. Thus, although the emphasis on agency is a significant feature of the capabilities approach, it does not distinguish capabilities from means in utility theory.

Rather, it must be the second condition which distinguishes a capability from a means. In

[110] For a clear example of this, see Jeannette Snowball and Kenneth Willis, 'Estimating the Marginal Utility of Different Sections of an Arts Festival: The Case of Visitors to the South African National Arts Festival', *Leisure Studies* 24(2006): 43-56. Snowball and Willis designed a willingness to pay questionnaire and distributed it to low-income households. The results showed that these households value aspects of the National Arts Festival that they do not attend as they lie beyond their income.

[111] Phillip Petit, "Symposium on Amartya Sen's Philosophy: 1 Capability and Freedom: A Defence of Sen", *Economics*

this regard, Sen argues that in contrast to utility theory, attaching "importance to the agency aspect of each person does not entail accepting whatever a person happens to value as being valuable (i) unconditionally, and (ii) as intensely as it is valued by the person."[112] The difference between the capabilities approach and utility theory therefore lies in its emphasis on a sub-class of preferences – commitments. As discussed, a commitment is (i) minimally normative, in the sense that it is based on reasons which can be better or worse and (ii) it is non-hypothetical, in the sense that a commitment does not rely on the particular preferences of an agent, but instead relies on preferences and/or reasons that can be shared by all. The capabilities approach differs from utility theory, in that it does not assign value to whatever means are necessary for an agents ends, but only assigns value to those means which are necessary to realise ends that we can all rationally commit to.

It seems that if this is the case, then utility theory already contains within it the tools to arrive at something like the capabilities approach. Recall that when utility is used to explain human behaviour as rational action, it is necessarily governed by the instrumental principle which has two properties. (i) It is minimally normative, in that it articulates the explainer's expectation that an agent gives deliberative weight to her ends and that her reason for acting therefore conforms to a particular structure. (ii) It is non-hypothetical, in that following the principle does not depend on a particular end, but on the possession of ends in general.

It is nevertheless clear that the normative and non-hypothetical properties associated with the instrumental principle differ from those in the capabilities approach. The instrumental principle is concerned with the relationship between means and ends, rather than ends in and of themselves. In order to apply the principle to ends, one would first have to consider ends as capable of rational deliberation, as based on reasons which are open to interrogation. One way in which we could do so, would be to consider a particular end in terms of its being conducive to some further end that an agent has in mind; that is, in terms of it being a means to some other end. Call this an instrumental

---

and Philosophy 17(2001): 4.

approach to ends.

We can see this instrumental approach at work in economic analyses of behaviour which focus on the way in which agents adapt to changing conditions. Consider, for instance, two competing accounts of the impact of climate change on agriculture in the United States.[113]

