Ex-Post Average Utilitarianism Can Be Worse for All Affected

Johan E. Gustafsson and Dean Spears*

Draft: February 27, 2024 at 2:34 p.m.

ABSTRACT. According to *Ex-Post* Average Utilitarianism, prospect X is at least as good as prospect Y if and only if the expected average well-being is at least as great in X as in Y. Relative to the ex-ante approach of taking the average of peoples' expectations, the *ex-post* approach has the advantage of not needing well-defined expectations of well-being for contingent people — people who exist in some but not all states of nature. Nevertheless, we show that *Ex-Post* Average Utilitarianism can oppose the interests of all affected people. Moreover, we show this without relying on any comparisons of expectations of well-being for contingent people.

Average utilitarianism is the view that the value of a final outcome is equal to the average well-being in that outcome.[1] This view about how to evaluate final outcomes can be extended in several ways to also evaluate prospects — that is, probability distributions over possible final outcomes. The most straightforward way to do so is to let the value of a prospect be equal to the prospect's expected value. This view is called *Ex-Post* Average Utilitarianism:[2]

> *Ex-Post Average Utilitarianism* Prospect X is at least as good as prospect Y if and only if the expected average well-being is at least as great in X as in Y.

A rival to *Ex-Ante* Utilitarianism is *Ex-Ante* Average Utilitarianism, which

* We would be grateful for any thoughts or comments on this paper, which can be sent to *johan.eric.gustafsson@gmail.com*.

[1] Sidgwick (1907, p. 415) distinguished average and total utilitarianism, although he (1907, pp. 415–16) favoured the latter. The name 'average utilitarianism' comes from Rawls 1971, p. 599.

[2] Harsanyi 1985, p. 44; 1986, p. 57. Nevertheless, in his correspondence with Ng (1983, p. 168), Harsanyi seems to defend *Ex-Ante* Average Utilitarianism. This conflicts with Harsanyi's (1955, p. 313) commitment to the expectation taking for the social ranking.

assess prospects by the average of each person's expected well-being conditional on that person existing:[3]

> *Ex-Ante Average Utilitarianism* A prospect x is at least as good as prospect y if and only if the average conditional-on-existence expected well-being for possible people in x is at least as great as the average conditional-on-existence expected well-being for possible people in y.

An advantage of *Ex-Post* Average Utilitarianism over *Ex-Ante* Average Utilitarianism is that the former maximizes an expectation, which means it satisfies expected utility theory for general betterness.[4] Nevertheless, we show that *Ex-Post* Average Utilitarianism can oppose the interests of all affected persons. Moreover, we can show this without assuming any well-defined expectations of well-being for contingent people.

A standard objection to *Ex-Post* Average Utilitarianism concerns the addition of lives of negative well-being. Like every form of average utilitarianism, the *ex-post* variant is vulnerable to the Sadistic Conclusion:[5]

> *The Sadistic Conclusion* When adding people without affecting the original peoples' well-being, it sometimes can be better to add

[3] Ng 1983, p. 168.

[4] *Ex-Ante* Average Utilitarianism violates expected utility theory. For instance, *Ex-Ante* Average Utilitarianism violates statewise dominance (even in non-sequential choices). Consider the following case where there are two possible states of nature S_1 and S_2 with an equal probability.

	Prospect A		Prospect B	
	S_1	S_2	S_1	S_2
	$1/2$	$1/2$	$1/2$	$1/2$
Ann	5	Ω	6	2
Bob	5	1	Ω	2

In S_1, the average well-being is 5 in A and 6 in B, and, in S_2, the average well-being is 1 in A and 2 in B. So B is better than A in every state of nature according to average utilitarianism. But, according to *Ex-Ante* Average Utilitarianism the value of A is 4 and the value of B is 3; so A is better than B. (The same counterexample also works against other *ex ante* approaches that would aggregate peoples' conditional-on-existence expected well-being while ignoring probabilities of their existence — such as an approach that values each prospect according to the minimum, among possible people, of individual expected well-being conditional on existence.)

