A new puzzle for limited aggregation

Kacper Kowalczyk

1. Introduction

Many ethicists are reluctant to aggregate – they would rather save one person from losing a life than any number of people from losing a finger – but they are also reluctant to abandon aggregation completely – they would rather save some number of people from losing an arm, say, than one person from losing a life. Many of them are, therefore, inclined to believe that we should minimize aggregate harm but only when the harms involved are close enough to each other. There has been much discussion of this view, often known as 'limited aggregation' (see e.g. Scanlon 1998: 238–41, Kamm 2000 and Voorhoeve 2014). It has given rise to several theoretical puzzles.[1] This paper describes a new puzzle which can cast doubt on limited aggregation, or, at least, bring the key choice points for limited aggregation into sharper focus.

2. A new puzzle

Imagine that you are in charge of a ship in the middle of the sea, with your private supply of vital medicine. You receive three distress calls: one from island A, one from island B, one from island C. Island A is the smallest, with population 1. Then there is island B, with population 1,000. Island C is the largest, with population 1,000,000. The people of each island urgently need your medicine. Without it, the sole A-islander will die, the 1,000 B-islanders will each lose an arm, and the 1,000,000 C-islanders will each lose a finger.

In line with limited aggregation and with no loss of generality, let us assume that the harm of dying is close enough to the harm of losing an arm, which is in turn close enough to the harm of losing a finger, which is not, however, close enough to the harm of dying. Likewise, let us also assume that 1,000 lost arms is a greater aggregate harm than one death and 1,000,000 lost fingers is a greater aggregate harm than 1,000 lost arms. Hence, if only islands A and B were involved, limited aggregation would recommend heading for island B. Likewise, if only B and C were involved, it would recommend

1 See, for example, Norcross 1997, 1998, Parfit 2003, Dougherty 2013, Gustafsson 2015, Horton 2017, Tomlin 2017, and Horton 2018, 2020. I should note, however, that many of these puzzles primarily target wholesale refusal to aggregate rather than limited aggregation more specifically.

Analysis Vol. XX | Number XX | XX 2024 | 1–9 doi: https://doi.org/10.1093/analys/anad054

heading for island C. Lastly, if only A and C were involved, it would recommend heading for island A.

Now imagine that the islands are covered in dense fog. You can see that you found yourself between two of the three islands, but you cannot see which two. None of the islanders can see where you are either. Imagine, in particular, that there is a 1/3 chance that you are next to island A, close enough to reach B with some delay, but with C unreachable; a 1/3 chance that you are next to island B, close enough to reach C with some delay, but with A unreachable; and a 1/3 chance that you are next to island C, close enough to reach A with some delay, but with B unreachable. Delay is, of course, stressful; so it would be better for any islander to be saved without delay.

The nautical chart in Figure 1 shows the three possible positions of your ship and how far it can travel from each of them.[2]

You might reason as follows:

> No matter where I am, I should head for the more remote island. If I am between A and B, the harms involved are close enough to each other and heading for the more remote B will prevent greater aggregate harm, despite the stressful delay for those saved. Likewise, if I am between B and C, the harms involved are again close enough to each other and

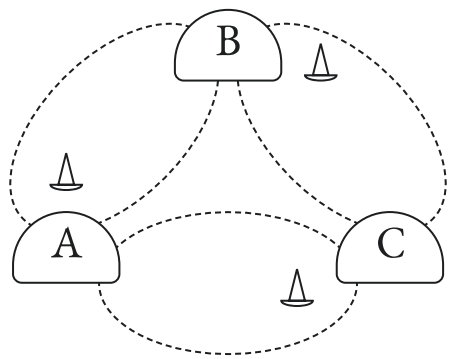

Figure 1 Three islands

2 Alternatively, we could consider the following example, where the same pattern of uncertainty arises from a more objective source: imagine that three qualitatively identical ships constantly cruise around our three-island archipelago on a circular route. The 1,000-passenger ship always follows the one-passenger ship and is in turn followed by the 1,000,000-passenger ship. All three ships are about to suffer a random mechanical failure. The position of your ship is fixed and known. One ship will disembark on the island closest to you, another on the island slightly further away, and the third on the island that you can no longer reach.