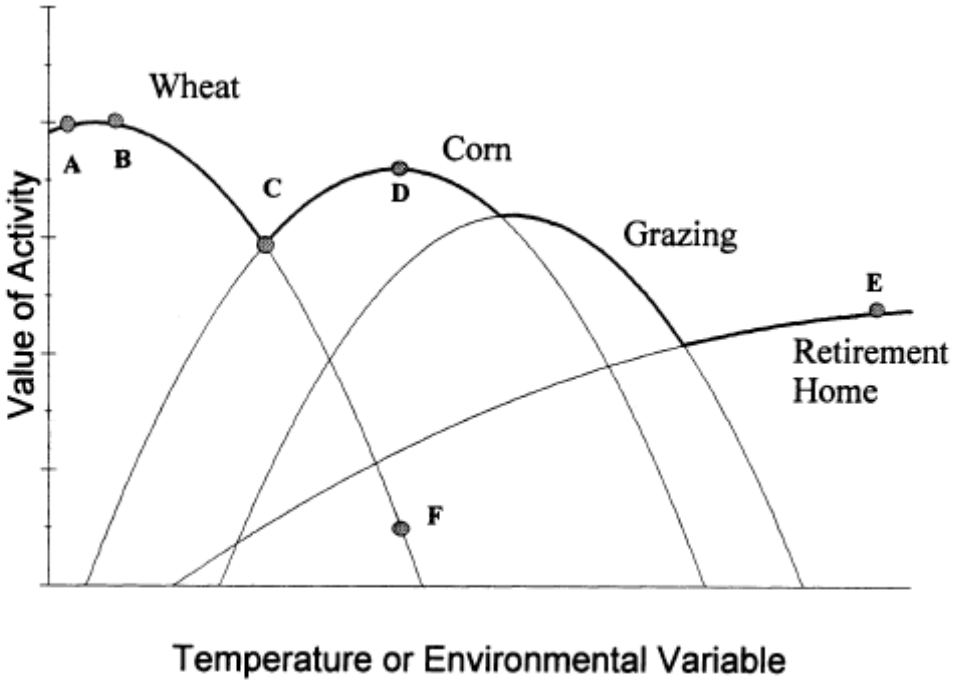

FIGURE 1. BIAS IN PRODUCTION-FUNCTION STUDIES

Figure 1 is an example of how the value of wheat, corn, grazing and retirement homes might look as a function of one environmental variable – temperature. Curve ACF is a hypothetical wheat-production function, and shows how the value of wheat falls when the temperature rises too high. ACF is typically the kind of analysis which operates in a "dumb farmer" scenario – it omits the variety of adaptations that farmers generally make in response to changing environmental and economic conditions. In a "smart farmer" scenario, however, farmers predict that wheat will

[112] Sen, *On Ethics and Economics* (Oxford: Baswell Blackwell, 1987): 43.
[113] This analysis is taken from Robert Mendelsohn, William D. Nordhaus and Daigee Shaw, ' The Impact of Global Warming on Agriculture: A Ricardian Analysis', *The American Economic Review* 84, 4 (1994): 754.

generate lower revenues than another crop, like corn. So at point C they switch their crop production to corn. This adaptation is represented by the curve ACD. Mendelsohn et al. argue that the "dumb farmer" scenario is implausible and leads to exaggeratedly high estimates of the agricultural costs associated with climate change, because it paints a false picture in which farmers blindly cultivate a crop until they are financially ruined (point F).

Notice that the only difference between the two accounts lies in whether preferences are taken as fixed. In the "dumb farmer" scenario, a preference for wheat growing never changes. In the "smart farmer" scenario, a farmer's preferences for a particular crop changes in response to temperature variation. The graph is intended to explain why land-use preferences change – as soon as one kind of crop fails to maximise profit, famers begin cultivating a new kind of crop. So this "smart farmer" explanation works by positing a principle preference for profit-making and a secondary preference for one or other kind of crop as a means to making profit.

The graph is essentially a kind of utility function for farmers, which maps the capacity of a crop to satisfy a primary preference for profit. When a crop fails to maximise profit, a farmer switches to a new crop. This means that the graph also maps the shift in secondary preferences – from wheat to corn to grazing, until finally they prefer to use the land for non-agricultural purposes, like retirement homes. We are invited to see the shift in secondary preferences as appropriate from the perspective of a profit-maximising farmer, so that a farmer who fails to change his preferences for land-use fails to act appropriately. He is like the "dumb farmer" who blindly continues along ACF. In this sense, we can understand the explanation as one which relies on a view of secondary preferences as amenable to rational deliberations concerning whether a particular land-use is the best means to the end of maximising profit. So, farmers' realisation that a change in temperature makes one kind of land-use unprofitable gives them a reason to change their preferences to another kind of land-use. More concisely, the explanation only works if we treat the secondary preferences as cognitive, rather than conative.

Economic explanations which use utility theory and take cognisance of human adaptability

and ingenuity therefore treat some ends as cognitive. They can do so without making reference to deontological considerations about one's duty or aretaic considerations about universally-shared preferences. They need only invoke the instrumental principle, and its stipulation about the right kind of fit between means and ends. In doing so, they bring economic practice (in the form of explanations that use utility theory) closer to the capabilities approach than Sen might have led us to believe.

