Preference Utilitarianism by Way of Preference Change?

Rabinowicz, Wlodek

Published in:
Preference Change: Approaches from Philosophy, Economics and Psychology

2009

Preference Utilitarianism by Way of Preference Change?[*]

Wlodek Rabinowicz

In this paper, I revisit Richard Hare's classical and much discussed argument for preference utilitarianism (Hare, 1981). The argument, which relies on the conception of moral deliberation as a thought-experimentation process with concomitant preference change, is problematic in several respects. Here, I shall mainly focus on one of these difficulties: an apparent gap in Hare's reasoning, which might be called The No-Conflict Problem. In a paper I wrote with Bertil Strömberg, we try to fill this lacuna (Rabinowicz & Strömberg, 1996). The way we do it, however, is not fully satisfactory, as I have come to realize. Below, I shall therefore also consider an alternative solution, which was suggested to me by Daniel Elstein.

In my paper with Strömberg, we also examine whether the gap is there to begin with: The problem should perhaps be *dis*solved rather than solved. This suggestion goes back to an idea of Zeno Vendler (Vendler 1988). Unfortunately, it turns out that Vendler's move does not save Hare from criticism. It dissolves the No-Conflict Problem but gives rise to another, potentially more serious difficulty.

1. The argument and the gap

Hare's argument rests on his interpretation of moral judgements as *universal overriding prescriptions*.[1] By the principle of universalizability, what a moral judgment prescribes for a given situation is also prescribed for the hypothetical variants of that situation, in which the individuals' roles have been reversed. In order to reach a moral judgment regarding the situation at hand, I must therefore take all these variants into consideration. To illustrate, suppose I contemplate an action that affects not only myself, but also some other persons, say, John and Mary. The claim that I ought to perform the action would amount to prescribing it both for the situation I am in *and* for the hypothetical situations in which I would be on the receiving end instead. To reach a moral judgment, I have to ask myself: What if I were in John's or Mary's shoes? How would it be like to be subjected to the action if I had been as he or she is now? Because of their universal

[*] I am indebted to Daniel Elstein, Christian List, Toni Rönnow-Rasmussen and Bertil Stömberg for helpful suggestions. An earlier version of this paper was presented at a workshop on belief change in Berlin 2006, arranged in connection with the meeting of Gesellschaft für Analytische Philosophie. I wish to thank the organizers, Till Grüne-Yanoff and Sven Ove Hansson, and the participants of this event.

[1] See, for example, Hare (1981), p. 55.

application, moral judgments must be based on, or tested by, this kind of thought-experiments.

”Being in somebody else’s shoes” is, by the way, a somewhat misleading expression in this context. When I imagine being as John is now, I must assume that, in this hypothetical situation, I not only take over John’s external circumstances but also his body, his psychological make-up, his character, beliefs, emotions and desires - not just the shoes but also what they are sitting on. I try to imagine how it would be like to be exposed to the action if I were just as he is in the actual situation. As Hare puts it:

> … if I now say that I ought to do a certain thing to a certain person, I am committed to the view that the very same thing be done to me, were I exactly in his situation, including having the same personal characteristics and in particular the same motivational state. (Hare 1981, p. 108)

To make the discussion that follows more concrete, let me add some detail to the example. Suppose I have agreed to meet John and Mary, two of my students, at the department today. We haven’t fixed any definite time for the meeting but the secretary phones me at home with the message that the students have already arrived and are waiting. Since the weather is beautiful, I would much prefer to go by bike to the office rather than to drive. The students, on the other hand, dislike waiting: They would prefer that I arrive as soon as possible. So, our preferences are in conflict. The preference-utilitarian solution would prescribe the action that best satisfies the aggregated preferences of the persons involved. The issue is decided by balancing: My preference for going by bike is weighed in proportion to its strength against the students’ opposing preferences. Suppose that each of the latter is weaker than my own but that together they weigh more. Under these circumstances, I ought to abstain from going by bike and take the car instead.

Preference utilitarianism pre-supposes that the strength of people’s preferences can be compared and measured on a common scale. Obviously, this is a highly contentious claim, but as Hare takes it more or less for granted (see ch. 7 in Hare, 1981, esp. p.124), I am going to do likewise, at least for the argument’s sake. Suppose then that the strength of my preference for going by bike is +4, while the intensities of John and Mary’s opposing preferences are -3 and -2, respectively. The signs, plus or minus, specify the direction of a preference - whether it is for or against the action under consideration. The preference-utilitarian calculus implies that I should abstain from the bike alternative: $+ 4 - 3 - 2 < 0$.

Now, Hare tries to establish that I will reach the same result if I seek to arrive at a moral judgment *via* thought-experiments in which I take on the positions of the different persons involved in the situation at hand. If I proceed in this way, and if I am rational,

well-informed and equipped with sufficient imagination, then I cannot avoid arriving at the same prescription as the one delivered by preference utilitarianism.

