Philosophical Studies
https://doi.org/10.1007/s11098-023-02082-7

A trilemma for the lexical utility model of the precautionary principle

H. Orri Stefánsson[1,2,3]

Accepted: 10 November 2023

Abstract
Bartha and DesRoches (Synthese 199(3–4):8701–8740, 2021) and Steel and Bartha (Risk Analysis 43(2):260–268, 2023) argue that we should understand the precautionary principle as the injunction to maximise *lexical* utilities. They show that the lexical utility model has important pragmatic advantages. Moreover, the model has the theoretical advantage of satisfying all axioms of expected utility theory except continuity. In this paper I raise a trilemma for any attempt at modelling the precautionary principle with lexical utilities: it permits choice cycles or leads to paralysis or implies that the smallest value difference that is possible in a context has extreme axiological implications.

Keywords Catastrophe · Lexical utility · Precautionary principle · Public health · Vagueness

1 Introduction

The precautionary principle (PP) has played an important role in both environmental law and chemical regulation over the last decades.[1] More recently, the principle has been appealed to—implicitly and explicitly—when deciding how to respond to the Covid-19 pandemic. For instance, both the imposition of restrictions on movement

[1] Two influential examples are Principle 15 of the United Nation's Rio Declaration, (https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_CONF.151_26_Vol.I_Declaration.pdf) and the European Commissions communication on the precautionary principle (https://op.europa.eu/en/publication-detail/-/publication/21676661-a79f-4153-b984-aeb28f07c80a/language-en).

✉ H. Orri Stefánsson
orri.stefansson@philosohy.su.se
http://www.orristefansson.is

[1] Stockholm University, Stockholm, Sweden

[2] Swedish Collegium for Advanced Study, Uppsala, Sweden

[3] Institute for Futures Studies, Stockholm, Sweden

Published online: 22 January 2024

between EU countries and rules about mandatory quarantines have been interpreted as applications of the principle (see, respectively, Goldner Lang (2021) and Raposo (2022)).

Although widely cited and purportedly applied, there is no agreement on the precise formulation of the precautionary principle.[2] But one influential formulation of the principle can be found in the Rio Declaration, principle 15 of which states that:

> In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.

The above is however far from being a clear and action guiding decision rule. In fact, it has been argued (Peterson, 2006; Stefánsson, 2019) that there is no coherent formulation of PP as a decision rule. For instance, suppose that one accepts a seemingly weak Archimedean (trade-off) condition:

Weak Archimedes Some (nonnegligible) increase in the relative likelihood of a catastrophic outcome can be at least counterbalanced by the relative likelihood of a non-catastrophic outcome being increased in relation to a strictly worse non-catastrophic outcome.

Then PP understood as a decision-rule either violates transitivity,[3] or gives conflicting recommendations depending on whether we apply it 'globally' or 'locally',[4] argues Stefánsson (2019) (based on work by Peterson (2006)).

Bartha and DesRoches (2021) and Steel and Bartha (2023) have recently responded to such arguments by suggesting that we should understand PP as the injunction to maximise *lexical* expected utility.[5] They defend the lexical utility model of the precautionary principle as recommending non-contradictory choice(s) in any decision scenario—thus avoiding the charge of incoherence—and, relatedly,

[2] Some, in fact, argue that it is a mistake to see it as *a* principle; see, e.g., Hartzell-Nichols (2013) and Sandin and Peterson (2019).

[3] A transitive decision-rule chooses A over C whenever it chooses A over B and B over C.

[4] Here is an example of latter problem: the principle may imply that we should not allow a set of chemicals even though we should allow each chemical in the set, and even though there are no negative interaction effects between the chemicals in the set. This gives rise to practical difficulties, as the precautionary principle cannot be safely applied by a risk manager unless she can foresee all future decisions and recall any past decision. (See further discussion in Stefánsson (2019).) Now, PP is by no means the only choice principle that gives conflicting recommendations depending on whether it is applied locally or globally; the literature on dynamic choice has shown that in some dynamic situations (say, where each action results in trivial harms that nevertheless add up to a significant harm), many choice rules can lead to such results. For an overview of problems of this kind, see Andreou (2020).

[5] This means that, unlike many philosophers (e.g., Gardiner (2006), Peterson (2006), Steel (2015), Stefánsson (2019)), Bartha, DesRoches and Steel don't primarily think of the precautionary principle as a *qualitative* decision rule to be used when quantitative probabilities and/or utilities are lacking. However, Bartha and DesRoches (2021, sec.: 5.2) show how their model can be generalised to a qualitative decision rule.

as satisfying all the standard axioms of expected utility theory except for continuity.[6,7] By allowing for violations of continuity specifically when it comes to gambles that may result in catastrophic outcomes, the lexical utility model does not satisfy the above Archimedean condition.

The key concept for the lexical utility decision rule—and, in fact, for the precautionary principle too—is a *harm threshold*, which intuitively distinguishes catastrophic outcomes from non-catastrophic outcomes. Beyond that distinction, no fixed interpretation is given to this threshold (Bartha and DesRoches 2021, 8704, 8720). In the simple two dimensional lexical utility model on which Bartha, DesRoches, and Steel mostly focus, the harm threshold is accounted for by using two independent utility functions, u_1 and u_2. The first function takes the value -1 if the outcome to which it is applied is catastrophic, but otherwise it takes the value 0. The latter is a standard utility function, summarising the outcome's more 'ordinary' (i.e., non-catastrophic) costs and benefits.