[5] Arrhenius and Bykvist 1995, p. 85 and Arrhenius 2000, p. 251. Parfit (1984, p. 406) also puts forward an objection based on the addition of lives with negative well-being in his two-hells case.

a number of people with negative well-being, rather than a number of people with positive well-being.

A potential response to this objection is that the average utilitarian can reject the meaningfulness of the concept of negative well-being. Negative well-being levels are plausibly those levels that are less good than the neutral well-being level.[6] But how are we to understand the neutral well-being level? One compelling idea is that the neutral well-being level is the level that is equal to the well-being level of non-existence. Yet it is dubious whether non-existing people would have any well-being at all.[7] And, crucially, *Ex-Post* Average Utilitarians are not committed to any comparisons of well-being between existence and non-existence.[8] So they may consistently deny the underlying assumption of the existence of negative well-being in these objections.[9] (Nevertheless, note that, although we claim that it is a dialectical advantage of our argument that it cannot be responded to by denying the existence of negative well-being, we do not here deny the existence of negative well-being.)

Another standard objection to average utilitarianism is the Egyptology objection.[10] It is the objection that, given average utilitarianism, the evaluation of to acts today may depend on how well off people were in ancient Egypt. If the ancient Egyptians were very well off, then the addition of a person with a certain level of well-being could be worse than not adding them (because they would lower the average) even though, if the ancient Egyptians were, instead, less well off, the addition of the person would be better (because they would raise the average). A potential response to this objection is that the average utilitarian could maintain that whether the well-being of the ancient Egyptians were high or low is morally relevant, so sensitivity to such facts is not a drawback.

A final standard objection to average utilitarianism is the utility

[6] Chisholm and Sosa 1966, pp. 247–8.

[7] Williams 1973, p. 87, Parfit 1984, p. 487, and Broome 1993, p. 77.

[8] See Harsanyi in Ng 1983, pp. 168–9.

[9] Note further that denying the existence of negative well-being also blocks many other conditions in population ethics, such as the Mere-Addition Principle (Parfit 1984, p. 420 and Ng 1989, pp. 537–8), which rules out Average Utilitarianism, and the Repugnant Conclusion (Parfit 1984, p. 388), which, in a variant that compares *additions* to populations, rules out Average Utilitarianism (Anglin 1977, p. 746 and Spears and Budolfson 2021, p. 574).

[10] McMahan 1981, p. 115 and Parfit 1984, p. 420.

monster.[11] The objection is that, even supposing that everyone in history would live very good lives, it would be better according to average utilitarianism if there had only been a single person that is just slightly happier. A weakness of this objection, in a dialectic between total and average utilitarianism, is that much the same objection also applies to total utilitarianism (if the single person is sufficiently happy to have more total well-being).

In this paper, we present a new objection to *Ex-Post* Average Utilitarianism, which avoids the drawbacks of the earlier objections.[12] Accordingly, the above responses to those objections do not apply to this new objection.

Consider the following prospects, where columns are risky states of nature, rows are people (including contingent people), and cells are outcomes which could include ordered well-being levels, represented by numbers, or non-existence, represented by Ω:

	Prospect A		Prospect B	
	S_1	S_2	S_1	S_2
	1/2	1/2	1/2	1/2
Ann	1	Ω	1	Ω
Bob	1	7	9	1

There are two possible states of nature S_1 and S_2 with an equal probability. In both prospects, Ann exists with a well-being of 1 in S_1 but she does not exist at all in S_2. Bob, on the other hand, exists in all states of nature in both prospects. In A, Bob has a well-being of 1 in S_1 and a well-being of 7 in S_2. In B, Bob has a well-being of 9 in S_1 and a well-being of 1 in S_2.