How does Hare describe the process that leads me towards a moral judgment? Let me start with a quote from *Moral Thinking*. There, he discusses a 'bilateral' example in which I – the subject - consider whether to move someone else's bicycle in order to create a parking space for my car. No other persons are involved. Preference utilitarianism implies that I ought to move the bicycle if, and only if, my preference for this action is stronger than the cyclist's preference against his bicycle being moved. Hare comments:

> … if I now say I ought to do a certain thing to a certain person, I am committed to the view that the very same thing ought to happen to me, were I in exactly his situation, including having the same personal characteristics and in particular the same motivational states. But the motivational states he actually now has may run quite counter to my own present ones. […] Suppose, for example, that all I think I ought to do to him is move his bicycle so that I can park my car, and he has a mild aversion to my doing this […] I can see no reason for not adopting the same solution here as when we do in cases when our own preferences conflict with one another. For example, let us change the case and suppose that it is my own bicycle, and that it is moderately inconvenient to move it but highly inconvenient not to be able to park my car; I shall then naturally move the bicycle thinking that that is what, prudentially speaking, I ought to do, or what I most want, all in all, to do. Reverting now to the bilateral case: we have established [section 5.3, pp. 94-6] that, if I have full knowledge of the other person's preferences, I shall myself acquire preferences equal to his regarding what should be done to me were I in his situation; and these are the preferences which are now conflicting with my original prescription [to move the bicycle]. So we have in effect not an interpersonal conflict of preferences or prescriptions, but an intrapersonal one; both the conflicting preferences are mine. I shall therefore deal with the conflict in exactly the same way as with that between two original preferences of my own.
>
> Multilateral cases [in which several persons are affected] now present less difficulty than at first appeared. For in them too the interpersonal conflicts, however complex and however many persons are involved, will reduce themselves, given full knowledge of the preferences of others, to intrapersonal ones. (Hare, 1981, sections 6.1 and 6.2, pp. 108ff)

Let us try to unpack this complex passage, now using the example that we have started with. I contemplate going by bike to the office in the situation at hand, call it s_1. I have a preference *for* this action, with strength 4. However, since moral judgments are universal, they prescribe exactly similar things for exactly similar situations. Consequently, a moral judgment concerning what I ought to do in s_1 would also apply to the hypothetical situations in which the roles were reversed. Therefore, I need to imagine being in John's shoes and in Mary's shoes, respectively. Thus, I envision two hypothetical situations, s_2 and s_3, in each of which I am on one of the receiving ends. I realize that if I were in John's position, with his desires etc., I would have the same preference as John has in the actual situation: *against* the action in question, with strength 3. Analogously, were I in Mary's shoes, I would have a preference *against* the action, with strength 2.

The next step in the deliberation process pre-supposes what Allan Gibbard has called the Principle of Conditional Reflection (Gibbard 1988). Hare himself introduces that principle without giving it any label.

Conditional Reflection: Insofar as I fully know what I would prefer *in* a hypothetical case, I must have the corresponding preference (same sign, same strength) *with regard to* that hypothetical case.[2]

In other words, my hypothetical preferences – if I know I would have them in a hypothetical case –are reflected in my actual preferences with regard to the case in question. Insofar as I now come to see that I would disprefer the biking alternative if I were in John's position, I acquire a preference against this action with regard to that hypothetical situation.

Hare takes Conditional Reflection to be a conceptual truth. The principle holds due to the alleged presence of a prescriptive element in the very concept of 'I': "The suggestion is that 'I' is not wholly a descriptive word but in part prescriptive." (*ibid.*, p. 96).

> … by calling some person 'I', I express at least a considerably greater concern for the satisfaction of his preferences than for those of people whom I do not so designate. Thus, in a normal clear-cut case, if I were asked, when somebody is being maltreated and dislikes it, 'How do you feel about being put forthwith in that situation with his preferences', I shall reply that if it would be *me*, I do now have the same aversion to having it done as he now has. (*ibid.*, p. 98)

The idea is that in thinking of hypothetical preferences as *mine*, I thereby endorse them. Is it a convincing claim? One might doubt this: Such endorsement does not seem to be operative when I consider hypothetical cases in which my preferences by my present lights would be corrupted or distorted in some way. If I had a sadistic disposition, I would wish to cause pain. But knowing this doesn't make me wish to cause pain if I were a sadist: I don't endorse my hypothetical preference if I now judge it to be corrupted or irrational.

This objection suggests that Conditional Reflection should at least be qualified in some ways in order to be acceptable. In addition, perhaps it should be interpreted as a requirement of rationality rather than as a conceptual truth. We might think of it as a condition on ideally self-integrated and self-confident preferrers.[3]

---

[2] Cf. Hare (1981), p. 99: "… I cannot know the extent and quality of others' suffering and, in general, motivations and preferences without having equal motivations with regard to what should happen to me, were I in their places, with their motivations and preferences." The same principle has also been called "The Principle of Hypothetical Self-Endorsement" (in Persson 1989), and "The Principle of Conditional Self-Endorsement" (in Rabinowicz 1989 and in Rabinowicz & Strömberg 1996).

[3] As is easily seen, Conditional Reflection, which is a constraint on preferences, is closely related to the well-known reflection principle for beliefs. According to the latter, knowing what one would believe in a hypothetical situation commits one to analogous and equally strong conditional beliefs - conditional on the obtaining of the situation in question. Bas van Fraassen has shown that a person who violates that principle is vulnerable to a Dutch Book, provided only that she assigns some positive probability to the hypothetical situation in question (cf. van Fraassen 1984). In Rabinowicz (1989), I suggest that a similar Dutch Book argument might be available for an analogous reflection principle for preferences. However, the probability of my occupying exactly the same position as someone else occupies in a situation at hand is zero. Therefore, it does not seem possible to set up a Dutch Book against someone who violates

But, at least for the time being, we can leave this principle as it stands. It is in any case clear that the ”distortion” objection does not apply in our example: The students' preferences in this example are perfectly reasonable.

Conditional Reflection implies that, after having considered what it would be like to be in my students' shoes, I end up with a set of preferences as regards the contemplated action – as many as the number of the situations I have considered. I still have my original preference for the bike alternative with strength 4, but now - after having considered the hypothetical situations s_2 and s_3 - I also acquire two preferences against this action, with strengths 3 and 2, respectively.

