The Possible Incommensurability of Utilities and the Learning of Goals

Bruce Edmonds,
Centre for Policy Modelling,
Manchester Metropolitan University,
Aytoun Building, Aytoun Street,
Manchester, M1 3GH.

1. Introduction

This is a short article to examine the following possibility (and a couple of its corollaries):

> *that a single agent might simultaneously have different utilities that are incommensurable.*

That is, not just that the different utilities the agent is seeking are independent or even conflicting but that there is *no* meaningful mapping from these different utilities to a single utility function, even a single ordinal utility ranking. In other words that the utilities may be of fundamentally different kinds so that any mapping to a single measure could only be made by losing the essential nature of one of them. This is in direct contradiction with the assumptions of many economic models e.g. [6] and goes beyond attempts to merely explain or fix intransitivity of choices (as in [1]).

For example, one can imagine an agent which wished to maximise both income and love. Now one can imagine that these two utilities could be incommensurable - that is that there might be *no* mapping between the two so that the agent could meaningfully translate amounts of love into income or *vice versa*. Now further suppose that in a certain situation there was no action that simultaneously maximised both indicators, but that there were actions that maximised one or the other - this is often the case. In such a situation there would be no *single* action that would represent an optimum. The agent might have to choose between a range of actions, some of which resulted in it gaining income and some which resulted in it gaining love. In other words there might be no single numerically valued utility function which the agent could be said to be attempting to optimise, i.e. it would not even be the case that the agent was acting *as if* it were seeking to optimise a single utility function.

2. Arguments Against the Possibility

One might seek to claim that in any particular situation for any particular agent there *must* be a mapping onto a single utility function, but it is entirely unclear why this should be necessarily true. This may be readily believable in some situations, but that is very far from the assumption that this must be possible in all situations.

It may also be argued that agents *do* choose actions in such situations. This argument can be put in two forms.

- *Firstly*, that this means that agents *must* reduce these to a single utility in order to have been able to make the decision. That this is not a strong argument can be shown by merely considering an alternative mechanism (e.g. as described in [7]).

- *Secondly*, that these choices define the mapping to a single utility function in this particular context. This however begs the question. The purpose of the utility function is to *explain and predict* the choice of action. So if the utility function is explained in terms of the actual choice made, we are left with a circular explanation. Thus this is useless unless the explanation in one context can be used to predict the motivation or action in another context and although this may be true for certain restricted domains there is no evidence that this holds in general[1].

This leaves the possibility that the direction of explanation was intended to be in the reverse direction, i.e. that the observed actions were being used to explain the motivations of the agent. In this case, one is left with the question: "Why assume that the motivation can be reduced to one utility function?". There are several possible answers.

- *Firstly*, simply because it is the simpler theory (a la Occam). This is an acceptable answer if the alternatives are equally plausible but I will argue otherwise below - namely that the existence of practically incommensurable utilities *does* explain some observed characteristics of action where a single utility theory would not.
- *Secondly*, for meta-theoretical reasons, for example the fact that in many situations the assumption that the action of agents can be characterised by the optimization of a single utility (or a set of utilities that can be practically reduced to a single utility) enables one to prove desirable outcomes such as single equilibria. In this case the strength of the assumption lies in the extent to which these outcomes are actually observed (otherwise one is merely studying an obscure branch of mathematics and not anything which could be called a natural science) and the evidence for many of these in many situations is, to say the least, weak.
- *Thirdly*, because many different goods and many kinds of wealth *are* routinely mapped into the single measure represented by money. However, this does not mean that the *motivation* for the decisions that agents make are reducible to a single utility function, just that for some kinds of transactions it is the only *medium* of exchange practically available. For example, it occurs that firms agree to exchange research intelligence rather than buy or licence it from each other. Now while one can legitimately delimit the subject of economics to transactions to do with money, this does not mean that these transactions can be adequately explained or predicted in terms of a single utility in monetary terms. For example marriage can have great monetary consequences but this does not mean a decision to marry can be adequately determined by a process of two individuals each seeking to maximise effectively single utilities[2].

In order to make this possibility of incommensurability more real, let us consider an example. Imagine two people playing chess. Each player's goal is to win and not lose. In the early stages of the game the players have to make decisions as to choices of action without there being any possibility of being able to work out all the consequences of their actions. However, this does not mean that they make arbitrary moves – typically they will judge the possible chess positions using a number of ad hoc *indicators* (number of pieces, how far up the board their pawns have progressed, how many central squares they control, time they have left on the clock etc.). Since the goal of winning is too abstract to motivate any

1. Of course one may choose to *define* economics as the study of those situations where this holds - but then one is excluding many situations that are relevant to economic policy makers, businesses and individuals inter-acting in a monetary way. One has justified the assumption of a single utility function at the expense of the relevance of a whole field.
2. The classic answer to this is that one must include *all* the benefits into the utility function, but this merely begs the question of whether this is possible - which is the point of this argument.

planning at this stage of the game, maximising these indicators take on the role of effective goals (ones that plans and actions can be meaningfully traced back to).

Now, is there any meaningful sense in which these indicators can be said to be commensurable? Could the actions of these players (at this stage of the game) be characterised in terms of a model of optimising a single utility function? I think the answer to both questions is "no". While it is true that a certain player in a certain position may judge there to be a certain trade-off between these various indicators, this trade-off may vary substantially for different positions, and against different players. It might even vary between different matches even if the game has reached the same position against the same player! This strongly suggests that the particular trade-off is *result* of the decision making process rather than representing its *motivation* in any meaningful way. Likewise a model of decision making in terms of a process of maximising a single utility that is so context-dependent is unlikely to have any explanatory or predictive value.