Part of the motivation for average utilitarianism is that it does not rely on comparisons of well-being between existence and non-existence. Indeed it is unclear whether a prospect could be better-for a person who would exist only in some possible outcomes and would not exist in others. We will say that a person is *affected* in a comparison between two prospects X and Y if and only if there is a state of nature in which that person's outcome (which could be either existence with a well-being level or non-existence) is different in X than in Y; otherwise, the person is *unaffected* in that comparison. Given this distinction, we can rely on a

[11] Nozick 1974, p. 41.

[12] Or, for a reader persuaded by these prior objections to Average Utilitarianism, our new example adds to their strength.

weakened form of *ex-ante*-Pareto dominance that only applies to people who are affected in a comparison and who are necessary in a comparison (that is, people who exist in all possible states):[13]

> *Necessity-Restricted All-Affected Stochastic Pareto* If (i) all contingent people are unaffected in a comparison between prospects X and Y and (ii), for each affected person, X stochastically dominates Y, then X is better than Y.

Notice that Necessity-Restricted All-Affected Stochastic Pareto does not invoke expectation-taking even for necessary people: It merely uses stochastic dominance.

Returning to our example, Ann is unaffected by a choice between A and B. But, for Bob, B stochastically dominates A. Hence, for the only affected person (in either existence or well-being), B is clearly better than A. So, according to Necessity-Restricted All-Affected Stochastic Pareto, B would be better than A. And yet, *Ex-Post* Average Utilitarianism entails that A is better than B, because the expected average well-being is 4 in A but just 3 in B.[14] Therefore, in the comparison of A and B, *Ex-Post* Average Utilitarianism opposes the interests of all affected persons.

This counter-example to *Ex-Post* Average Utilitarianism avoids the drawbacks of the earlier objections. It does not rely on negative well-being. Unlike the unaffected ancient Egyptians in the Egyptology

[13] It may be helpful to compare our approach to that of Thomas 2022, pp. 280–2. Thomas shows that Anteriority, combined with some further assumptions, rules out average utilitarianism. Here, Anteriority is the following principle (McCarthy, 2017, p. 226):

> *Anteriority* If every potential person faces the same prospect in X as in Y, then X and Y are equally good.

(We have added 'potential' to highlight that the involved people need not exist.) An advantage of relying on Necessity-Restricted All-Affected Stochastic Pareto rather than Anteriority is that the former (unlike the latter) does not assume that we are indifferent between prospects which are the same except that some contingent person (who exists with the same well-being and probability) exists in different states of nature. As Blackorby et al. (1998, p. 10) point out, contingent people do not have well-defined expectations, because they do not have a well-being level in all potential final outcomes. So Anteriority cannot be supported by the claim that each person has the same expectation in the prospects. Given a variant of our example that replaces the 9 with a 7, Anteriority rules out *Ex-Post* Average Utilitarianism without further assumptions.

[14] Note, moreover, that *Ex-Post* Average Utilitarianism would still favour A even if Ann's well-being in S_1 of B were 2.

objection who could be relevantly better or worse off, this counter-example does not rely on any morally relevant change to the unaffected Ann — as her prospect is exactly the same in both prospects. Finally, unlike the utility-monster objection, this counter-example does not work against total utilitarianism.[15]

We wish to thank Kacper Kowalczyk for valuable comments.

References

Anglin, Bill (1977) 'The Repugnant Conclusion', *Canadian Journal of Philosophy* 7 (4): 745–754.

Arrhenius, Gustaf (2000) 'An Impossibility Theorem for Welfarist Axiologies', *Economics and Philosophy* 16 (2): 247–266.

Arrhenius, Gustaf and Krister Bykvist (1995) *Future Generations and Interpersonal Compensations: Moral Aspects of Energy Use*, Uppsala: Uppsala University.

Asheim, Geir and Stéphane Zuber (2014) 'Escaping the Repugnant Conclusion: Rank-Discounted Utilitarianism with Variable Population', *Theoretical Economics* 9 (3): 629–650.

Blackorby, Charles, Walter Bossert, and David Donaldson (1998) 'Uncertainty and Critical-Level Population Principles', *Journal of Population Economics* 11 (1): 1–20.