Hare seems to suggest that the last step in the process of arriving at a moral judgment involves *preference balancing*. Here I am, with preferences that pull me in opposing directions –towards the action and away from it. My rational preference “all in all”, as he puts it, is a function of these preferential inputs. In our example, this means that, all things considered, I come to prefer not to go by bike, despite the fact that I have originally preferred that action: My preferences against it are jointly stronger. By engaging in thought-experiments that lead me to acquire new preferences and then by balancing these against my original preference, I thus seem to have reached the same solution as the one delivered by preference utilitarianism. Hare is anxious to point out that his re-construction of the process of moral deliberation transforms the original *inter*personal preference conflict into an *intra*personal one. The latter conflict is then solvable in the standard way - by straightforward balancing.

Now, as Shueler (1984) and Persson (1989) have pointed out, this argument – as presented above – contains an important gap. Hare's comparison with standard decision problems in which the subject experiences a conflict of preferences is misleading. In the standard case, my conflicting desires concern one and the same situation – the one in which I am about to act. In Hare's argument, however, the various preferences I have acquired via thought-experiments are not related in this way. I have a preference for going by bike with regard to the actual situation s_1, a preference against this action with regard to the hypothetical situation s_2, in which I am in John's shoes, and yet another preference against this action with regard to s_3, in which I occupy Mary's position. These desires of mine concern different situations and for that reason they do *not* oppose each other. Unlike in the standard case, there is here no conflict of preferences to begin with, which would need to be solved by balancing. Thus, suppose I were to decide to go by bike to the office. This action would satisfy my preference as regards the actual situation

---

Conditional Reflection with regard to hypothetical cases that are considered by Hare in the course of his argument for preference utilitarianism.

s_1, but it would in no way frustrate my preferences regarding the purely hypothetical situations s_2 and s_3. This, in a nutshell, is the 'No-Conflict Problem' that threatens Hare's argument.

2. Preference extrapolation

But haven't we forgotten something? We have omitted to make use of a crucial premise – the principle of universalizability. Universalizability requires that my prescription with regard to the different situations under consideration, s_1, s_2 and s_3, must be *uniform* in order to be moral. Thus, as long as long as my preferences regarding s_1 differ from those regarding s_2 and s_3, I haven't yet arrived at a moral judgment. If I understand him correctly, Hare suggests that the uniform prescription can be reached by a process of *tentative extrapolation*: I try to extrapolate my preference regarding one situation to other situations. The question is then whether the extrapolated preference is strong enough to survive any conflicts of preference that might be created by this move. If it is not, then I try to extrapolate one of my other preferences instead – one of those I entertain with regard to the situations in which the roles are reversed. Can this tack help us here?

The extrapolation manoeuvre does help, but only when we deal with *bilateral* cases. If there is just one student, say John, who is waiting for me at the department, I only have two situations to worry about, the actual situation s_1 and the hypothetical situation s_2, in which I am in John's shoes. Now, I can successfully extrapolate my preference for going by bike from s_1 to s_2, since this preference is stronger than my opposing preference regarding s_2, which I have acquired in accordance with Conditional Reflection. Had the latter preference been stronger, then I would be able to successfully extrapolate that preference instead. Consequently, I can uphold a uniform prescription with regard to both situations and the prescription is of the right utilitarian kind.

In fact, this is how Hare himself deals with the bilateral example of a motorist who contemplates moving another person's bike in order to make room for his own car: The motorist's preference is stronger than the opposing preference of the cyclist. Therefore, the former can successfully extrapolate his preference for moving the bike to the hypothetical case in which he is in the cyclist's shoes: The preference against moving the bike, which the motorist holds with regard to that hypothetical case, is weaker. Therefore, upon extrapolation, the former preference is victorious and a uniform prescription can be upheld. (Cf. Hare 1981, p. 110.)

However, the proposed solution would lead us astray in multilateral cases (cf. Persson 1989). Thus, consider again the example with two students who wait for me in the department. It is easily seen that my preference for biking as regards s_1 can be successfully extrapolated to both s_2 and s_3. It is stronger than each of the opposing

preferences I have regarding these two situations, even though it is weaker than both of them taken together. The extrapolated preference wins because it only meets one opposing preference at a time. Opposing preferences never have an opportunity to join forces, so to speak. The uniform prescription that I go by bike therefore remains undefeated, even though it obviously is counter-utilitarian in its spirit. This means that the extrapolation manoeuvre is inappropriate in the multilateral cases.

Persson (1989) suggests that the gap in Hare's argument might instead be filled in by introducing a "veil of ignorance" - a device that has been made famous by John Rawls and John Harsanyi. Persson's veil of ignorance is essentially the same as Harsanyi's in the latter's "equiprobability model" (see Harsanyi 1953 and 1977): After having acquired preferences concerning the three situations s_1, s_2, and s_3, I should now pretend that I am ignorant as to which of these three situations is the actual one. I should treat them as though they were equiprobable and then apply the standard principle of expected utility maximization in order to identify the action to be performed.[4] As a result, I will abstain from going by bike to the office. This action would satisfy my preferences if s_1 is actual, but it would frustrate them if one of the other two situations obtains instead; which is twice as probable, given my pretence of ignorance.

Persson's solution has not been adopted by Hare, who unlike Rawls and Harsanyi wants to avoid any elements of pretence, of make-believe in his reconstruction of moral reasoning. As Persson himself points out, the veil-of-ignorance approach would represent an alien element in Hare's thought, given Hare's project to base ethics on rational grounds:

> "… the addition of PEP [= The Principle of Equal Probability] to Hare's premisses appears highly problematic, since while rationality [...] demands that preferences be formed on the basis of all relevant information available to one, PEP requires one to seal oneself off from certain pieces of information (concerning the numerical identity of the particulars involved). (Persson 1989, p. 170)

One might also put it like this: pretence in, pretence out. With premises we only pretend to accept, the conclusion wouldn't be accepted for real. Therefore, in his comments on Persson's paper, Hare tries to fill the gap in a different way (cf. Hare, 1989). To save space, I won't discuss that proposal. Let me just say I find it quite unsatisfactory.[5]

---

[4] Harsanyi thought that such pretence of ignorance was appropriate for an ideal observer. It is more difficult though to understand how such pretence could even in principle be possible in the context of practical moral deliberation: When I deliberate whether I ought to perform an action, I cannot at the same time pretend to myself that I might be at the receiving end of the action I consider performing!