> heading for the more remote C will prevent greater aggregate harm, despite the stressful delay for those saved. Lastly, if I am between A and C, the harms involved are no longer close to each other and heading for the more remote A will prevent a much greater individual harm, despite the stressful delay for those saved.

In this way, if you happen to accept limited aggregation, you could support the decision to head for the more remote island. If you are so inclined, you could invoke the following general principle:

> (Statewise Dominance) If, conditionally on every possible resolution of uncertainty about the state of the world, one should take one action rather than another, then, even before that uncertainty is resolved, one should take the first action rather than the second.[3]

But now a puzzle arises because the following line of thought is also plausible:

> The more remote island might, in fact, turn out to be any of the three islands. So, if I head for that island, there is a 1/3 chance that I will be saving the A-islander, a 1/3 chance that I will be saving the B-islanders and a 1/3 chance that I will be saving the C-islanders, but whoever I thereby save will have had to wait. But the closer island might also, in fact, turn out to be any of the three islands. So, if I head for that island, there is a 1/3 chance that I will be saving the B-islanders, a 1/3 chance that I will be saving the C-islanders and a 1/3 chance that I will be saving the A-islander, but without stressful delay. Hence it would be better, from the perspective of every islander, if I headed for the closer island rather than the more remote island. So I should not head for the more remote island.

To support the claim that heading for the closer island would be better from the perspective of every islander, you could invoke the following general principle:

> (The Better-Prize Principle) If one outcome is better than another, then, other things being equal, it is better to have a given chance of getting the first than to have the same chance of getting the second (cf. Resnik 1987: 91–92).

The decision to head for the more remote island could then be ruled out by means of the following general principle:

3 Similar principles, typically stated in terms of individual or social preference, are often regarded as basic axioms of rationality. See, for example, Fleurbaey 2010: 655.

(Ex-Ante Pareto) If the prospect of one option is worse for everyone affected than the prospect of another option, then one should not take the first option if one can take the second.[4]

It seems that a violation of this principle could not be justified to the people affected by your decision. If the islanders were rational and self-interested, they would unanimously vote to replace you with a captain who would commit to heading for the closer island.

At this point, it might be helpful to consider Table 1, which shows which islanders are saved by which option and how fast.

Table 1 The puzzle

	A close, B far Chance 1/3	**B close, C far Chance 1/3**	**C close, A far Chance 1/3**
A-islanders (1)	saved	dead	dead
B-islanders (1,000)	no arm	saved	no arm
C-islanders (1,000,000)	no finger	no finger	saved
		Save the closer island	
A-islanders (1)	dead	dead	saved, stressed
B-islanders (1,000)	saved, stressed	no arm	no arm
C-islanders (1,000,000)	no finger	saved, stressed	no finger
		Save the more remote island	

In this table, columns correspond to the possible states of the world, which have the probabilities shown, while rows correspond to the different groups of people affected. Ex-Ante Pareto would have us view this table person by person, that is, row by row. Then the option of saving the closer island does look superior. Statewise Dominance, on the other hand, would have us view this table state by state, that is, column by column. Then, at least for a proponent of limited aggregation, the option of saving the more remote island looks superior instead.

Thus we have a puzzle: proponents of limited aggregation have to choose between two compelling accounts of what to do in our example. Or, in a more theoretical register, proponents of limited aggregation have to deny one of the following: Ex-Ante Pareto, the Better-Prize Principle or Statewise Dominance.

4 A similar principle, but stated in terms of individual and social preference, appears in Harsanyi 1955. A version of it is defended by Frick (2013), while another version is disputed by Mahtani (2017). Ex-Ante Pareto might be considered controversial by those who support deontological side-constraints; compare Hare 2016. However, for the purposes of this discussion, the principle could be limited to cases that do not involve harming or using others as a means.

