Ethics & Uncertainty ‘Behind the Veil’

January 2, 2018

Harsanyi’s Arguments for Utilitarianism

Harsanyi produced two different (but related) arguments for Utilitarianism. Both draw conclusions about how to aggregate well-being across people from premises concerning how to rationally evaluate prospects.

The earlier argument is from 1953, the later argument is from 1955.

1. **The Veil of Ignorance Argument.** Consider a range of social situations. Imagine you are offered a choice between them.

A *social situation* describes, for each person, how well-off they are.

People:	1	2	3	…	h
A	a_1	a_2	a_3	…	a_h
B	b_1	b_2	b_3	…	b_h
C	c_1	c_2	c_3	…	c_h
⋮	⋮	⋮	⋮	⋮	⋮

However, you make your choice “behind a veil of ignorance": you don’t know which life you will lead. Furthermore, you assign equal probability to each: $Cr\,(\text{I am Person }1) = Cr\,(\text{I am Person }2) = Cr\,(\text{I am Person }3) = \cdots = Cr\,(\text{I am Person }h) = \frac{1}{h}$.

Because your decision is made behind a veil of ignorance, it’s guaranteed to be *impersonal* and *impartial*.

You’re choosing between prospects. If you’re rational, you’ll weakly prefer one lottery to another if and only if the former gives you at least as great an expectation of your good as the latter.

For example, if you’re rational, you’ll prefer the lottery corresponding to A to the lottery corresponding to B if and only if:

$$\sum_i \frac{1}{h} \cdot g(a_i) > \sum_i \frac{1}{h} \cdot g(b_i)$$

$$\frac{1}{h} \cdot g(a_1) + \frac{1}{h} \cdot g(a_2) + \cdots + \frac{1}{h} \cdot g(a_h) > \frac{1}{h} \cdot g(b_1) + \frac{1}{h} \cdot g(b_2) + \cdots + \frac{1}{h} \cdot g(b_h)$$

Because the probabilities are all the same, they cancel out; and so you’ll prefer the former to the latter if and only if:

$$\sum_i g(a_i) > \sum_i g(b_i)$$

That is: the total sum of good in A is greater than the total sum of good in B.

In general, every rational person behind the veil of ignorance prefers X to Y if and only if the total sum of good in X is greater than the total sum of good in Y. *Claim:* If every rational person behind the veil of ignorance prefers X to Y, then X is better than Y.

The argument makes several claims that could be resisted:

(a) Does rationality require you to maximize the *expectation* of your *good*?

(b) Isn’t it rational for different people to have different attitudes toward the same potential life?

(c) Why should it matter what every rational person would prefer “behind the veil of ignorance"? (Why privilege what we prefer *ex ante* to what we would prefer *ex post*?)

2. **The Aggregation Argument.** Say that an ordering is *coherent* if it satisfies the axioms of expected utility theory. Consider the following principle.

Harsanyi took the relevant orderings to be *preferences*. Broome argues against doing so, and resurrects the argument using 'better than' instead.

Pareto: If two alternatives are equally good for everyone, they are equally good. If one alternative is as good or better for everyone and is strictly better for someone, then it is better.

Broome calls this *the Principle of Personal Good*. The Pareto Principle, traditionally, is put in terms of preference (not the good): "If everyone is *indifferent* between two alternatives, they are equally good. If everyone weakly prefers one alternative to another and someone strictly prefers it, it is better." Broome argues that the traditional Pareto Principle has problems that his version does not.

Harsanyi proves the following theorem:

The Aggregation Theorem. Assume that (1) everyone's personal goodness-ordering is coherent, (2) the general goodness-ordering is coherent, and (3) the general goodness-ordering satisfies Pareto. Then there are expectational utility-functions, $U_1, U_2, \ldots, U_h$, representing each person's goodness-ordering, and an expectational utility-function, W, representing the general good, such that for any prospect L:

$$\begin{aligned} W(L) &= U_1(L) + U_2(L) + \cdots + U_h(L) \\ &= \sum_i U_i(L) \end{aligned}$$

In other words, how good a prospect is generally is the *sum* of how good it is for each individual.

Should the general goodness-ordering be coherent in this sense? (Consider *Diamond's Example:* You should be indifferent between giving a cookie to Alice and giving it to Bob; but you should, for the sake of fairness, prefer flipping a fair coin to determine who gets the cookie. This (seemingly) violates the Sure-Thing Principle.

The Pareto Principle: ex ante *vs* ex post

While the Pareto Principle seems fairly plausible when applied to *outcomes*, it's less plausible when applied to *prospects.* Harsanyi's argument requires the stronger (less plausible) version. Is there a good argument for it?

Counterexample to *ex ante* Pareto:

	HEADS	TAILS	U_1	U_2
L_E	$\langle 4,4\rangle$	$\langle 2,2\rangle$	3	3
L_P	$\langle 5,1.5\rangle$	$\langle 1.5,5\rangle$	3.25	3.25

ex ante Pareto recommends L_P over L_E. But egalitarian considerations suggest that L_E is a better prospect than L_P because it is guaranteed to result in more egalitarian distribution. (In fact: if we value egalitarian distributions enough, L_E is guaranteed to result in a better distribution than L_P no matter how the world turns out to be.)

Some Arguments:

1. *The Argument from Presumed Consent.* Were you to ask everyone affected by your decision what they would like you to do, they would all want the Pareto-dominate option. If you know what everyone would want you to do, you should do it.
2. *The Argument from Composition.* Break your action up into sub-actions that each only affect one person. For each of these sub-actions, you should prefer performing it (irrespective of whatever else you might do). If you should perform each of these sub-actions (irrespective of whatever else you might do), you should perform them all. (Otherwise, there would be nothing you could such that, were you to do it, you would have done everything you ought to have done.) Therefore, you should perform an option if it *ex ante* Pareto-dominates all the others.

What if you know more about the situation than the people involved? What if they know more about their situation than you?

We won't prove this, but here's an example to bring out the basic idea. Suppose there are only two people. And consider the following three prospects:

	$\frac{1}{2}$	$\frac{1}{2}$	U_1	U_2
L_1	$\langle 5,5\rangle$	$\langle 5,5\rangle$	5	5
L_2	$\langle 10,0\rangle$	$\langle 0,10\rangle$	5	5

	$\frac{1}{10}$	$\frac{9}{10}$	U_1	U_2
L^*	$\langle 10,0\rangle$	$\langle 0,0\rangle$	1	0

By Pareto, $W(L_1) = W(L_2)$.

Thus, $f(\langle 5,5\rangle) = \frac{1}{2} \cdot f(\langle 10,0\rangle) + \frac{1}{2} \cdot f(\langle 0,10\rangle) = f(\langle 10,0\rangle)$.

Because $W(L^*) = f(\langle 1,0\rangle) = \frac{1}{10} \cdot f(\langle 10,0\rangle)$, $f(\langle 10,0\rangle) = 10 \cdot f(\langle 1,0\rangle)$.

So, $f(\langle 5,5\rangle) = 10 \cdot f(\langle 1,0\rangle) = 10$.

So, $W(L_1) = f(\langle 5,5\rangle) = 10 = 5 + 5$.

The short example proof relies on a couple of lemmas that I won't prove:

1. *Overall Good is a Function of Personal Good:* $W(L) = f(\langle U_1(L), U_2(L)\rangle)$.
2. *Anonymity:* For any x and y, $f(\langle x,y\rangle) = f(\langle y,x\rangle)$.
3. By convention, let $f(\langle 0,0\rangle) = 0$ and $f(\langle 1,0\rangle) = 1$