[5] For a critical discussion, see Rabinowicz & Strömberg (1996), section 3.

3. Preference Revision

Instead, let me move to the proposal outlined in my paper with Strömberg. Go back to the point at which my thought experiments have led me to acquire a set of preferences concerning the three situations, s_1 - s_3. My preference profile with respect to the action under consideration can now be represented by a vector,

(+4, -3, -2)

in which the first component specifies the strength of my preference regarding s_1, the second specifies the strength of my preference regarding s_2, and so on. The signs, plus or minus, specify the direction of a preference – whether it is for or against the contemplated action. On the basis of this profile, I must now arrive at a moral judgment, i.e., a universal prescription: I must come to prescribe the same action – either to go by bike or to abstain – for each of the three situations.

The main idea behind our proposal may be formulated as follows: The universal prescription to be reached should agree as much as possible with the subject's original preference profile. This idea can be made more precise in several ways, two of which we outline in our paper. We distinguish between what we call the "preference revision" approach and the "final verdict" approach. Here I will only focus on the former.

Prescribing and preferring are for Hare essentially the same thing. "[A]ll prescriptions, including moral ones, are expressions of preferences or of desires in a wide sense" (Hare 1981, p. 185, cf. also p. 107)[6] Thus, when I try to arrive at a uniform prescription for the three situations in our example, what I am after is a uniform preference with regard to these situations.[7] In other words, I try to revise my original preferences, which differ with respect to the three situations, in order to reach a new preference state with a uniform profile:

(x, x, x)

In this vector, the same (positive or negative) value appears at each place. In the preference state I am after, I have exactly the same preference (for or against the action) as regards each of the three situations, s_1 - s_3.

---

[6] See also Hare (1987), p. 73: "To want something to happen is to be in a state of mind which, if it had to be put into words, could be expressed by saying that one accepts the prescription that it happen." This idea goes back to Hare (1963), section 9.4.

[7] "To accept a universal prescription" is consequently the same as "to form a universal preference" (Hare 1989, p.172). This identification of a universal prescription with a universal preference was also assumed by the extrapolation manoeuvre.

How should I proceed in order to change my preferences in this way? What is the appropriate value for *x*?

Preference revision may be seen as a process analogous to revision of *beliefs*. As the ruling principle of belief revision one usually takes the Principle of Minimal Change. When I have to revise my beliefs in order to make room for new information, or – more generally – in order to get them in line with some constraint I need to satisfy, I should be conservative: I should deviate as little as possible from what I have originally believed. To put it differently, the distance between my old beliefs and my new beliefs ought to be minimized given the task at hand. Cf. Gärdenfors (1988), p. 8:

> when evaluating changes of belief, we require that the change be the *minimal* one needed to accommodate the epistemic input that generates the change.

If Minimal Change is taken as the ruling principle for revision of preferences, then it follows that the uniform preference state to be reached by the subject should diverge as little as possible from his original non-uniform preference state. To paraphrase Gärdefors, we require that the change of preferences be the minimal one needed to satisfy the uniformity constraint that generates the change. Thus, the value for *x* should be chosen in such a way that the distance between the two preferences states be minimized.

But how are we to determine such distances? If one represents preference states as vectors, as we have done, then each state may be seen as a point in a vector space. A space point is describable by its numerical coordinates, which specify its position in the different spatial dimensions. In our example, we work with three-place vectors, i.e. with points in a three-dimensional space. Generally speaking, the number of dimensions is determined by the number of situations I – the subject – need to consider, i.e., by the number of persons involved in the actual situation. For each person, I need to consider a situation – actual or hypothetical – in which I would be in that person's shoes. Had the number of persons involved been smaller, say, just me and John, only two situations would need to be taken into account, instead of three, and my preference state would then be representable as a point in a two-dimensional space.

Given a vector space, what measure of distance between vectors is it appropriate to accept? It is clear that this measure should not be 'partial'. In particular, it should not favour the subject's preference with regard to the actual situation, s_1, at the expense of his preferences with regard to hypothetical situations, s_2 and s_3. Such partiality would clearly go against the spirit of universalizability that inspires Hare's enterprise. Thus, we take it that universalizability makes its appearance at two places in Hare's argument: first, as a uniformity constraint on the posterior preference state - as a demand that the posterior preference with regard to each situation be the same whatever position one is supposed to occupy in the situation in question; second, as an impartiality constraint on the distance

measure – as a requirement that the distance between points in a vector space be invariant under permutations of dimensions.

Consider an *n*-dimensional space of preference states. As we already know, *n* is the number of situations to be considered, i.e., the number of persons involved. *If* we suppose that the distance measure on that space is of the standard Euclidean type, one that we are used to deal with in other contexts, then the distance between two preference states, $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$, equals the square root of the sum-total of the squared differences between the corresponding components of *v* and *w*:

Euclidean distance: $[\Sigma_{i = 1, \ldots, n}(v_i - w_i)^2]^{½}$

This makes our task solvable: We can determine what value *x* must take if the Euclidean distance between the prior preference state and the posterior uniform state $(x, \ldots, x)$ is to be minimal.