The recommendation of the lexical utility decision rule is always to maximise lexical expected utility. To state this more precisely, let's suppose that there are n states of the world, $s_1, \ldots, s_n$, representing the uncertainty of the decision situation.[8] Let $p(s_i \mid A)$ be the probability of state s_i when alternative A is chosen while $u_j(A\&s_i)$ is the utility (according to function j) of choosing alternative A when state s_i obtains. The lexical utility decision rule is based on the expected utilities (EUs) that are generated by both u_1 and u_2:

$$EU_1(A) = \sum_{i=1}^{n} u_1(A\&s_i)p(s_i \mid A)$$

$$EU_2(A) = \sum_{i=1}^{n} u_2(A\&s_i)p(s_i \mid A)$$

Finally, the lexical utility decision rule says that, for any alternatives A and B, the latter should be chosen over the former just in case either

$$EU_1(A) < EU_1(B),$$

or

$$EU_1(A) = EU_1(B) \text{ and } EU_2(A) < EU_2(B)$$

[6] One could of course generalise the lexical utility model further, for instance, by weakening the independence axiom to get a lexical *risk-weighted* (Buchak, 2013) utility theory.

[7] Continuity implies that no outcome o is so bad that one should not be willing to to accept a gamble G that has some probability of resulting in o as long as G also offers sufficient probability of an outcome o^+ that is preferred to the status quo.

[8] In other words, the decision-maker may not know which state obtains, but for each state s_i and each alternative A, she knows for sure which outcome she gets if she chooses A and s_i obtains.

Note that, since u_1 takes the value of either 0 or -1—the catastrophe either occurs or not[9]—$EU_1(A)$ is simply the negative of the probability that A results in a catastrophe. So, informally, the lexical utility decision rule says that for any two alternatives A and B, you should choose B over A in case either:

- B offers a lower probability of the catastrophe, or
- the probability of the catastrophe is the same given either alternative, but B offers a higher expected utility according to u_2.

The lexical utility decision rule thus gives absolute priority to minimising the long-run (Steel and Bartha 2023: 265–266) risk of catastrophe. If B offers an all-things-considered lower probability of catastrophe than A, then B should be chosen rather than A irrespective of how the two alternatives otherwise compare.

Is this type of absolute (lexical) priority plausible? For instance, is it plausible that, say, no aggregate welfare benefits can ever outweigh any chance of a public health catastrophe? I find it very hard to believe that the true (or correct) prospect axiology could have such a lexical structure. However, Bartha and DesRoches (2021) offer a *pragmatic* defence of the lexical utility model (based on remarks by Christiansen (2019)), which may show that the use of the lexical utility decision rule could be justified even if the true axiology does not have this lexical structure.

> First: we may face a high-stakes decision where the relevant utilities and probabilities are not sufficiently determinate to settle the issue of continuity, and not sufficiently determinate to allow a decision using standard decision theory. Second: it may be that our preferences are *locally discontinuous*, given the realistic range of utilities and probabilities. [...] In both sorts of case, the need for action can justify a model that represents our preferences as discontinuous and, in these cases, lexical utility maximization provides a framework for articulating a simple and straightforward rationale for precautionary measures. (Bartha and DesRoches 2021, 8728)

Similarly, Steel and Bartha (2023) point out that a lexical utility model may—even if theoretically not quite right—provide a "sufficiently good approximation and likely to produce better results in practice than a model wherein utilities are continuous and the probabilities must be specified with precision" (262).

[9] In Steel and Bartha's (2023) slightly more general lexical utility model, which allows for multiple co-occurring catastrophes each of which is equally bad, this condition becomes: we should minimise the expected number of such catastrophes. However, when it comes to what they call "terminal" catastrophes, whose distinguishing feature is that "once one occurs, it makes no difference what happens afterward" (264), the model implies that we should minimise the probability of such a catastrophe.

Although perhaps pragmatically appealing, the lexical utility model faces some theoretical difficulties (as Bartha, DesRoches, and Steel are of course aware[10]). In this paper I raise a new theoretical challenge for any attempt at using a lexical utility model to formalise the precautionary principle. The basic problem is that such a version of the precautionary principle faces a trilemma—in fact, it faces two versions of this same trilemma for two independent reasons:

- Marginal value differences (in fact, the smallest value difference possible, given the context) result in extreme axiological differences,[11] or
- a vague threshold is used in a way that either implies a choice cycle, or
- leads to paralysis, that is, a situation where no choice is permissible.

Here is an illustration of the first horn of the trilemma. Suppose we think that not imposing 'lockdown' in response to the spread of a respiratory virus could result in at most n patients needing intensive care within a particular time period. Then some economic gain would justify avoiding lockdown, according to the lexical rule. However, if we then come to think that there is *any* chance, no matter how small, that not locking down could result in $n + 1$ patients needing intensive care, then *no* economic gain could justify avoiding lockdown (because $n + 1$ patients needing intensive care would be catastrophic).

In the lockdown example, the second horn of the dilemma arises if we decide that the threshold for what counts as a catastrophic number of patients in intensive care, which no economic gain could offset, is vague. As we shall see in the next section, this can lead to choice cycles. That seems particularly problematic given, first, that the lexical utility model of the precautionary principle was partly formulated as a response to the charge that any decision-rule version of the principle is incoherent. It is hard to see that a decision rule is coherent if it implies that one should choose A over B, B over C, and C over A. Second, the lexical utility model of the precautionary principle was advertised as a decision rule that satisfies transitivity,[12] which means that it should not permit cyclical choices.