— (2005) *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*, Cambridge: Cambridge University Press.

[15] By the same reasoning as in our example, Necessity-Restricted All-Affected Stochastic Pareto is also incompatible with some notable *ex-post* extensions to social risk of other non-separable population axiologies defined elsewhere in the literature, such as variable-value utilitarianism (Hurka 1983, Ng 1989, pp. 244-250, formalized as 'number-dampened utilitarianism' by Blackorby et al. 2005, pp. 144–7); rank-discounted utilitarianism (Asheim and Zuber, 2014, pp. 632); variable-population extensions of equally-distributed-equivalent egalitarianism (Fleurbaey 2010, pp. 657-658 for fixed-population cases; Spears and Zuber forthcoming); and variable-population expected maximin, which would value a prospect according to the expectation across states of the minimum well-being among people alive in that state. We explore these extensions further by using a principle related to Necessity-Restricted All-Affected Stochastic Pareto to characterize an additively-separable family of variable-population social welfare functions in a companion paper for the economic theory literature that cites the priority of this paper.

Broome, John (1993) 'Goodness Is Reducible to Betterness: The Evil of Death Is the Value of Life', in Peter Koslowski and Yuichi Shionoya, eds., *The Good and the Economical: Ethical Choices in Economics and Management*, pp. 70–84, Berlin: Springer.

Chisholm, Roderick M. and Ernest Sosa (1966) 'On the Logic of "Intrinsically Better" ', *American Philosophical Quarterly* 3 (3): 244–249.

Fleurbaey, Marc (2010) 'Assessing Risky Social Situations', *Journal of Political Economy* 118 (4): 649–680.

Harsanyi, John C. (1955) 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', *The Journal of Political Economy* 63 (4): 309–321.

— (1985) 'Does Reason Tell Us What Moral Code to Follow and, Indeed, to Follow Any Moral Code at All?', *Ethics* 96 (1): 42–55.

— (1986) 'Utilitarian Morality in a World of Very Half-hearted Altruists', in Walter P. Heller, Ross M. Starr, and David A. Starrett, eds., *Social Choice and Public Decision Making: Social Choice and Public Decision Making: Essays in Honor of Kenneth J. Arrow, Volume I*, pp. 57–73, Cambridge: Cambridge University Press.

Hurka, Thomas (1983) 'Value and Population Size', *Ethics* 93 (3): 496–507.

McCarthy, David (2017) 'The Priority View', *Economics and Philosophy* 33 (2): 215–257.

McMahan, Jeff (1981) 'Problems of Population Theory', *Ethics* 92 (1): 96–127.

Ng, Yew-Kwang (1983) 'Some Broader Issues of Social Choice', in Prasanta K. Pattanaik and Maurice Salles, eds., *Social Choice and Welfare*, pp. 151–173, Amsterdam: North-Holland.

— (1989) 'What Should We Do about Future Generations? Impossibility of Parfit's Theory X', *Economics and Philosophy* 5 (2): 235–253.

Nozick, Robert (1974) *Anarchy, State, and Utopia*, New York: Basic Books.

Parfit, Derek (1984) *Reasons and Persons*, Oxford: Clarendon Press.

Rawls, John (1971) *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Sidgwick, Henry (1907) *The Methods of Ethics*, London: Macmillan, seventh edn.

Spears, Dean and Mark Budolfson (2021) 'Repugnant Conclusions', *Social choice and welfare* 57:567–588.

Spears, Dean and Stèphane Zuber (forthcoming) 'Foundations of Utilitarianism under Risk and Variable Population', *Social Choice and Welfare*.

Thomas, Teruji (2022) 'Separability and Population Ethics', in Gustaf

Arrhenius, Krister Bykvist, Tim Campbell, and Elizabeth Finneron-Burns, eds., *The Oxford Handbook of Population Ethics*, pp. 271–295, New York: Oxford University Press.
Williams, Bernard (1973) *Problems of the Self: Philosophical Papers 1956–1972*, Cambridge: Cambridge University Press.