It can be proved that Euclidean distance is minimized if *x* is the *average* of the values in the original preference profile.[8] This averaging solution is, of course, very much in the spirit of preference-utilitarianism: The average of the preferences I have acquired, in accordance with Conditional Reflection, with regard to the situations in which I occupy the positions of different individuals equals the average of the preferences these individuals entertain in the actual situation. And preference utilitarianism implies that the action ought to be performed if and only if the latter average is positive.[9]

Thus, in our example, the average of my preferences in the state

(+4, -3, -2)

equals -1/3. Consequently, if the right measure for the distance between preference states is Euclidean, the revised uniform state would be:

(-1/3, -1/3, -1/3)

This means that the moral prescription is to abstain from going by bike, just as preference utilitarians would have it.

4. Questions

[8] For the proof, see Rabinowicz & Strömberg (1996).

[9] This holds if the choice problem is binary, i.e., if the only alternatives are performing the action or abstaining. In such a choice, we may assume that positive preference for an action is mirrored by an equally strong negative preference for its alternative. For a discussion of choices between several action alternatives, see the next section. Note also that it doesn't matter whether we go by the average or by the sum-total of preferences, as long as we only consider situations with a fixed number of persons involved.

There are, of course, several controversial elements in this proposal. Here are some of the questions that would need to be examined:

(i) Questions about *prescriptions*: Is prescribing really the same thing as preferring? This seems to presuppose a rather simplistic view of our mental life. According to philosophers like Michael Bratman (see Bratman 1987), we should carefully distinguish between such mental phenomena as desires and intentions. Quite possibly, one might equally well argue that preference and acceptance of a prescription are distinct mental states. This would create problems for the proposal, which follows Hare in his treatment of moral prescriptions as universal preferences.

(ii) Questions about *minimal change*: Is the analogy between belief revision and revision of preferences justified? Is the principle of minimal change as plausible in the latter case as in the former? In the case of beliefs, conservatism in adding beliefs is grounded in a requirement of epistemic responsibility and conservatism in giving up beliefs is justified insofar as, *ex ante*, stopping to believe amounts to an epistemic loss: One needs to stop believing what one currently takes to be *true*. In the case of preferences, a corresponding justification is absent, unless preferences are interpreted on cognitivist lines, as beliefs that certain things (= objects of preference) are valuable. But a cognitivist account of preferences is highly questionable. An alternative justification might be found in a principle of mental economy. Changing preferences is not easy, and larger changes might be more difficult to bring about than smaller ones.

(iii) Questions about choices between *several alternative actions*: Often, the agent's choice problem isn't simply whether to perform a given action or to abstain. Instead, the task is to choose between a number of alternative actions. The subject's preferences with respect to different situations are then representable by a *matrix* rather than by a single vector: If the number of situations to be considered is n and the number of action-alternatives is m, a preference state can be represented as a matrix with n columns (one for each situation) and m rows (one for each action). The numerical values in each cell of the matrix specify preference intensities. Thus, the value that appears in the jth row in the ith column specifies the strength of the agent's preference concerning action j with regard to situation i. We may suppose that preference strength is measured on an interval scale. Values in the matrix are therefore invariant up to positive linear transformations. (We no longer need to determine whether the subject is for or against an action j as regards a given situation. It is enough to determine whether he prefers or disprefers that action to other alternatives, and by how much. Thus, the zero point of the scale may now be arbitrarily chosen.)

By universalizability, in order to take a moral stand, the subject needs to move from his prior preference state to a new one, representable by a matrix with *uniform rows*. I.e., for every action *j*, the row for *j* in the new matrix must have the same preference values for each situation: $(x_j, x_j, \ldots\ldots, x_j)$. In addition, by the principle of minimal change, the new matrix should deviate from the original one as little as possible. There is a natural way to generalize the Euclidean measure to distances between $n \times m$-matrices: We let the distance between two matrices be the square root of the sum-total of the squared differences between the values in the corresponding cells of the matrices in question. Given this generalization, it can be shown that the distance is minimized if and only if the preference value for each action *j* in the posterior matrix is the average of the preference values in *j*'s row in the prior matrix.[10] This means that the action that ought to be performed is the one that maximizes the degree of preference satisfaction for the persons who are involved in the situation at hand, just as preference utilitarians would have it.

In what follows, however, I shall for simplicity's sake revert to binary choice problems in which there is just one action that the agent has to care about. Thus, instead of distances between matrices, we only need to consider distances between vectors.

(iii) Questions about *distance measure*: Why suppose that the correct measure of distance must be Euclidean? Obviously, it's just one possibility among many. What are then the adequacy criteria for a 'reasonable' measure of distance between preference states? We have mentioned one such criterion, the requirement that the measure be impartial. Another plausible criterion is that the distance between two preference states, *v* and *w*, should be an increasing function of the absolute differences between the corresponding preference components in v and w. But these constraints by themselves do not take us very far.

The simplest distance measure one might use in this context is the so-called "city block"-distance, which goes by the sum-total of the absolute differences between vectors v and w on each of the *n* dimensions:[11]

City-block distance: $\Sigma_{i = 1, \ldots, n} \left| v_i - w_i \right|$

Such a measure, however, does not always yield a unique solution for the distance minimizing task. In fact, in the two-dimensional case, the averaging solution is *only one*

---

[10] For the proof, see Rabinowicz & Strömberg (1996).