An instance of the third horn of the dilemma would be that both imposing and not imposing lockdown is impermissible. Similarly to the second horn, this horn seems particularly problematic in light of the fact that Bartha and DesRoches (2021: 8731–8732) argue that their lexical utility model of PP does not lead to the kind of paralysis that for instance Sunstein (2005) criticises PP for. However, if Bartha and

[10] For instance, Bartha and DesRoches (2021) point out that the lexical utility model can have counterintuitive "absolutist" implications, which they however think can be justified (see, e.g., page 8736). Moreover, in the concluding section they say: "We acknowledge that there is more work to be done to establish the theoretical credentials of the lexical utilities approach to [the precautionary principle], but we maintain that it is viable." (8737–8738)

[11] This is an instance of what Pummer (2018, 2022) calls *hypersensitivity* that is, "when a slight difference in one sort of property makes a radical difference in another sort of property" (2022: 510). As Pummer points out, such sensitivity is deeply puzzling, in particular when a slight difference in a non-evaluative property (say, the number of patients in intensive care) makes a radical difference in an evaluative property (say, the moral badness of the situation).

[12] See e.g. Bartha and DesRoches (2021: 8704).

DesRoches' version of the principle is to avoid the first two horns of the trilemma, then it would seem they have to insist that there may be situations where their rule implies that none of the available options can permissibly be chosen.

The lexical formulation of the precautionary principle faces two versions of the above trilemma. In the next section I formulate and discuss the first version, in the section after I formulate and discuss the second version.

Before considering the trilemma in more detail, it may however be worth saying a few words about the scope—or, rather, the limit—of my analysis. As previously mentioned, my aim is to identify a new *theoretical* problem for the lexical utility model. *In practice*, someone applying the lexical utility model might nevertheless do better, even in some extremely important decision-contexts, than someone who applies (or tries to apply) the expected utility model or standard cost-benefit analysis. As Bartha and DesRoches (2021) point out (e.g., pp. 8728–8729), there are some decision-problems—in particular, problems involving potential catastrophes and limited information about the relevant probabilities—where trying to figure out the exact probabilities and utilities, in order to calculate expected utility, would be futile. Similarly, "computational convenience" (Bartha and DesRoches 2021: 8723) in many cases speaks in favour of the lexical model over the expected utility model. So, despite the theoretical problems I identify, we may well be justified in often using the lexical utility model. Still, this paper highlights some potential costs of doing so; and, similarly, identifies situations in which one must proceed carefully with the lexical utility model.

Moreover, it is possible that the lexical utility model is a faithful formalisation of PP despite the theoretical problems I identify. Perhaps the above mentioned trilemma simply is an inherent feature of the principle. And, in fact, "descriptive accuracy" is the main defence that Bartha and DesRoches (2021) offer in favour of their lexical utility model: "the lexical approach faithfully models precautionary reasoning in its handling of actual examples" (8704). If that is so, then my paper can be read as highlighting the theoretical costs of PP and as identifying situations in which one must proceed carefully when applying the principle.

2 A threshold for graded harms

In some applications of the precautionary principle, the harm threshold may reflect something that is truly binary, say, human extinction, a climatic tipping point, or an ecosystem collapse. In those cases it may be that a miss is as good as a mile. For instance, the difference between, on the one hand, humanity going extinct, and, on the other hand, humanity *almost* going extinct may not be a difference in degree; it may be a difference in kind. So, it may make sense to treat human extinction as being categorically different from "almost extinction".[13]

[13] This claim could be questioned (as a referee for this journal points out). Doesn't extinction vs. non-extinction really come down to how large the total (time-extended) human population will be? If so, aren't we dealing with a graded rather than binary harm? This in fact raises the interesting question of whether there really are any truly binary catastrophes. For the sake of the argument, I however grant Bartha and DesRoches (2021) the claim that it makes sense to represent some catastrophes—or our attitudes

The above is clearly not true of all applications of the precautionary principle. Sometimes the harm threshold reflects some cut-off point in a variable that, if not continuous, at least comes in degree. An example is the application of the principle when evaluating a public health policy, such as a strategy against the coronavirus (which Steel and Bartha (2023) in fact use to illustrate their model). In such applications, the harm threshold will typically be something like: "x number of patients need intensive care" (within some particular time period), or perhaps "x people die from the coronavirus disease". I shall assume the former in what follows.

Any such cut-off raises the question: why this number? Why not $x + 1$? Any choice may seem arbitrary.[14] That is not, however, the problem I shall be concerned with. Instead, the problem I shall focus on is that, if one takes the lexical utility model seriously, we face the problem that marginal value differences result in extreme axiological differences.[15] Recall the lockdown example from the introduction; let me spell it out a little.

Suppose that, early in the Covid-19 pandemic, the government has gathered their economic and health policy advisors to try to decide if it should lock down the economy. To simplify greatly, suppose that all the economic advisors agree that the expected economic cost of locking down—or, to put it differently, the expected economic gain from not locking down—is X billion dollars. All the public health experts in the room agree that not locking down would at most result in $x - 1$ additional patients needing intensive care within the relevant period . As the government is about to decide that, in that case, the gain from not locking down is worth the cost, a fringe health expert enters the room, late as usual. He tends to be controversial, just for the sake of it, and the government knows this. The problem is that he expresses the opinion that the result of not locking down will be that x patients need intensive care within the relevant period, which happens to be the number deemed catastrophic. Everyone agrees that this maverick is the least trustworthy of all the health experts. Still, there is *some* chance that he is right, so the government feels it must take his opinion into account. But then the lexical utility model implies that no possible economic gain could justify not locking down.