Whereas cognitive preferences, or commitments, are a central, explicit feature of the capabilities approach, at most they play an implicit role in some economic explanations. There is only room for cognitive preferences in these explanations, however, in virtue of these explanations' reliance on the instrumental principle. The principle is normative in the sense that it articulates our expectation that an agent has a reason for acting, where this reason aims to connect means to ends. So it allows us to see some preferences as based on instrumental reasons. The principle is non-hypothetical, in that the structure to which these reasons ought to conform is independent of any particular end she might have, but is instead embedded in the possession of an end in general. Insofar as we possess ends then, these reasons are open to public deliberation and judgment. If farmers were to change from a highly-profitable and sustainable land-use to a ruinously expensive land-use, then the instrumental principle would allow us to judge their reasons as bad ones, and we might seriously buy into the "dumb farmer" scenario.

This similarity between the capabilities approach and explanations of human adaptability in terms of utility does not render the two identical. It does not, for instance, undercut the plurality of functionings that Sen offers in place of a single over-arching criterion of utility. Neither does it make room for Sen's conviction that instrumental reasoning is but one of many ways of reasoning. While I do not wish to underplay these significant differences, it seems to me that the capabilities approach is best viewed as being on a continuum with utility theory. On the one end, lie explanations of behaviour in terms of utility theory, which have implicit non-hypothetical and normative properties. Further along down the line lies the capabilities approach, which makes these

properties explicit and then develops them in interesting directions.

The capabilities approach represents, I think, an attempt to reconcile two directions in economics. In the first chapter of this thesis, I pointed to the way in which the discipline of neoclassical economics emerged as an attempt to explain agents' actions without buying into a codified system of ethics like utilitarianism. At the same time, however, part of what makes economic analyses interesting is their direct bearing on issues of wellbeing. By framing the capabilities approach as an epistemic project concerned with identifying how we go about assessing wellbeing, it looks as if Sen manages to incorporate ethical thinking into economic analyses, without requiring schematic accounts of precisely what it takes for an end or an action to count as moral. I take this to be expressive of a belief that, by and large, we generally do know what is good for us, and we do not need an overarching moral theory to decide or justify for us our moral beliefs. Nevertheless, by showing how legitimate communal judgements are possible, the capabilities approach offers us a way of comparing and developing our moral beliefs, and thereby offers us a way of comparing and developing our responses to suffering and deprivation.

The two fundamental properties which allow for such an approach are non-hypotheticality and normativity. That utility theory exhibits these properties, and therefore contains within it the tools to arrive at something like the capabilities approach, is a mark of its capacity to be interesting, to have a profound bearing on our understanding and pursuit of wellbeing.

CODA

Under the Carpet

In this thesis, I have tried to analyse the way in which utility theory works in relative isolation from the plethora of economic and philosophical considerations which have a bearing on utility. Sweeping these considerations under the carpet has the advantage of allowing one to focus closely on the way in which utility explanations work. Some of these considerations, however, have a direct bearing on the analysis and should be investigated further. Below is a brief discussion of two potential avenues of exploration with regards to expected utility theory and behavioural economics.

1. Expected Utility

This thesis was concerned primarily with the concept of marginal utility. The discussion of classical and axiomatic utility essentially served as a counterpoise to marginal utility, in order to bring out the way in which marginal explanations are (i) normative but not moral and (ii) concerned with behaviour as an instance of rational choice. I purposefully abstained from discussing the application of utility to predictions about human behaviour, and in particular, evaded the question of our beliefs about the future. The concept of *expected utility* has been developed to capture our beliefs concerning the probability of an expected outcome occurring. It raises the interesting and complicated question of how we assign probabilities to outcomes, and in this way connects theoretical reasoning (in the form of beliefs about what will come to pass) with practical reasoning (in the form of intentions about how to act).

