Continuity and Catastrophic Risk

H. Orri Stefánsson[*]

Abstract

Suppose that a decision-maker's aim, under certainty, is to maximise some continuous value, such as lifetime income or continuous social welfare. Can such a decision-maker rationally satisfy what has been called "continuity for easy cases" while at the same time satisfying what seems to be a widespread intuition against the full-blown continuity axiom of expected utility theory? In this note I argue that the answer is "no": given transitivity and a weak trade-off principle, continuity for easy cases violates the anti-continuity intuition. I end the note by exploring an even weaker continuity condition that is consistent with the aforementioned intuition.

KEYWORDS: *Continuity, Catastrophic Risk, Rationality, Trade-off, Transitivity*

1 Introduction

The *continuity* axiom of expected utility theory entails that no outcome is so bad that one should be unwilling to risk it for the sake of a high enough chance of a marginal improvement on the status quo. More precisely, the axiom states that for any triple of outcomes, A, B, and C, such that A is preferred to B which is preferred to C, there is some probability p such that a rational decision-maker is indifferent between on the one hand getting B for sure and on the other hand a lottery (or gamble) that with probability p results in A but otherwise (with probability 1-p) results in C.

Many people seem to have the intuition that continuity cannot hold in full generality. For instance, what if B is some outcome where you enjoy a high

[*] Department of Philosophy, Stockholm University, Universitetsvägen 10 D, 114 18 Stockholm, Sweden; and Swedish Collegium for Advanced Study, Thunbergsvägen 2, 752 38 Uppsala, Sweden. Email: orri.stefansson@philosohy.su.se. URL: www.orristefansson.is.

level of lifetime well-being and A is only marginally better than B, while C is an outcome that involves your death? (Luce and Raiffa 1967: 27) Or, to take another example, what if A, B, and C denote lifetime income, where B is some very high income and A is income B plus an added dollar each year, whereas C is a lifetime income of $0 per year? (Temkin 2001)

Those who hold what I shall call the anti-continuity judgement typically take some examples like these to undermine continuity. The aim of this note is to explore whether continuity can be weakened, to satisfy the anti-continuity judgement, while still assuming that the decision-maker of interest wants (under certainty) to maximise some continuous value. So, an example of the type of decision-maker that I shall be concerned with is one whose aim is to maximise their lifetime income. Another example is a social planner whose aim is to maximise some continuous measure of "social welfare".

That the above two decision-makers want to maximise some continuous value (income or continuous social welfare) as far as risk-free *outcomes* are concerned, does not imply that they must satisfy the continuity axiom when evaluating lotteries. More generally, an *outcome* axiology that maximises some continuous value is consistent with a *lottery* axiology that violates the continuity axiom of expected utility theory.

My inquiry is inspired by the discussion of Temkin (2001), Arrhenius and Rabinowicz (2005), and Jensen (2012). They explore weakening continuity, as I do below, and find that, given some additional assumptions, the full-blown continuity axiom of expected utility theory follows. However, these authors assume what is essentially the independence axiom of expected utility theory,[1] which I think we have independent reason to reject (as I have argued in Stefánsson and Bradley 2015). Hence, I shall not assume independence.

However, I show that essentially the same result can be obtained by replacing independence by some weak trade-off assumptions, which seem hard to deny given that the decision-maker of interest wants to maximise some continuous value as far as risk-free outcomes are concerned. I end the note by examining even weaker continuity conditions than the above-mentioned

[1] Informally, independence states that when comparing two risky prospects, one can ignore outcomes on which the two prospects confer the same probability. Strictly speaking, the above-mentioned authors do not assume independence, but a condition that entails independence assuming that the preference relation is a weak order.

authors consider, and find that these conditions are consistent with the anti-continuity judgement.

2 The case against continuity

Let $\mathbf{O}$ be a set of outcomes, and let $[A, p, C]$ be a "lottery" that with probability p results in A and with probability $1 - p$ results in C. $\prec$ denotes the strict preference relation, $\sim$ the indifference relation (and later I'll use $\precsim$ for the weak preference relation). $[0, 1]$ denotes the *closed* interval between (and including) 0 and 1; $(0, 1)$ the corresponding *open* interval (which neither contains 0 nor 1 but does contain every number in between 0 and 1). The version of continuity with which I will be primarily concerned—which is entailed by expected utility theory—can now be stated as:

Axiom (Continuity)**.** *For any $A, B, C \in \mathbf{O}$ such that $C \prec B \prec A$, there is a $p \in (0, 1)$ such that:*

$$B \sim [A, p, C]$$

In addition to the two *personal* catastrophes discussed in the introduction, some examples involving *social* catastrophes might seem to undermine continuity. An example of a social catastrophe is an *existential catastrophe*, which we can define as an outcome that brings about "the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development" (Bostrom 2013: 15). Next consider:

Example. *Supposes that C is an existential catastrophe while B is some fantastic social outcome, and let A be some social outcome that is only marginally better than B (perhaps B plus one person living an additional minute of happy life). Would a rational social planner necessarily be willing to give up B for a (non-trivial, that is, p strictly between zero and one) lottery between A and C?*