3. Practical Examples

That this is not such a far-fetched example for economics, consider the example where firms effectively engage in complex marketing games against each other, successively placing products and advertisements, where the aim is merely to survive. They may well use a variety of ad hoc indicators as to the extent they will avoid extinction (like profit, predictability of cash-flow, customer loyalty, turn-over, diversity of product range etc.) but it is likely that any firm which simplifies their strategy down to a single indicator will not do very well, however sophisticated it is.

Now quite apart from the *theoretical* possibility of mapping the agents' utilities onto one utility, it is even less sure that this is remotely *practical*. The trade-offs between the different utilities might vary so critically from context to context that it would be foolish for the agent (or us) to model it by mapping them onto a single utility function (even a single ordinal utility ordering). Thus *even if* one supposed that one must be able, in theory, to map these utilities onto a single utility, this might well be beyond the capabilities of the agent or us in many contexts.

4. Implications for Goals

Beyond these negative considerations about the weakness of the assumption of an effectively single utility function, the introduction of multiple and effectively incommensurable utilities *does* explain some aspects of actually observed decision making behaviour, namely the learning of goals and the discontinuity of action as compared with stimulus.

If one is trying to simultaneously maximise more than one indicator and these are incommensurable, then this does not represent a goal in the sense of something that is a *direct* guide to action. In this case an agent in such a situation will have to *learn* its goals, or in other words, to model its indicators. This is a qualitatively different process from *sub-goaling*. In sub-goaling, one is looking to reduce the main goal to more less ambitious sub-goals by systematic means, whereas in modelling indicators one might accept approximate and partial realisations of the indicators.

This is particularly relevant when one changes goals. If you are using the single utility optimization model you have to change the utilities, in this model one can induce a new goal. For example[3], if one were a soldier who had certain aims: to reach a certain location in addition to some other aims such as

3. Taken from [3].

survival, and harming the enemy and it came to pass that the first aim was impossible then it makes more sense to say that we have *changed the goals* rather than that the utility of reaching the location has changed[4].

It sometimes happens that the actions of an agent (or even groups of agents) can change unboundedly quickly, even when faced with a smooth and slowly changing environment (e.g. the stock market crash of 1987 which occurred in the apparent absence of any significant economic news). Indeed sometimes a continuous, slowly changing and uni-directional environmental aspect can cause an individual to “flip” back-and-forth between alternative courses of action. This is difficult to model using a single utility function optimization procedure - you have to choose a highly complex and counter-intuitive function to capture it. Using the model of learning goals this becomes much simpler - the agent could have two different equally-successful models of its goals - the choice of which is fairly arbitrary. The flipping could then be explained by merely postulating that the agent was “trying out” these alternative models, finding neither satisfactory for any length of time.

One possible mechanism for effective decision making in the face of incommensurable goals is given in [7] in terms of a model of “deliberative coherence”. Another possibility is that such goal learning could take place in a more general framework where learning is represented by a process of modelling by agents, as described in [5].

5. Conclusion

There is no a priori reason to support the assumption that agents act as if they maximise a single utility. Further if one assumes that agents do act as if they have different utilities that are not meaningfully mapped onto a single utility function, then this *does* explain some of the observed behaviour of agents including some of those who are interacting in situations of monetary transaction.

Thus support for this assumption can only come from empirical evidence which we deliberately have not touched on here. We will only note that, at best, such evidence is highly equivocal (for a recent example see [2]) and so is not able to preform a strong justificatory role for the assumption at the present time.

4. “Hasn’t” given any intuitive reading of the term ‘utility’, that is. One could say that impossible goals can be modelled by either attributing it a zero utility or an infinite cost. However this abstracts the meaning of the term away from any conception of usefulness and changes the single utility optimization model into more of an abstract “black-box” model. In this form it would cease to correspond to how agents actually make decisions - in which case there would be no reason not to attribute multiple incommensurable utilities if this was convenient.

References

[1] Kirchsteiger, G. and Puppe, C. (1996). Intransitive Choices Based on Transitive Preferences - The Case of Menu-Dependent Information. *Theory and Decision*, 41, 37-58.

[2] Li, S. (1996). An Additional Violation of Transitivity and Independence between Alternatives. *Journal of Economic Psychology*, 17, 645-650.

[3] Millgram, E. (1995). Rational Goal Acquisition in Highly Adaptive Agents. AAAI Fall Symposium on Rational Agency, Cambridge, MA, 1995.

[4] Millgram, E. (forthcoming). Incommensurability and Practical Reasoning. In Chang, R. (ed.), Incommensurability and Value, Harvard University Press, Boston, MA.

[5] Moss, S. J. and Edmonds, B. (forthcoming). Modelling Economic Learning as Modelling. *Systems and Cybernetic*s.

[6] Von Neuman, J. and Mogenstern, O. (1947). *The Theory of Games and Economic Behaviour.* Princeton University Press, Princeton.

[7] Thagard, P. and Millgram, E (1997). Inference to the best plan: A coherence theory of decision. In Leake, D. and Ram, A. (eds.). *Goal-Driven Learning*. MIT Press, Cambridge, MA, 439-454.