The interplay between theoretical and practical reasoning is especially fertile ground for the debate between reasons internalism and externalism. This is a debate about how to read sentences of the form "An agent has a normative reason to φ". We can read such sentences as implying that "An

agent has some motive which will be served or furthered by her φ-ing", so that, if there is no such motive, it will not be true that "An agent has a normative reason to φ". This is the *internal* interpretation of such sentences. We can also read sentences of the form "An agent has a normative reason to φ" as not implying this, but as saying that she has a reason to φ even if none of her motives will be served or furthered by her φ-ing. This is the *external* interpretation of such sentences.

The basic idea of internalism is that we cannot have genuine reasons to act if those reasons have no connection to anything that we care about.[114] This thesis challenges certain natural and traditional ways of thinking about ethics. When we tell someone that he should not rape women, we usually understand ourselves as telling him he has a reason not to rape women. If internalism is correct, however, then the would-be rapist can show us that he has no such reason, because he does not care about anything which could be achieved by abstaining from sexual violence. So we seem to reach the conclusion that morality's rules are essentially indiscernible from trivial games like chess – they only apply to those who choose to join in by obeying them. This is typically a kind of neo-Humean position on morality.[115]

Notice, however, that this neo-Humean position requires either conativism about motives or a weak form of cognitivism about motives. The dumb farmer/smart farmer scenario in the fourth chapter provides an example of how ends can be amenable to rational deliberation in virtue of their being means to other ends. This is a form of weak cognitivism, as it does not imply that considerations external to a farmer's motivation set can provide a farmer with a reason to act; only considerations internal to the farmer – profit-making – can legitimately function as reasons.

If internalism is defined in terms of strong cognitivism about motives, however, then internalism can include more traditional positions on the nature of practical reason, for it essentially amounts to the claim that a reason to act must be directed towards some goal that an agent can come

[114] This interpretation of reasons internalism is taken from Bernard Williams, 'Internal and External Reasons', in *Moral Luck*.
[115] It is not limited to neo-Humeanism. For a persuasive account of how a contractarian is forced into this position, see

to have. This includes goals that are formed by or responsive to facts about the world (heavy-duty ethical realism), or rules that one lives by (deontology) or the reasoned demands of universally-held goals (virtue ethics). Consequently, one way of responding to internalism is to argue for strong cognitivism about motives, so that deontological and aretaic considerations fall under the compass of reasons internalism, without implying the disturbing conclusion that moral rules constitute just another set of parlour games. This robs internalism of its ability to challenge and unsettle our thinking about morality, and in this sense, constitutes a serious challenge to the internalist thesis. So although strong cognitivism about ends is not strictly speaking incompatible with internalism, it essentially dresses up the sentiments of externalism in an internalist guise. We can therefore take strong cognitivism about motives to be an under-cover version of externalism.

I have tried to frame the argument in such a way that it is consistently neutral between internalism and externalism. In the first place, the argument has been concerned with the reasons that agents actually act on (motivating reasons), rather than the reasons they ought to act on (normative reasons). Secondly, reasons for acting are broadly characterised in terms of their relation to ends, rather than motivations, and this leaves it open as to whether (i) ends just are conative motivations or (ii) ends can be cognitive in a weak sense.

The introduction of expected utility, however, makes this fence-sitting position a far more precarious one. When an agent deliberates about how to act, she forms beliefs about expected outcomes, so that reasons based on probabilistic belief are a paradigmatic case of practical reasoning; reasons about how one should have acted are atypical, as they are only indirectly concerned with action insofar as they dwell on how to interpret past actions. Expected utility theory therefore has a bearing on the reasons internalism/externalism debate, because it poses the question of whether belief about expected outcomes is conative or cognitive. Say Xoli believes it will rain two months from today. Subjectivism about probabilistic belief regards this belief as a subjective judgement that is independent of the truth-value of the statement “it will rain two months from

---

E. M. Zemach. ‘Ought, Is, and a Game Called *Promise*’. *The Philosophical Quarterly* 21, 82 (1971): 61-3.

today", and is instead dependent on something else, like prior subjective probabilities and preferences.[116] On this view, probabilistic beliefs have a strong conative element. Objectivism about probabilistic belief, on the other hand, sees Xoli as having a belief which aims to conform with the way the world is by following objective rules of probability (like decreasing one's conviction in the likelihood of an event occurring as the sample size of similar events occurring decreases).