3 An old puzzle

It is worth noting that, in contrast with other recent puzzles, my puzzle does not presuppose that rational self-interested people are even minimally tolerant of risk. These other puzzles appeal to cases which are structurally like the following one.[5]

A cruise ship has sunk, and everyone on board is now stranded on two islands. The ship carried a large number of passengers; call this number '*n*'. One passenger washed up on island A and the rest on island B. Your ship happens to be halfway between the two islands, carrying your private supply of potable water, which is now urgently needed by the stranded passengers. This is because island A is a desert and island B is semi-arid. The lone castaway on island A will soon die from severe dehydration, while the castaways on island B will suffer a migraine from mild dehydration. With your supply of water, the castaway on island A would suffer only a migraine from mild dehydration, while the castaways on island B would be completely fine. The trouble is that you can reach one but not both islands. The two islands are currently covered in dense fog, so you are unable to see who ended up on island A and who on island B. In particular, you find it equally likely that any one of the passengers is now on island A. The stranded passengers, let us suppose, do not know which island they are on either.[6]

It follows from scepticism about aggregation – whether limited or wholesale – that, conditionally on every possible resolution of uncertainty, you should head for island A, as that would prevent an individual harm much greater than any individual harm that could be prevented by heading for island B. It therefore follows from Statewise Dominance that you should head for island A, even before this uncertainty is resolved.

But, from the perspective of any passenger, heading for island A means a $1/n$ chance of being saved, while heading for island B means a $(n - 1)/n$ chance of being saved, which, if n is large enough, is close to certainty. Hence, heading for island A might seem to be worse for everyone ex-ante, thus contravening Ex-Ante Pareto. But that follows only if we assume that

5 See Norcross 1998, Horton 2017 and Horton 2020. A somewhat different puzzle is presented in Dougherty 2013. See also Frick's (2015) example called 'Mass Vaccination (Unknown Victims)'. My puzzle is instead structurally analogous to Fishburn's (1991) decision-theoretic puzzle for individuals with cyclic preferences (except that it involves social morality rather than individual preference), as well as to Fleurbaey's (2009) version of Harsanyi's (1955) theorem (except that my puzzle does not presuppose transitivity of the social preference relation, nor anonymity, nor expected-utility theory for individuals).

6 Alternatively we could consider the following example, where the same pattern of uncertainty arises from a more objective source: imagine that you need to decide which island to visit before the passengers wash up on either of them. Each passenger will soon be carried, by random sea currents, to one of the two islands – one to the desert island A, the rest to the semi-arid island B.

rational, self-interested people are at least somewhat tolerant of risk. In particular, it would be sufficient to assume that $(n - 1)/n$ is a probability large enough to witness the following general principle:

> (Risk Tolerance) If one outcome is better than another, which is in turn better than a third, then, for some non-zero, non-unit probability, getting the first outcome with that probability and the third otherwise is better than getting the second one for sure.[7]

While this is a plausible principle, it is less compelling than the Better-Prize Principle, which was sufficient to generate my puzzle.[8] After all, even an infinitely risk-averse person would rather replace a worse outcome with a better one with the same probability, since that could not decrease the minimum possible payoff.[9]

4 *Implications*

I expect to see two kinds of responses to my puzzle. One takes it as a refutation of limited aggregation, as it appears implausible both to deny Ex-Ante Pareto (because then one foregoes the humane ideal of justifiability to each) and to deny either Statewise Dominance or the Better-Prize Principle (which seem to be basic rules of rationality).