The reason why I think the above implication is implausibly extreme, is that the only way to make sense of the shift that happens when the fringe expert expresses his opinion, would be to think of the difference between x patients needing intensive care and $x - 1$ patients needing intensive care as being *extremely* important.[16] If there is some economic benefit that could offset risking $x - 1$ patients needing

Footnote 13 (continued)
to them—as being "binary" in the above sense (see also Christiansen (2019)). But those who disagree can simply ignore the next section, and focus on the trilemma presented in this section.

[14] A non-arbitrary reason for considering it catastrophic when the number of patients needing intensive care within a time frame exceeds a particular threshold, is that it would force intensive care units to apply rationing. I return to this issue later in this section.

[15] As some readers may recognise, the following result is similar to those in Arrhenius and Rabinowicz (2005, 2015) (in particular, Observation 3 in the 2005 article and Observation 5 in the 2015 article).

[16] Another way to put this is that this reflects hypersensitivity (Pummer, 2018, 2022) to the number of patients needing intensive care, but only at the point between $x - 1$ and x.

intensive care, but no economic benefit that could offset risking x patients needing intensive care, then that seems to suggest that the difference between x patients needing intensive care and $x-1$ patients needing intensive care is extremely important. How could we otherwise justify treating these two outcomes so differently, for instance, being willing to risk $x-1$ patients needing intensive care if the potential economic benefit is sufficiently hight while not being willing to risk x patients needing intensive care no matter the potential economic benefit? But, from a social point of view—which surely is the relevant point of view when evaluating public health policies—the evaluative difference between x patients needing intensive care and $x-1$ patients needing intensive care is trivial. So, the lexical utility formulation of the precautionary principle seems to have implausible implications, when applied to public health.[17]

In response, the defender of the lexical utility model might argue that, actually, the evaluative difference between x patients needing intensive care and $x-1$ patients needing intensive care is *not* trivial, even from the social point of view. After all, it is the difference between a catastrophic outcome and a non-catastrophic one![18] Although this is certainly a possible response, I suspect that defenders of the lexical utility model would find it too counterintuitive. It would at least require a convincing explanation of why x is so important.[19]

The only plausible reason that I can think of for why x could be so important, is if $x-1$ is the number of patients that can be treated by society's intensive care units (within the relevant period). Then the difference between $x-1$ and x patients needing intensive care may not be quite trivial, even from a social point of view. After all, that means that some very difficult (and tragic) choices will have to be made about to whom to provide intensive care. Still, it is hard to see that, even in this case, the difference between $x-1$ and x patients needing intensive care is sufficiently important to justify sometimes being willing to risk the former while never being willing to risk the latter. After all, even if the intensive care capacity is $x-1$ patients, the worse outcome only means that a single patient will not be granted needed intensive care. And although that may be tragic, it is hard to see that it could

[17] Instead of using the lexical utility model, we might say that quantities of patients in intensive care have increasing marginal negative value that becomes infinite at some quantity x but is continuous up to that point. So, $x-1$ patients in intensive care will then also be extremely bad. Does this avoid the trilemma I am raising for the lexical utility model? It does not. It will still be the case that one additional patient has extreme axiological implications. Since any positive probability multiplied by infinity is still infinity, the model now under consideration will also imply that there is some benefit that justifies taking some gamble that might result in $x-1$ patients needing intensive care, but no benefit that could justify any gamble that might result in x patients needing intensive care. (For a structurally similar discussion, see Eyal (2020), e.g. pp. 145–146.)

[18] I am grateful to a referee for bringing this response to my attention.

[19] While discussing an example of Covid-19 triage, Steel and Bartha (2023) say that "each patient's death might be considered a catastrophe" (262). That may be true from the point of view of, say, the management or staff of a (small) intensive care unit, and of course from the point of view of the dying person and their family. But a single death is not a catastrophe from the social point of view, which is the view from which we (should) evaluate public policies. But more importantly for the current argument, the claim that each death is catastrophic is more plausible, I contend, than the claim that for some $x > 1$, x deaths are catastrophic even though $x-1$ deaths are not.

justify treating these gambles so extremely differently. From the point of view of each patient, the difference between the two outcomes is that in the latter outcome, each patient faces an average $\frac{1}{x}$ probability of not getting the needed intensive care. But recall that $x - 1$ is the number of patients that can receive intensive care (within some time period). For any modern society,[20] $\frac{1}{x}$ is a very small probability (no matter how small the aforementioned time period). So, we still find that the lexical utility model of the precautionary principle seems to have implausible implications, when applied to public health.

Alternatively, the defender of the lexical utility model might respond that if indeed the evaluative difference between x patients needing intensive care and $x - 1$ patients needing intensive care is trivial, from a social point of view, then if $x - 1$ patients needing intensive care is not a catastrophe, we were simply mistaken in treating x as a catastrophe.[21] The problem, of course, is that to avoid the above result—without claiming that each patient in intensive care is a catastrophe—they would have to say in addition that for *any* integer y, y patients needing intensive care is not a catastrophe if $y - 1$ patients needing intensive care is not a catastrophe. And that assumption also leads to trouble, as we shall see in a moment.