If it turns out that probabilistic belief is subjective, then reasons for acting are typically relative to conative beliefs for those cases that matter most to practical reason – deliberations about how to act. Considerations about expected utility could therefore push us towards reasons internalism. This is an important issue that should be dragged out from under the carpet, dusted off and examined in much greater depth than I have given it here.

2. Behavioural Economics

The main claim of this thesis is that communal judgements about the nature of correct reasoning play a crucial role in our explanations of individuals' behaviour as rational choice. I have largely restricted the analysis of communal judgements to the instrumental principle, because this principle is considered to be less problematic than other rules of practical reasons. Considerations from behavioural economics suggest that isolating the instrumental principle in this way is not ultimately defensible.[117] As I understand it, behavioural economics is primarily concerned with investigating the nature of communal judgements by testing individuals' choice behaviour in as close to laboratory conditions as possible. The phenomena of hyperbolic discounting and framing effects are particularly important for understanding communal judgements about what counts as rational, because they suggest that norms for valuation and prudence may play a role distinct to the instrumental principle. If this is correct, then an understanding of the way in which utility

[116] For a concise and critical summary of subjective probability theory, see Michael Goldstein, 'Subjective Bayesian Analysis: Principles and Practice', *Bayesian Analysis* 1, 3(2006): 403-420.
[117] As the preceding analysis of utility theory did not assume that agents are perfectly rational, this analysis can take on

explanations work will inevitably have to draw on these experimental findings.

Faced with the option of two similar rewards, people tend to prefer the one that arrives sooner rather than later. In behavioural economics, hyperbolic discounting is a mathematical model thought to approximate this discounting process, in which valuations fall rapidly over small delay periods, but then fall more slowly for longer delay periods.[118] For instance, when offered the choice between R100 now and R200 a year from now, many people will choose the immediate R100. However, given the choice between R100 in five years or R200 in six years, almost everyone will choose R200 in six years, even though this is the same choice seen at five years' greater distance.

This temporal distancing effect may have implications for what we consider to be rational behaviour. If an agent chooses a current small reward which prevents her from realising a slightly larger reward in the near future, we may be inclined to judge this appropriate. Choosing to drink tonight over having a hangover-free morning the next day is one such example. If, however, an agent chooses a small reward over considerable rewards in the distant future, then we may be inclined to think of her as suffering from *akrasia* or weakness of will. A variety of studies, for instance, have found that addicts tend to discount the delayed consequences of their actions far more steeply than non-addicted individuals, and it looks as if this steeper discount curve accounts for our judgement that addicts are weak-willed.[119]

It does not seem as if the difference between the two lies in a failure of instrumental reasoning; after all, in both cases the agent takes the means to her immediate ends and fails to take the means to her more distant ends. Instead, it seems to lie in the permutation of our valuations with respect to time – it looks as if changes in time can give one a reason to value a good more or less. If this is correct, then a norm for valuation needs to be more carefully interrogated within the context of practical reason, and compared with existing discussions on the nature of value and its role in guiding actions. We need to ask whether this norm for valuation belongs to practical reason or is

---

board considerations from behavioural economics.

[118] See, for instance, George Ainslie and Nick Haslam, 'Hyperbolic Discounting' in *Choice over Time*: 57-92.

[119] Warren Bickel and Matthew Johnson, 'Delay Discounting: A Fundamental Behavioral Process of Drug Dependence',

better understood as an arational phenomenon; if it does belong to practical reason, it may be worth chewing over its relation to the instrumental principle, as it looks like it might be distinct from the principle.