[11] The name derives from the fact that, from the Euclidean perspective, this measure gives as the distance between two points the length of the shortest path from one point to the other *provided* that we always move along the axes. It is as though we were constrained to travel along city streets that form a regular cross-pattern.

of the infinitely many that are possible. If the original vector is of the form (v_1, v_2), then any x between v_1 and v_2 will fill the bill. Thus, for example, if the prior preference state is (+3, -2), the averaging solution would be (+½, +½). But the uniform vectors that minimize city-block distance from the vector (+3, -2) form a continuum that ranges from (+3, +3) to (-2, -2). If the number of dimensions is larger than two, using city-block distance might sometimes yield a unique minimizing solution, *but* there is no guarantee that this solution will be the averaging one. Here is an example in three dimensions: Suppose that the original vector is (+6, 0, -3). The uniform vector that minimizes city-block distance from (+6, 0, -3) is (0, 0, 0), while the averaging solution would be (+1, +1, +1). (Note, by the way, that (0, 0, 0) would still minimize city-block distance if we replace the first component in (+6, 0, -3) by any value higher than 6.) The conclusion is that if we opt for the city-block as our distance measure, the argument for preference utilitarianism doesn't go through. So, why is the Euclidean measure to be recommended?

These two measures, city-block and the Euclidean distance, are members of a large family of distance measures, all of which have the form:

Minkowski distance: $[\Sigma_{i = 1, \ldots, n} \mid v_i - w_i \mid^k]^{1/k} \qquad (k \geq 1)$

If the coefficient k equals 1, we get the city-block; if it is 2, we obtain the Euclidean distance; and so on. The higher k is, the greater weight is given to larger absolute differences between corresponding vector components, as compared to smaller differences. Only when k equals 1, as in the city block, all the differences between the components are weighted equally, independently of size. But already for $k = 2$, as in the Euclidean measure, the larger absolute differences are made 'disproportionately' larger by exponentiation, as compared with the smaller differences.[12]

Now, to give greater weight to larger differences between the corresponding components of preference states looks like a consideration of *fairness*: One thereby favours posterior states that show smaller variance in the extent to which they deviate from component preferences in the prior state. I.e., ultimately, one favours posterior states that show smaller variance in the extent to which they deviate from the preferences of the persons who are involved in the situation at hand. This gives rise to a puzzle. It is notorious that fairness considerations are alien to the utilitarian outlook. For a preference utilitarian, the only thing that matters is that the overall degree of preference satisfaction is maximized (the average or the sum-total; it doesn't matter which as long as the population is kept fixed). Whether this goal is accomplished by letting the preferences of

---

[12] At the limit, all weight is placed on the largest difference. A very simple distance measure that is sensitive only to the largest differences can be defined as follows: the distance between v and $w = \max\{|v_i - wi|: i = 1, \ldots n\}$.

some individuals to be frustrated to a much larger extent than the preferences of others is irrelevant: Achieving a fair distribution of preference satisfaction doesn't matter. So how can we explain that it is the Euclidean measure rather than the city-block that gives us the utilitarian averaging solution, if it is the former and not the latter that makes allowances for the considerations of fairness? I wish I knew the answer to this puzzling question.[13]

(iv) Questions about *Harean exegesis*: How faithful is this proposal to Hare's own formulation of his argument? Well, there is one big difference between the two: They implement the universalizability requirement in different ways. Preference extrapolation, which in Hare's argument functions as a universalization device, does not come into play in our proposal at all. Instead, it is replaced by preference revision, in which universalizability is implemented in two ways: as the uniformity constraint on the outcome of revision and as the impartiality constraint on the measure of distance. This also means that Hare's idea of arriving at moral prescriptions by transformation of interpersonal preference conflicts into intrapersonal ones is not preserved.

5. Simultaneous extrapolation

Can we reconstruct the argument in a way that's closer to the original? Let's again go back to the point at which I entertain a set of preferences with varying strengths and signs with respect to a given action: one regarding the actual situation and the remaining ones regarding the hypothetical situations in which the roles are reversed. As we remember, Hare's suggestion was that a uniform prescription can be reached at that point by a process of tentative extrapolation: I try to extrapolate my preference regarding, say, the actual situation to its hypothetical variants. If the extrapolated preference is strong enough to survive any conflicts of preference that might be created by this move, then I am home. If it is not, then I try to extrapolate one of my other preferences instead – one of those I hold with regard to the situations in which the roles are reversed. As we have seen, however, this proposal can only deal with bilateral cases, but not with multilateral ones.

An alternative would be to employ what might be called a *simultaneous preference extrapolation*. This suggestion is due to Daniel Elstein.[14] Let's illustrate how this

---

[13] I am indebted to Christian List for pressing this point. Note, by the way, that if we instead increase the role of fairness considerations to the limit, so to speak, and determine the distance only by the largest absolute difference between vector components (see footnote 12 above), then we don't get averaging either. The uniform vector that minimizes such 'minimax' distance from (6, 0, -3) is (1,5, 1,5, 1,5) and not (1, 1, 1), as averaging would have it.

[14] In private communication.

procedure is supposed to work in our example. I have acquired a preference concerning the action under consideration with regard to each of the situations s_1 - s_3. These preferences form a vector,

(+4, -3, -2)

To satisfy the universalizability requirement, I now simultaneously extrapolate each of my preferences in this profile to all the three situations. We can think of this step as a move in which each preference I have is universalized so as to become a moral prescription. I thus arrive to a complex preferential state in which each of the preferences in the state (+4, -3, -2) is now being entertained with respect to each situation:

(<+4, -3, -2>, <+4, -3, -2>, <+4, -3, -2>)

In this new state, the first component, i.e. <+4, -3, -2>, specifies my preferences regarding s_1, the second component, which is exactly the same, specifies my preferences regarding s_2, and so on. One might say that in this state I simultaneously accept three prescriptions that uniformly apply to all the three situations: one prescription *for* the action under consideration, with strength 4, and the other two *against* the action, with strengths 3 and 2, respectively.