The other sees it as another example where we are led astray by either ex-ante reasoning or by ex-post reasoning.[10] It would be much less plausible, I think, to blame the Better-Prize Principle, which I therefore set aside.[11] Proponents of limited aggregation will thus be divided depending on which of the other two principles they wish to uphold in the context of my puzzle. To see this in more detail, it will be helpful to have a simple but concrete example of a limited-aggregation view, restricted at first to riskless cases. Consider, for example, the following view:

> (Relevant Aggregation) We should minimize aggregate relevant harm, that is, the total of individual harms which are greater than, or close

7 This corresponds to one half of the standard continuity axiom of expected-utility theory. See Jensen 1967.

8 Norcross (1998), Dougherty (2013) and Horton (2020) give some reasons in favour of risk tolerance. Temkin (2012: 258–60), however, entertains denying it, precisely in response to a case analogous to my shipwreck case described in the main text.

9 It is also worth noting that, in this older type of puzzle, everyone is, as it were, behind a veil of ignorance: everyone has a chance to end up in anyone else's shoes. In this respect, my puzzle is different: the A-islander, for example, can only end up either dead or saved, but has no chance at all of suffering the smaller harms of losing a finger or losing an arm.

10 See, for example, Fleurbaey 2010 for a discussion of this issue in the context of egalitarianism.

11 This principle might be considered controversial by those sympathetic to value incommensurability; compare Hare 2010. However, the issue of value incommensurability appears irrelevant to the truth of limited aggregation.

> enough to, the greatest individual harm that would be suffered otherwise.[12]

To obtain its ex-ante version, we first define a person's *expected harm* for a given option as the result of subtracting their level of expected welfare from the level of expected welfare they would have enjoyed if the alternative option had been chosen instead, except when this difference is negative, in which case we set that person's expected harm to zero.[13] Then we define:

> (Ex-Ante Relevant Aggregation) We should minimize the aggregate of relevant expected harm, that is, the total of expected individual harms which are greater than, or close enough to, the greatest expected individual harm that would be suffered otherwise.

In the context of my puzzle, this ex-ante view implies that you should head for the closer island, in accordance with the Better-Prize Principle and Ex-Ante Pareto.

To obtain the ex-post version, we first define a person's *harm in a state of the world* for a given option as the result of subtracting their level of welfare in that state from the level of welfare they would have enjoyed, conditionally on the same state, if the alternative option had been chosen instead, except when this difference is negative, in which case we set that person's harm in that state to zero. Then, in any given state of the world, we identify relevant harms as those greater than, or close enough to, the greatest individual harm that would have been suffered, conditionally on the same state, if the alternative option had been chosen instead. Lastly, we take the expectation of the total of these relevant harms, across all possible states of the world. Then we define:

> (Ex-Post Relevant Aggregation) We should minimize the expected aggregate of relevant harm, that is, the expectation of the total of individual harms that, in a given state of the world, are greater than, or close enough to, the greatest individual harm that would be suffered otherwise, conditionally on the same state.

If we make assumptions about relevance and aggregate harm consonant with our previous discussion, this ex-post view implies, in the context of my puzzle, that you should head for the more remote island, in accordance with the core ideas of limited aggregation and the principle of Statewise Dominance.[14]

12 This is a version of Voorhoeve's 'Aggregate Relevant Claims' view; see Voorhoeve 2014. I will follow Voorhoeve in setting aside the issue of how the view is to be understood in choice situations that involve more than two options. For a lucid account of other possible forms of limited aggregation in riskless cases, see Brown 2020.

13 Hence if x is a person's expected welfare given one option and y is their expected welfare given the alternative, their expected harm given the first option is equal to $\max\{y - x, 0\}$.

14 This view might face other problems in virtue of the fact that it determines the relevance of a harm on a state-by-state basis; see Steuwer 2022.

These ex-ante and ex-post views differ, therefore, in the relative placement of the word 'expected'. The distinction lies in whether one minimizes the *expected* aggregate of relevant harm or minimizes the aggregate of relevant *expected* harm.[15] If we want to uphold limited aggregation in response to my puzzle, we not only must choose between otherwise appealing principles but might also have to determine the placement of the word 'expected' within the statement of our own view. I leave these difficult choices to those inclined to limit aggregation.[16]

Funding

This work was supported by a grant from Longview Philanthropy.