Suppose that despite the above problem, we want to maintain that the lexical utility framework is a good model for the precautionary principle in cases where the relevant harm threshold concerns a variable that comes in degree. (In the next section I consider what might be truly "binary" catastrophes.) What are then our options, if we still want to avoid extreme results like that above? I think that our only feasible option is to say that the harm threshold is *vague*.[22]

Let's assume, then, that the threshold is some vague or fuzzy range that includes x. In that case, the fact that the maverick health expert expresses his opinion that the number of patients needing intensive care will be x rather than $x - 1$, need not make such an extreme a difference. If the harm threshold is vague, and if $x - 1$ patients fall below the threshold, then plausibly x patients will not be (determinately) above the threshold.

Unfortunately we are not free from trouble yet. Suppose there is a second fringe health expert, who also arrives late and expresses her opinion that, if the government doesn't lock down, then that will result in $x + 1$ patients needing intensive care. And so on, imagining a third, fourth, etc., fringe health experts, respectively expressing their opinion that not locking down will result in the number of patients needing intensive care being $x + 2$, $x + 3$, etc. To avoid the above extreme result, we might be tempted to make the assumption that for any integer y, if there is some potential economic gain Y that can compensate for y patients needing intensive care, then there is some potential economic gain (greater than Y) that can compensate for

[20] Why is the relevant point of view not that of an intensive care unit? Because what we are evaluating is the *public* policy of locking down.

[21] I am grateful to a referee for bringing this response to my attention.

[22] Andersson (2022) suggests a similar response to an analogous worry about "spectrum" arguments, specifically, the question of how two pains that differ only marginally in intensity could nevertheless differ in kind.

$y + 1$ patients needing intensive care. More generally, if y patients needing intensive care is (determinately) not a catastrophe, then $y + 1$ patients needing intensive care is also not a catastrophe. And that may seem very natural, given the stipulation that the threshold is vague. (Compare: if the threshold for being bald is vague, then it would be natural to say that, for any integer y, if having y hairs means that a person is not bald, then having $y + 1$ hairs means that a person is not bald.)[23]

But it is not hard to see that a problem is looming. Even if the threshold is vague, there must surely be some integer z such that, even though there is some potential economic gain that can compensate for y patients needing intensive care, no potential economic gain can compensate for $y + z$ patients needing intensive care. Otherwise the lexical utility model would be trivialised, and we would be back to standard cost-benefit analysis, which the precautionary principle (and the lexical utility model) is meant to be a genuine alternative to.

The above means that we face a Sorites paradox. For any number y of patients needing intensive care that does not constitute a catastrophic outcome, no single additional patient in intensive care makes the outcome catastrophic; still, there is some number of patients in intensive care that constitutes a catastrophe. While puzzling, these types of paradoxes are very familiar, and not what I shall be focusing on. Instead, I shall focus on how this puzzling feature of vague thresholds can result in choice cycles, when choices are guided by the lexical utility model.

As an illustration of how such a cycle can arise, consider the following (admittedly somewhat unrealistic) example. Suppose that policy 0 is predicted to result in at most y patients needing intensive care, which is determinately non-catastrophic, and is also expected to result in economic gain of magnitude Y. Policy 1 however is expected to result in additional economic gain of magnitude Y^1 (so, gain $Y + Y^1$ in total) and is predicted to result in at most one additional patient needing intensive care (so, $y + 1$ patients in total). By previous assumptions, $y + 1$ patients needing intensive care is not catastrophic. So, there must be some Y^1 for which policy 1 should chosen over policy 0.

We can continue this reasoning. Ultimately we then reach a policy z, the implementation of which means that the total expected economic gain is $Y + Y^1 + \cdots + Y^z$ and the number of patients needing intensive care is predicted to be at most $y + z$. The assumption that the harm/catastrophe threshold is vague seems to imply that for each i, policy $i + 1$ should be chosen over policy i for *some* additional economic gain Y^{i+1}. But suppose that $y + z$ patients needing intensive care is a determinately catastrophic outcome. In contrast, we assumed that y patients in intensive care was determinately not a catastrophic outcome. Then there is no economic gain that offsets risking $y + z$ patients needing intensive care rather than y patients. So, policy 0

[23] It may be worth emphasising that "limited aggregation" (Bartha & DesRoches, 2021: 8735) doesn't apply here; it only applies "in a context that involves risks of one harm that aggregate in a clearly foreseeable way to be comparable to another harm". In the present example, limited aggregation would apply if the *economic* harms were to aggregate "in a clearly foreseeable way to be comparable to" the public health harm.

should be chosen over policy z, for any economic gain $Y + Y^1 + \cdots + Y^z$. In other words, we have a choice cycle.

There may be ways of avoiding such choice cycles, without ending up back at the first horn of the trilemma. Perhaps there is an integer j such that, although $y + j$ patients needing intensive care is determinately not catastrophic, it is *indeterminate* whether $y + j + 1$ patients needing intensive care is catastrophic. And if this seems too close to the first horn of the trilemma, we can appeal to second order vagueness and say that it is *indeterminate* whether this shift happens at integer j rather than some other integer in a (fuzzy) interval around j (cf. Andersson (2022)). The implication would then be that for some integer(s) j, it is indeterminate whether some economic gain justifies choosing policy $j + 1$ over policy j (in the sequence from the last paragraph). Moreover, where this indeterminacy starts is indeterminate.