But how is it possible to accept prescriptions that are mutually incompatible? How can I accept both that I ought to go to the office by bike *and* that I ought not to do so? The answer is that the relevant ought-judgments are *pro tanto*: Each of them reflects just one relevant aspect of the case. In other words, they prescribe or forbid an action *insofar as* it has this-or-that feature. Thus, going by bike is prescribed insofar I originally prefer this action regarding s_1, it is forbidden insofar I originally disprefer it regarding s_2, in which I am in John's position, and it also is forbidden insofar I originally disprefer it regarding s_3, in which I am in Mary's position.

Unlike oughts all-things-considered, *pro tanto* oughts are not overriding. The novelty of simultaneous extrapolation lies precisely in that it employs the universalizability requirement at an earlier stage than Hare himself: at the stage at which we do not yet commit ourselves, not even tentatively, to an overriding moral judgment all-told.

The remainder of the argument is unproblematic. In the state

(<+4, -3, -2>, <+4, -3, -2>, <+4, -3, -2>)

I have mutually conflicting preferences with regard to each situation s_1 - s_3. This intrapersonal preference conflict is then dealt with by straightforward balancing,

+4 – 3 – 2 = -1

Consequently, I end up with the same preference all-told with regard to each situation:

(-1, -1, -1)

My overriding moral prescription all-things-considered, which is reached by the balancing of moral prescriptions *pro tanto*, is thus that I ought not to go to the office by bike, just as preference utilitarianism would have it: The stronger preference loses against joined forces of the weaker preferences.

This reconstruction of the argument preserves Hare's conception of an ideal moral deliberation as a process in which

> the interpersonal conflicts, however complex and however many persons are involved, will reduce themselves, given full knowledge of the preferences of others, to intrapersonal ones. (Hare 1981, p. 110)

However, the simultaneous extrapolation approach departs from Hare's moral theory at one crucial point: with respect to the overridingness issue. To be sure, Hare himself registers the possibility of overridable moral judgments, but these are according to him always *prima facie*.[15] They seem to hold 'at first sight', but may turn out to be invalid, in a particular case, upon further reflection. It is in this sense that they can be overridden: They are based on *prima facie* moral principles that are general, but admit of exceptions.[16] As far as I know, Hare never considered the possibility of moral judgments *pro tanto*, which retain their weight and validity even in those cases when they are overridden (= outweighed) by other moral considerations.[17] Simultaneous extrapolation therefore requires that we go beyond Hare at this point.

While allowing for *pro tanto* oughts is as such unproblematic, it is less clear what on this approach justifies the step from a preference I entertain regarding some situation to its extrapolation – i.e. a corresponding *pro tanto* moral prescription. The answer cannot simply be that I am trying to reach a moral judgment, which requires universality. It's certainly true that I am after a universal prescription *all-told*, but on what grounds do I first frame universal prescriptions *pro tanto*? I don't really know how to deal with this issue.

6. Vendlerian twist

Let me now turn to Zeno Vendler's comments on Hare's argument. It seems that, if Vendler is right, we have all along been on a wild-goose chase. If he is right, the No-Conflict Problem is spurious. It should be dissolved rather than solved.

---

[15] Unless they are what he calls 'inverted commas' moral judgments, "implying merely that a certain act is required in order to conform to the moral standards current in society" (Hare 1981, p. 58). 'Inverted commas' moral judgments are not genuinely moral, since they lack prescriptive force.

[16] Cf. Hare (1981), p. 59f.

[17] For a distinction between *pro tanto* and *prima facie*, as applied to reasons, see Kagan (1989), p. 17.

As we remember, the problem in question arises because the thought-experiments needed for Hare's argument concern purely *hypothetical* situations. When I ask myself what it would be like to be in someone else's shoes, and what preference I now have regarding that situation, I am supposed to consider a hypothetical state of affairs, which differs from the actual one in the role I occupy. My preference regarding what is to be done in such a hypothetical situation does not, on the face of it, conflict with my preference with respect to the actual situation, and this is what the No-Conflict Problem is all about.

The whole picture would change, if – as one might argue – the envisioned situation is not really distinct from the actual one, Suppose that what I consider is still *the actual situation, but now viewed from the perspective of another person*. If this is the case, then the preference I form regarding what is to be done in that situation does conflict with the preference I have when I view the same situation from my own point of view. This would mean that the resulting preference conflict can be solved in the standard way - by balancing. The action to be prescribed is the one that satisfies my conflicting preferences to the largest extent.

That imagining being exactly like someone else is in the actual situation does not take us to another possible world is a view put forward by Vendler:

> If I imagine being you, I do not imagine 'transporting' something into your body, or 'mixing' two entities. What I do is assume, as far as it is in the power of my imagination, the coherent set of experiences corresponding to your situation (your 'Humean self', as it were). But, as Hume pointed out, there is no specific experience of an 'I' in that totality. Nor is there one in mine. The 'I' as such has no content: it is the empty frame of consciousness indifferent to content. Consequently, by constructing in my fancy the image of what I would experience in your situation, I *ipso facto* represent your experiences. (Vendler, 1988, p. 176)
>
> ... in fancying being you or Castro, I do not touch the world: I merely switch perspectives on it. This is the reason, by the way, for my maintaining throughout this paper that imagining being in exactly the same qualitative conditions as another person is the same thing as imagining being that person. (ibid., p. 182)
>
> ... there *seem* to be two different situations envisioned [...] Hare says this, 'Note that although the two situations are different, they differ only in what *individuals* occupy the two roles; their *universal* properties are all the same' (*MT* 111). 'No', I say, it is the same situation, with the same individuals; the only difference is which of them is I: in imagining being he, I imagine the same situation from a different perspective. (ibid., p. 178)

Vendler’s position is very attractive. It does seem reasonable to say that Hare’s thought experiments do not target new situations, objectively speaking. Instead, they only effect a shift of subjective perspective. What shifts from one position to another is not myself, the person I am, but only the ”transcendental I”, to use Vendler’s Kantian terminology, i.e. a mere frame of consciousness that in principle can be filled with any content. Remember, that when I imagine being as John is now, I am not only supposed to take over his

external circumstances but also his psychological make-up: his beliefs, emotions, desires, and so on.