A second phenomenon in behavioural economics is that of framing effects. Consider the stock example of the Asian disease problem.[120] In this study, a group of participants were asked to choose between two programs to counter the outbreak of an unusual Asian disease expected to kill 600 people:

A. 200 people will be saved
B. there is a one-third probability that 600 people will be saved, and a two-thirds probability that no people will be saved

In a second scenario, participants were asked to choose between:

C. 400 people will die
D. there is a one-third probability that nobody will die, and a two-third probability that 600 people will die

On the first version of the problem, 78 percent of participants thought that Programme A should be adopted. On the second version, however, 72 percent chose Programme D despite the fact that the outcome described in A is identical to the one described in C. The difference between the two scenarios lies in the way in which they are described or framed. When the programmes were presented in terms of lives saved, the participants preferred the secure program A. When the programmes were presented in terms of expected deaths, participants chose the gamble D. This

in *Time and Decision* (New York: Russell Sage Foundation, 2003).
[120] Amos Tversky and Daniel Kahneman, 'The Framing of Decisions and the Psychology of Choice', *Science* 211(1981): 453-458.

suggests that people tend to be more risk-averse when the options are framed as a likely gain; when the options are framed as a loss, they tend to be more risk-seeking.

As such, it appears likely that there is an inherent bias to maintaining the status quo, in the sense that actions seem to be orientated towards preservation of self and others rather than change. This seems to be something like the principle of prudence, which essentially stipulates for self-preservation. If this is correct, then something like prudence plays an important complementary role to the instrumental principle, and it is well worth asking what the nature of this relationship is, given the contemporary focus on the indispensability of the instrumental principle and comparatively little focus on prudence.

Although this thesis has been concerned with applying considerations from the philosophical discipline of practical reason to utility theory, it should be clear from this brief discussion of behavioural economics and expected utility theory that questions in practical reason can benefit from considerations in economics.

BIBLIOGRAPHY

Ainslie, George and Haslam, Nick. 'Hyperbolic Discounting'. In *Choice over Time*. George Loewenstein and Jon Elster (eds.). New York: Russell Sage Foundation, 1992.

Akerlof, George and Kranton, Rachel. 'Identity and the Economics of Organizations'. *Journal of Economic Perspectives* 19, 1(2005): 9-32.

Anderson, Elizabeth. 'Unstrapping the Straightjacket of 'Preference': A Comment on Amartya Sen's Contributions to Philosophy and Economics'. *Economics and Philosophy* 17 (2001): 21-38.

Aristotle. *Nicomachean Ethics*. Roger Crisp (trans., ed.). Cambridge: Cambridge University Press, 2000.

Becker, Gary. 'A Theory of the Allocation of Time'. *The Economic Journal*, 75, 299(1965): 493-517.

Becker, Gary. 'A Theory of Marriage: Part 2'. *Journal of Political Economy* 82, 2(1974): 11-26.

Bentham, Jeremy. 1954. 'The Psychology of Economic Man.' In *Jeremy Bentham's Economic Writings, Vol III*. Werner Stark (ed.). London: George Allen and Unwin, 1954.

Bickel, Warren and Johnson, Matthew. 'Delay Discounting: A Fundamental Behavioral Process of Drug Dependence'. In *Time and Decision*. George Loewenstein, Daniel Read and Roy Baumeister (eds.). New York: Russell Sage Foundation, 2003.

Blau, Julian. 'Liberal Values and Independence'. *Review of Economic Studies* 43(1975): 395-401.

Bowles, Samuel and Gintis, Herbert. 'Homo Reciprocans: Altruistic Punishment of Free Riders'. *Nature* 415(2002): 125-128.

Broome, John. *Ethics out of Economics*. Cambridge: Cambridge University Press, 1999.

Broome, John. 'Normative Requirements'. In *Normativity*. Jonathan Dancy (ed.). Oxford: Blackwell: 2000.