Now, whether the above move prevents choice cycles depends on what one ought to do when what one ought to choose is indeterminate. But it seems we should say one of the following.[24] First, we might say that if it is indeterminate whether option A should be chosen over B, then it is permissible to choose either option when only these are available. It is not hard to see that this would permit choice cycles (e.g., by the argument from the last section).[25] Second, we could say that if it is indeterminate whether option A should be chosen over B, then it is indeterminate whether it is permissible to choose either option when only these are available. But that just raises the question of how one should choose when what is permissible is indeterminate, to which analogous answers to those considered here would seem to be the only plausible options. Third, we might say that if it is indeterminate whether option A should be chosen over B, then one is permitted to choose neither. But that would constitute a particularly unwelcome instance of incompleteness,[26] and in fact paralysis, when A and B are the only options.

In other words, we can construct a sequence of options such that the lexical utility model, with a vague harm threshold, implies one of the following:

1. The model permits cyclical choices, or
2. the model leads to paralysis.

Without a vague threshold, however, the lexical utility model implies that trivial value differences result in extreme axiological differences. Thus, the trilemma.

[24] Other answers are of course *possible,* for instance: if it is indeterminate whether option A should be chosen over B, then one ought to choose one over the other. But such answers seem too implausible to consider.

[25] What should determine the ranking of public policies are properties of the alternatives rather than properties of the chooser. Therefore, what the chooser "wills" (see, e.g., Chang (2013)) does not suffice to make cyclical choices impermissible in this case, nor do commitments created by previous choice.

[26] The axiom of completeness in decision theory is typically understood to mean that any two alternatives are ranked by the theory's *preference relation* But even when completeness, thus understood, fails to hold between two alternatives, it does not follow that it is impermissible to choose between them.

3 A probabilistic threshold

I now turn to a second reason for why the lexical utility decision rule leads to a trilemma structurally very similar to the one raised in the last section. It might be worth acknowledging right away that the trilemma discussed in this section may be less troubling—and, perhaps, less interesting—than the trilemma discussed in the last section. The trilemma discussed in the last section arises because, first, lexical utilities were meant to model the precautionary principle, and, second, the principle is often applied in contexts where the harm threshold is a cut-off point in some graded variable. In contrast, the trilemma discussed below derives *directly* from the lexical utility model, irrespective of the uses to which it is put. Therefore, it should not come as a surprise to those who are attracted to the lexical utility model. Nevertheless, since the trilemma in this section is structurally almost identical to the previous trilemma, I hope it is worth discussing.

Let us now focus on catastrophes, and corresponding harm thresholds, that are what Steel and Bartha (2023: 264) call *terminal*, that is, an outcome such that "once [it] occurs, it makes no difference what happens afterward". Maybe human extinction is the most plausible example of such a catastrophe. So, it may be quite right—not just in practice but even in theory—to treat human extinction as being categorically different from any outcome that is close to but not quite extinction.

The lexical utility decision rule then implies that if we are evaluating two alternatives, and one has *any* chance—no matter how small—of resulting in human extinction whereas the other has no chance of resulting in human extinction, then we should choose the latter, irrespective of other costs and benefits. But note that this is a version of the first horn of the trilemma discussed in the last section: marginal value difference[27] results in extreme axiological difference.

To take an example, recall that before CERN's test run of the Large Hadron Collider in 2008, concerns were raised that this might create a black hole that could mean the end of humanity.[28] Law suits were even filed, including to the European Court of Human Rights, to stop the project. Courts dismissed the claims, and CERN's own safety review concluded:

> [Large Hadron Collider] collisions present no danger and [...] there are no reasons for concern. Whatever the [Large Hadron Collider] will do, Nature has already done many times over during the lifetime of the Earth and other astronomical bodies.[29]

CERN and the courts were no doubt right in concluding that the experiment was justifiable. But could they say with absolute certainty that there was no chance that

[27] I am assuming that the badness of human extinction is finite. (Otherwise, even the tiniest possible increase in the chance of extinction would arguably not be a marginal value difference.) As Broome (2013: S30) puts it: "We are a finite species living on a finite planet. There has to be a finite limit on the badness of anything that can happen here."

[28] For an example of the reporting of these worries by respectable media, see, e.g., https://www.theguardian.com/news/blog/2008/sep/07/1.

[29] See https://press.cern/science/accelerators/large-hadron-collider/safety-lhc.

critics, such as the German biochemistry professor Otto Rössler—one of the more vocal critics of the experiment—were mistaken? That is, were they justifiably 100% certain that the experiment would not create a black hole? I think they were not. In fact, it seems to me that the arguments by Otto Rössler and others, while *almost* certainly mistaken, should have had (and maybe did have) some, but perhaps *very* small, positive impact on the degree to which it was believed possible that the experiment would create a black hole. But, if we take the lexical utility decision rule seriously, it then follows that the experiment should have been abandoned, irrespective of how small impact the argument did or should have on the relevant agents' degrees of belief (i.e., on their subjective probabilities) and irrespective of how beneficial they predicted the experiment would be. So, we get the first horn of the trilemma again: marginal value difference results in extreme axiological difference.