University College London
UK
kacper.kowalczyk@ucl.ac.uk

References

Brown, C. 2020. Is close enough good enough? *Economics and Philosophy* 36: 29–59.
Dougherty, T. 2013. Aggregation, beneficence, and chance. *Journal of Ethics and Social Philosophy* 7: 1–19.
Fishburn, P. 1991. Nontransitive preferences in decision theory. *Journal of Risk and Uncertainty* 4: 113–34.
Fleurbaey, M. 2009. Two variants of Harsanyi's aggregation theorem. *Economics Letters* 105: 300–302.
Fleurbaey, M. 2010. Assessing risky social situations. *Journal of Political Economy* 118: 649–80.
Frick, J. 2013. Uncertainty and justifiability to each person: response to Fleurbaey and Voorhoeve. In *Inequalities in Health: Concepts, Measures, and Ethics*, eds. N. Eyal, S. Hurst, O. Norheim and D. Wikler, 129–46. New York: Oxford University Press.
Frick, J. 2015. Contractualism and social risk. *Philosophy and Public Affairs* 43: 175–223.
Gustafsson, J.E. 2015. Sequential dominance and the anti-aggregation principle. *Philosophical Studies* 172: 1593–601.
Hare, C. 2010. Take the sugar. *Analysis* 70: 237–47.
Hare, C. 2016. Should we wish well to all? *Philosophical Review* 125: 451–72.
Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–21.

15 A similar account of the distinction between ex-ante and ex-post forms of limited aggregation can be found in Horton 2020. Recently, hybrid and disjunctive views have also been proposed by Lazar (2018) and Walen (2020), respectively. These views aim to preserve more of our intuitions but may inherit issues from both ex-ante and ex-post forms without inheriting a unified rationale.

16 I would like to thank Johan Gustafsson, Karolina Watroba, Todd Karhu, Tomi Francis, Jonas Hertel and two anonymous referees.

Horton, J. 2017. Aggregation, complaints, and risk. *Philosophy and Public Affairs* 45: 54–81.
Horton, J. 2018. Always aggregate. *Philosophy and Public Affairs* 46: 160–74.
Horton, J. 2020. Aggregation, risk, and reductio. *Ethics* 130: 514–29.
Jensen, N.E. 1967. An introduction to Bernoullian utility theory: I. Utility functions. *Swedish Journal of Economics* 69: 163–83.
Kamm, F. M. 2000. Nonconsequentialism. In *The Blackwell Guide to Ethical Theory*, ed. H. LaFollette, 205–26. Malden, MA: Blackwell.
Lazar, S. 2018. Limited aggregation and risk. *Philosophy and Public Affairs* 46: 117–59.
Mahtani, A. 2017. The *ex ante* Pareto principle. *Journal of Philosophy* 114: 303–23.
Norcross, A. 1997. Comparing harms: headaches and human lives. *Philosophy and Public Affairs* 26: 135–67.
Norcross, A. 1998. Great harms from small benefits grow: how death can be outweighed by headaches. *Analysis* 58: 152–8.
Parfit, D. 2003. Justifiability to each person. *Ratio* 16: 368–90.
Resnik, M.D. 1987. *Choices: An Introduction to Decision Theory*. Minneapolis, MN; London: University of Minnesota Press.
Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press.
Steuwer, B. 2022. Limits to aggregation and uncertain rescues. *Utilitas* 34: 70–83.
Temkin, L.S. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. New York: Oxford University Press.
Tomlin, P. 2017. On limited aggregation. *Philosophy and Public Affairs* 45: 232–60.
Voorhoeve, A. 2014. How should we aggregate competing claims? *Ethics* 125: 64–87.
Walen, A. 2020. Risks and weak aggregation: why different models of risk suit different types of cases. *Ethics* 131: 62–86.