Broome, John. 'Reasons'. In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, Jay Wallace, Michael Smith, Samuel Scheffler and Philip Pettit(eds.). Oxford: Oxford University Press, 2004.

Carroll, Lewis. 'What the Tortoise Said to Achilles'. *Mind* 4, 14(1895): 278-280.

Cartwright, Nancy. '*Ceteris Paribus* Laws and Socio-Economic Machines'. In *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press, 1999.

Dancy, Jonathon. 'Enticing Reasons'. In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz.* J Wallace, P Pettit, S Scheffler and M Smith (eds.). Oxford: Clarendon Press, 2006.

Davidson, Donald. 'Actions, Reasons and Causes'. In *Essays on Actions and Events*. Donald Davidson (ed.). Oxford: Clarendon Press, 1980.

Dennett, Daniel. ‘True Believers’. In *The Intentional Stance*. Cambridge, Mass: MIT Press, 1995.

Dreier, James. ‘Humean Doubts about the Practical Justification of Morality’. In *Ethics and Practical Reason*. In *Ethics and Practical Reason.* Garret Cullity and Berys Gaut (eds.). Oxford: Clarendon Press, 1997.

Fine, Ben. ‘Economics Imperialism and the New Development Economics as Kuhnian Paradigm Shift?’. *World Development* 30, 12(2002): 2057-2070.

Geach, Peter. 'Good and Evil'. In *Theories of Ethics.* Phillipa Foot (ed.).Oxford: Oxford University Press, 1967.

Goldstein, Michael. ‘Subjective Bayesian Analysis: Principles and Practice’. *Bayesian Analysis* 1, 3(2006): 403-420.

Hausman, Daniel M and McPherson, Michael S. *Economic Analysis, Moral Philosophy and Public Policy*. Cambridge: Cambridge University Press, 2006.

Heal, Jane. ‘Replication and Functionalism’. In *Language, Mind and Logic.* James Butterfield (ed.). Cambridge: Cambridge University Press, 1986.

Hicks, John and Allen, R.G.D., ‘A Reconsideration of the Theory of Value. Part 1’. *Economics* 1(1934): 52-76.

Irwin, T.H. ‘Practical Reason Divided’. In *Ethics and Practical Reason*. Garret Cullity and Berys Gaut (eds.). Oxford: Clarendon Press, 1997.

Jevons, William Stanley. ‘Theory of Utility’. In *The Theory of Political Economy.* London: Macmillan Press, 1911.

Korsgaard, Christine. ‘Skepticism about Practical Reason’. *The Journal of Philosophy* 83, 1(1986): 5-25.

Korsgaard, Christine. *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996.

Korsgaard, Christine. ‘The Normativity of the Instrumental Principle’. In *Ethics and Practical Reason*. Garret Cullity and Berys Gaut (eds.). Oxford: Clarendon Press, 1997.

Lancaster, Kevin. ‘A New Approach to Consumer Theory’. *Journal of Political Economy* 74, 2(1966): 132-157.

Lucas, Robert. 'Unemployment Policy'. *American Economic Review* 68(1978): 353-357.

Mendelsohn, Robert, Nordhaus, William D., and Shaw, Daigee. ‘The Impact of Global Warming on Agriculture: A Ricardian Analysis’. *The American Economic Review* 84, 4 (1994): 753-771.

Mill, John Stuart. *Utilitarianism*. Roger Crisp (ed.). New York: Oxford University Press, 1998. Originally published in 1861.

Mill, John Stuart. ‘On the Definition of Political Economy and the Method of Investigation Proper to It’. In *Collected Works of John Stuart Mill*, vol. 4. Toronto: University of Toronto Press, 1967.

Nagel, Thomas. *The Possibility of Altruism*. Princeton: Princeton University Press, 1970.

Nagel, Thomas. *The Last Word*. Oxford: Oxford University Press, 1997.

Okun, Arthur. *Equality and Efficiency: The Big Trade Off*. Washington: Brookings, 1975.