Now, the courts and CERN may well have been justifiably confident that the catastrophic risk in question was what in legal contexts is often called *de minimis* (Adler, 2007; Peterson, 2002), or small enough to be ignored. The lexical utility model does not contain such a *de minimis* threshold,[30] but it of course could, in which case we would not get the troubling results from the previous example. However, the exact same problem would re-appear. Suppose that a second dissenting professor had come to the same conclusion as Otto Rössler. Further suppose that the testimony of this (or some subsequent, individiual) dissenting scientist moved CERN's and the judges' degrees of belief that the experiment would result in a black hole *very* slightly above the *de minimis* threshold. For instance, let's say a catastrophic risk is *de minimis* when the probability of the catastrophe is no greater than δ, and suppose that, before the second professor made their argument, the agents in question believed to degree $\gamma \leq \delta$ that the experiment would result in a black hole, while after the professor had made their argument, the agents in question believed to degree $\epsilon > \delta$ that the experiment would result in a black hole. Then if the agents applied a lexical utility model, they would have had to abandon the experiment, irrespective of how beneficial they predicted the experiment would be and no matter how close ϵ is to γ. So, yet again, we get the first horn of the trilemma: Marginal value difference results in extreme axiological difference.[31]

Perhaps Bartha, DesRoches, and Steel would simply bite the bullet and accept the above implication of their model (in fact, I suspect they would, as I alluded to at the beginning of this section). However, if we want to avoid this problem, while maintaining the lexical utility model, then the obvious solution would seem to be

[30] Steel and Bartha however suggest that in practice, both standard decision models and lexical utility models include implicit *de minimis* thresholds: "since models are simplified approximations of complex reality, they often assume that some unlikely yet possible events will not occur" (2023: 261).

[31] Note that the lexical view developed by Lee-Stronach (2018) faces the same problem. Instead of an absolute threshold, Lee-Stronach stipulates that a state (and a corresponding outcome) can be ignored if the ratio of its probability to the probability of the most likely alternative state (/outcome) is below a threshold. Therefore, since the threshold is assumed to be precise, the implication is that marginal value difference results in extreme axiological difference. An analogous problem arises for Smith's (2022) *normic de minimis decision theory*. Smith suggest that we can ignore outcomes that are sufficiently "abnormal". Given his assumptions about the measure of abnormality, the resulting notion is sufficiently precise so that the problem in question arises for this theory too.

to assume a **vague** *de minimis* threshold. In other words, we might stipulate some vague or fuzzy probability interval such that if the probability that some alternative results in a catastrophe determinately falls below that interval, then that catastrophic risk is *de minimis* and may be ignored,[32] but not if the probability is determinately above the interval.

The problem is that while this move avoids the first horn of the trilemma, it either leads to paralysis or permits choice cycles. Imagine that a sequence of seemingly independent fringe scientists come forward with support of professor Rössler's argument, one after the other. For each scientist, CERN and the courts revise upwards their degrees of belief that the experiment could result in a black hole by some tiny magnitude r^-. To avoid the first horn of the trilemma, we might want to say that if a risk is *de minimis* when the relevant probability is ζ, then so is a risk when the probability is $\zeta + r^-$. So, then if the courts and CERN though that the risk of the experiment turning out catastrophic was *de minimis* when the *n*th professor expressed their agreement with Rössler's argument, then they should also think that the risk was *de minimis* when the $(n + 1)$th professor expressed their agreement with Rössler's argument. But, obviously, given a long enough sequence of such professors, we eventually reach a sufficiently high probability such that the risk is determinately not *de minimis*. So, again we face a Sorites paradox.

The explanation for why we in addition either face choice cycles or paralysis is structurally very similar to the argument in the last section. If a risk that an alternative results in a catastrophe is *de minimis*, then the alternative should—according to the lexical utility decision rule with a *de minimis* threshold—be judged on its non-catastrophic costs and benefits. Let a risk with probability γ be determinately *de minimis*. Now imagine that God makes the government the following offer: she will raise the GDP by $\$X$, but the cost is that the chance of human extinction will increase to γ. Using the lexical utility model with a *de minimis* threshold, the government accepts the offer. Now God offers to instead raise the GDP by $\$X^+ > X$, which comes with additional $\gamma + r^-$ chance of human extinction. Using the lexical utility model, the government finds this option to be even better than God's previous offer.

The problem, of course, is that God can continue this sequence, repeatedly offering the government options that seem better to them than the previous option, but where eventually the risk of human extinction is determinately not *de minimis*. So, if we are to avoid the first horn of the trilemma—i.e., to avoid marginal differences having extreme axiological differences—when it comes to any pair of adjacent offers by God, then it seems we will (by reasoning analogous to that in the last section) have to accept either a choice cycle or that no choice is permissible between some adjacent options in the sequence of options offered by God. Thus, yet again, the trilemma:

[32] To avoid problematic violations of statewise dominance, we might want to add: except when comparing the alternative to another alternative *exactly* like the first except that it has *no* chance of resulting in the catastrophe (see Lundgren and Stefánsson 2020).

- Marginal value difference results in extreme axiological difference, *or*
- a vague threshold is used in a way that either implies a choice cycle, or
- leads to paralysis, that is, a situation where no choice is permissible.

Now, as alluded to in the introduction, it might well be that (something like) the above trilemma is really what we should expect when applying the precautionary principle to such an extreme example as the Large Hadron Collider. In that case, Bartha and his colleagues are correct when they claim that the lexical utility model is a faithful representation of the principle. And this section then illustrates problems that can arise when applying PP, but that are not unique to the lexical utility model of the principle.

Before concluding this section, I should address the objection that *all* options could result in human extinction; hence, the fact that one option has some chance of resulting in extinction need not imply that it should be ruled out, according to the lexical utility model. However, the problem is that the lexical utility model implies that, if the two available options can both result in whatever is designated as a (terminal) catastrophe, then the option that has a lower chance of resulting in that catastrophe should be chosen.[33] This means that we are back in the same trilemma. Either a tiny difference in the risk of a catastrophe has extreme axiological effect; or we impose a vague *de minimis* condition—applied to either differences in probabilities or absolute probabilities—in a way that leads to choice cycles or paralysis.