Subjectively speaking, then, the situation changes when it is viewed from different perspectives. But objectively, it still is the same situation. But then, if I form preferences that reflect the motivations I ('the transcendental I') would have in different positions, all these preferences concern one and the same objective situation. As such, they can conflict with each other – the No-Conflict Problem is spurious.

Vendler himself thinks that this 'anti-metaphysical' move makes Hare's argument an easy sailing. However, it seems to me that the opposite may be the case. While the No-Conflict Problem disappears, we now get a new, more serious problem instead.[18] I still need to form preferences reflecting the ones 'I' would have in different positions; I need to entertain all these preferences together, from one and the same perspective, in order to balance them against each other. For this I have to rely on something like Conditional Reflection. But, and here comes the catch, that principle does not really seem to be applicable in contexts like this. Why not? Well, Conditional Reflection is an expression of *self-concern* – a fundamental attitude of caring for oneself that each well-integrated person is supposed to have. Self-concern applies not only to the actual situation; it also extends to hypothetical circumstances in which one might be placed. It manifests itself in the endorsement of the preferences one would have in hypothetical cases, just as Conditional Reflection has it.

Now, one might ask, who is it I am concerned about when I am concerned about *myself*? Is it "the transcendental I" – a mere frame of consciousness that can be filled with an arbitrary content – or is it rather a definite *person*, the person I am? If it's the latter, then self-concern has no role to play in the radical thought-experiments of the kind Hare invites us to consider. In these experiments, what I envision is really being someone else, a different person. So it is not the question of imagining oneself – the very person one is – being placed in some hypothetical circumstances. But then, if self-concern does not extend to transcendental 'perspective shifts', such shifts remain outside the domain of application of Conditional Reflection. This means that Hare's argument cannot go through, contrary to what Vendler might have thought. The preferences belonging to different subjective perspectives concern the same objective situation, but, if Conditional Reflection is inapplicable, they cannot be reflected in one perspective: They do not give rise to a co-existing set of preferences that are being entertained together, in one

---

[18] Cf. section 7 in Rabinowicz & Strömberg (1996).

preference state. Consequently, they do not give rise to an intrapersonal conflict that can be solved by balancing.

To conclude, if Vendler is right, then Hare's argument doesn't go through, because Conditional Reflection is not meant to apply to transcendental perspective-shifts. If he is wrong, on the other hand, or if thought-experiments with role reversals could be given a less radical reading than the one Hare has in mind, then such experiments would manage to take the subject beyond the actual situation to other, merely hypothetical situations in which he finds himself at the receiving end of the action under consideration. Then Conditional Reflection would apply, but we would face the No-Conflict Problem. Its solution requires one's preferences be universalized, which can be implemented either by the preference revision approach or by the device of simultaneous preference extrapolation. In the former, a moral judgment all-things-considered is arrived at directly, while in the latter it is only reached by mediation of moral judgments *pro tanto*. Both approaches depart from Hare's own presentation of the process of moral deliberation, but the indirect approach appears to be closer to Hare and somewhat less question-begging.

References

Bratman, M., 1987, *Intentions, Plans, and Practical Reason*, Harvard Univ. Press, Cambridge, Mass..

Gibbard, A., 1988, "Hare's Analysis of 'Ought' and its Implications", in D. Seanor and N. Fotion (eds.), *Hare and Critics: Essays on Moral Thinking*, Clarendon Press, Oxford, pp. 57 - 72.

Hare, R.M., 1963, *Freedom and Reason*, Oxford University Press, Oxford.

Hare, R.M., 1981, *Moral Thinking: Its Level, Method and Point*, Clarendon Press, Oxford.

Hare, R.M., 1987, "Why Moral Language?", in P. Pettit, R. Sylvan, and J. Norman (eds.), *Metaphysics &Morality*, Basil Blackwell, Oxford.

Hare, R.M., 1989, "Reply to Ingemar Persson", *Theoria 55*, pp. 171-7.

Harsanyi, J. C., 1953, "Cardinal utility in welfare economics and in the theory of risk-taking", *Journal of Political Economy 61*, pp. 434–435

Harsanyi, J. C., 1977, "Morality and the Theory of Rational Behaviour", *Social Research 44*, pp. 623-56; reprinted in A. Sen and B.Williams (eds.), *Utilitarianism and Beyond*, Cambridge Univ. Press, Cambridge, 1982, pp. 39 - 62.

Kagan, S., 1991, The Limits of Morality, Clarendon Press, Oxford.

Persson, I., 1989, "Universalizability and the Summing of Desires", *Theoria 55*, pp. 159-70.

Rabinowicz, W., 1989, "Hare on Prudence", *Theoria 55*, pp. 145-51.

Rabinowicz, W. and Strömberg, B., 1996, "What if I were in his shoes? On Hare's argument for preference utilitarianism", *Theoria 62*, pp. 95-123.

Schueler, G.F., 1984, "Some Reasoning about Preferences", *Ethics* 95, pp. 78-80.

van Fraassen, B., 1984, "Belief and the Will", *Journal of Philosophy 81*, pp. 235 -56.

Vendler, Z., 1988, "Changing Places?", in D. Seanor and N. Fotion, *Hare and Critics*, Clarendon Press, Oxford, pp. 171-84.