Petit, Phillip. 'Capability and Freedom: A Defence of Sen'. *Economics and Philosophy* 17(2001): 1-20.

Railton, Peter. 'On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Actions'. In *Ethics and Practical Reason,* Garret Cullity and Berys Gaut (eds.). Oxford: Clarendon Press, 1997.

Robbins, Lionel. 'Interpersonal Comparisons of Utility'. *Economic Journal* 48, 192(1938):635-641.

Robbins, Lionel. 'Ends and Means'. In *On the Nature and Significance of Economic Science*. London: Macmillan Press, 1969.

Robbins, Lionel. 'The Subject Matter of Economics'. In *On the Nature and Significance of Economic Science*. London: Macmillan Press, 1969.

Ross, Don. *What People Want: The Concept of Utility from Bentham to Game Theory*. Cape Town: University of Cape Town Press, 1999.

Samuelson, Paul A. 'The Empirical Implications of Utility Analysis'. In *Econometrica 6.* Oxford: Blackwell Publishers, 1938.

Samuelson, Paul. 'A Note on the Pure Theory of Consumer Behaviour'. *Economica*, New Series, 5, 17(1938): 61-71.

Scanlon, Thomas. *What We Owe to Each Other*. Cambridge, Mass: Harvard University Press, 1995.

Scanlon, Thomas. 'Sen and Consequentialism'. *Economics and Philosophy* 17(2001): 39-50.

Sen, Amartya. 'The Impossibility of the Paretian Liberal'. *Journal of Political Economy* 78(1970): 152-157.

Sen, Amartya. `Behavior and the Concept of Preference'. *Economica*, 41(1973): 241-259.

Sen, Amartya. 'Rational Fools: A Critique of the Behavioral Foundations of Economic Theory'. *Philosophy and Public Affairs* 6, 4(1977): 317-344.

Sen, Amartya. 'Utilitarianism and Welfarism'. *The Journal of Philosophy* 76, 9(1979): 463-489.

Sen, Amartya. *Inequality Reexamined*. Cambridge, Mass: Harvard University Press 1992.

Sen, Amartya. 'Positional Objectivity'. *Philosophy and Public Affairs* 22 (1993): 126-145.

Sen, Amartya. 'Consequential Evaluation and Practical Reason'. *The Journal of Philosophy* 96(2000): 477-502.

Sen, Amartya. ‘Why Exactly is Commitment Important to Rationality?’ *Economics and Philosophy* 21(2005): 5-13.

Sen, Amartya. ‘Liberty and Social Choice’. In *Rationality and Freedom*. New Delhi: Oxford University Press, 2008.

Sen, Amartya. ‘Maximisation and the Act of Choice’. In *Rationality and Freedom*. New Delhi: Oxford University Press, 2008.

Sen, Amartya. *The Idea of Justice.* London: Allen Lane, 2009.

Skidmore, James. 'Skepticism about Practical Reason: Transcendental Arguments and Their Limits'. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 109, 2 (2002): 121-141.

Smith, Michael. *The Moral Problem*. Oxford: Blackwell, 1995.

Snowball, Jeannette and Willis, Kenneth. ‘Estimating the Marginal Utility of Different Sections of an Arts Festival: The Case of Visitors to the South African National Arts Festival’. *Leisure Studies* 24(2006): 43-56.

Sturgeon, Nicholas. ‘Ethical Intuitionism and Ethical Naturalism.’ In *Ethical Intuitonism: Re-evaluations*. Phillip Stratton-Lake. Oxford: Clarendon Press 2002.

Tversky, Amos and Kahneman, Daniel. ‘The Framing of Decisions and the Psychology of Choice’. *Science* 211(1981): 453-458.

Williams, Bernard. 'Internal and External Reasons'. In *Moral Luck*. Cambridge: Cambridge University Press, 1981.

Zemach, E. M. ‘Ought, Is, and a Game Called *Promise*’. *The Philosophical Quarterly* 21, 82 (1971): 61-3.