4 Concluding remarks

The lexical utility decision rule appears to face a trilemma. Either it is extremely sensitive to trivial value differences, or else it leads to choice cycles or paralysis. But, to conclude, I would like to revisit the point I made in the introduction about the problems raised in this paper being merely theoretical. To derive the problematic results, I had to make some presumably unrealistic assumptions, for instance, about the exactness of the predicted value of the variable underlying the harm threshold, and about the types of alternatives with which a policy maker can be faced. So, despite the arguments in this paper, Bartha, DesRoches, and Steel may well be correct that the lexical utility model leads to better results than its alternatives in many practically important situations, in particular given the resource- and time-constraints of ordinary decision-makers. As previously mentioned, the standard expected utility model may for instance require information that is not available and computations that are infeasible. Still, the results in this paper highlight some theoretical costs of the lexical utility model, and identifies some situations where its use requires extra caution.

[33] The lexical utility model thus implies Bostrom's (2013) maxipok rule (when the catastrophe is interpreted as human extinction).

Acknowldegements Thanks to Paul Bartha for very helpful correspondence about the topic of this paper, and to the audiences of Incommensurability and Population-Level Bioethics (Rutgers University, May 2022) and Pandemic Ethics (Institute for Futures Studies, September 2023) for useful comments and suggestions. Finally, thanks to two reviewers for helping me improve the paper. Financial support from Riksbankens Jubileumsfond is gratefully acknowledged.

Funding Open access funding provided by Stockholm University.

References

Adler, M. D. (2007). Why de minimis? University of Pennsylvania, Institute for Law and Economics, Research Paper no. 07-12.

Andersson, H. (2022). Spectrum arguments, indeterminacy, and value superiority. In H. Andersson & A. Herlitz (Eds.), *Value Incommensurability: Ethics, Risk, and Decision-Making* (pp. 109–125). Routledge.

Andreou, C. (2020). Dynamic choice. In E. N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition.

Arrhenius, G., & Rabinowicz, W. (2005). Millian superiorities. *Utilitas, 17*(2), 127–146.

Arrhenius, G., & Rabinowicz, W. (2015). Value superiority. In I. Hirose & J. Olson (Eds.), *The Oxford Handbook of Value Theory* (pp. 225–248). Oxford University Press.

Bartha, P., & DesRoches, C. T. (2021). Modeling the precautionary principle with lexical utilities. *Synthese, 199*(3–4), 8701–8740.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy, 4*(1), 15–31.

Broome, J. (2013). A small chance of disaster. *European Review, 21*(S1), S27–S31.

Buchak, L. (2013). *Risk and Rationality*. Oxford University Press.

Chang, R. (2013). Commitments, reasons, and the will. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics.* (Vol. 8). Oxford University Press.

Christiansen, A. (2019). Rationality, expected utility theory and the precautionary principle. *Ethics, Policy and Environment, 22*(1), 3–20.

Eyal, N. (2020). Is there an ethical upper limit on risks to study participants? *Public Health Ethics, 13*(2), 143–156.

Gardiner, S. M. (2006). A core precautionary principle. *Journal of Political Philosophy, 14*(1), 33–60.

Goldner Lang, I. (2021). "Laws of Fear'' in the EU: The precautionary principle and public health restrictions to free movement of persons in the time of COVID-19. *European Journal of Risk Regulation*. https://doi.org/10.1017/err.2020.120

Hartzell-Nichols, L. (2013). From 'the' precautionary principle to precautionary principles. *Ethics, Policy and Environment, 16*(3), 308–320.

Lee-Stronach, C. (2018). Moral priorities under risk. *Canadian Journal of Philosophy, 48*(6), 793–811.

Lundgren, B., & Stefánsson, H. O. (2020). Against the de minimis principle. *Risk Analysis, 40*(5), 908–914.

Peterson, M. (2002). What is a de minimis risk? *Risk Management, 4*(2), 47–55.

Peterson, M. (2006). The precautionary principle is incoherent. *Risk Analysis, 26*(3), 595–601.

Pummer, T. (2018). Spectrum arguments and hypersensitivity. *Philosophical Studies, 175*(7), 1729–1744.

Pummer, T. (2022). Sorites on what matters. In J. McMahan, T. Campbell, K. Ramakrishnan, & J. Goodrich (Eds.), *Ethics and Existence: The Legacy of Derek Parfit* (pp. 498–523). Oxford University Press.

Raposo, V. L. (2022). Quarantines: Between precaution and necessity. A look at COVID-19. *Public Health Ethics, 14*, 35–46.

Sandin, P., & Peterson, M. (2019). Is the precautionary principle a midlevel principle? *Ethics, Policy and Environment, 22*(1), 34–48.

Smith, M. (2022). Decision theory and de minimis risk. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00624-9

Steel, D. (2015). *Philosophy and the Precautionary Principle: Science, Evidence, and Environmental Policy*. Cambridge University Press.

Steel, D., & Bartha, P. (2023). Trade-offs and the precautionary principle: A lexicographic utility approach. *Risk Analysis, 43*(2), 260–268.

Stefánsson, H. O. (2019). On the limits of the precautionary principle. *Risk Analysis, 39*(6), 1204–1222.

Sunstein, C. R. (2005). *Laws of Fear: Beyond the Precautionary Principle. The Seeley Lectures*. Cambridge University Press.