"No", someone who worries about existential risk might say[2]—which would then be another instance of what I called the "anti-continuity judgement". More generally, the judgement is that some outcome C is so catastrophic

[2] For instance, Bostrom's *maxipok* rule, which instructs us to maximise the probability avoiding existential catastrophe (Bostrom 2013: 19), violates continuity when existential risk is concerned. Applied to the above example, maxipok selects B over any non-trival lottery between C and A, since any such lottery reduces the probability of avoiding existential catastrophe.

that, for some outcomes A and B, where A is better than B which is better than C, one would never be indifferent between B and a non-trivial lottery between A and C. More formally:

Anti-continuity judgement. *For some* $A, B, C \in \mathbf{O}$ *such that* $C \prec B \prec A$, *there is no* $p \in (0, 1)$ *such that:*

$$B \sim [A, p, C]$$

In response to the anti-continuity judgement, it could be pointed out that we constantly take risks that increase the chance of a catastrophe. For instance, realising that we dropped a £5 bill before crossing the street, we might go back to pick it up (A), which at best results in an outcome that is marginally better than our life without the £5 (B), but also yields some chance p that we will be killed by a car while crossing the street (C).

However, the fact that we do take such risks does not mean that it would *never* be *rationally permissible* for a decision-maker to violate continuity (Arrhenius and Rabinowicz 2005).[3] The aim for the reminder of this paper is to explore the idea that the anti-continuity judgement is rationally permissible[4]—that is, to examine whether violating continuity can be rationally permissible—by seeing whether it can be sustained when combined with assumptions and principles that even those who hold the anti-continuity judgement should accept.

3 Weakening continuity

The problem with continuity—if there is a problem—might be that it is meant to hold for *any* A, B, C such that $C \prec B \prec A$, no matter how much worse C is than B compared to how much (or little) worse B is than A. As a first attempt, we can weaken continuity in light of this potential problem. Start by ranking the outcomes in $\mathbf{O}$ and label them accordingly, from the worst outcome, O_1, to the best, O_n. (Two equally good outcomes are treated as one and the same

[3] Moreover, it is, of course, possible that the potential catastrophe doesn't even occur to people when engaging in this type of behaviour.

[4] Note that the anti-continuity judgement itself is not about rationality; it is simply a claim about preference between outcomes and lotteries that might be true of some agents but not others. I thank a referee for making me see the need to clarify this.

outcome.)[5] I shall assume that for each decision-maker we consider, O_1 is catastrophic while O_n is very good.

Now consider the following weakening of continuity, which Arrhenius and Rabinowicz (2005) call *continuity for adjacent elements*, and which by itself is consistent with the anti-continuity judgement:

Axiom (Continuity for adjacent elements)**.** *For any triple $O_{i-1}, O_i, O_{i+1} \in \boldsymbol{O}$ there is a $p \in (0,1)$ such that:*

$$O_i \sim [O_{i+1}, p, O_{i-1}]$$

Continuity for adjacent elements is similar to Temkin's (2001) "continuity for easy cases", although Temkin states his principle in terms of some stipulated quantitative measure of how good/bad are the outcomes. Arrhenius and Rabinowicz (2005) and Jensen (2012) show that *given independence* and transitivity,[6] continuity for adjacent elements implies continuity. But, as previously mentioned, I don't want to assume independence.

However, by assuming a weak *trade-off* principle and transitivity of indifference (see fn. 6), continuity for adjacent elements violates the anti-continuity judgement. Consider:

Assumption 1 (Downward trade-off)**.** *For any triple $O_{j-1}, O_j, O_i \in \boldsymbol{O}$, where $j < i$, and for any $p \in (0,1)$, there is a $q \in (0,1)$ such that:*

$$[O_i, p, O_j] \sim [O_i, q, O_{j-1}]$$

Downward trade-off could be viewed as a structural constraint on the *set of outcomes* (to adapt Suppes 2002 terminology), not a rationality constraint on preference. But it is hard to see how a decision-maker of the type that interests us here could deny it. Whether the decision-maker is interested in maximising

[5] Although I will be working with a set of outcomes that corresponds to *finitely* many value levels—in other words, I am assuming that there is a finite number of equivalence classes in **O** generated by $\prec$—my argument is consistent with the set of *all possible* outcomes corresponding to infinitely many value levels. (We can, for instance, think of **O** as the set of outcomes that actually could result from the decision-problems of interest.) Moreover, a simple re-interpretation of the principles and assumptions that I introduce below make them consistent with **O** being infinite. (I thank John Broome for making me see the need to emphasise this.)

[6] Transitivity states that if A is (weakly) preferred to B which is (weakly) preferred to C, then A should be (weakly) preferred to C. Transitivity of indifference, which will play a role in my argument, says that if one is indifferent between A and B and also between B and C, then one should be indifferent between A and C.

lifetime income or continuous social welfare, we can safely assume that for any outcome O_j, there is an adjacent outcome O_{j-1} that is only a tiny bit worse, according to this decision-maker. Given how fine-grained probabilities are, it would seem plausible that a decision-maker of this kind should then accept that, for any $p \in (0, 1)$, there is a (greater) $q \in (0, 1)$ such that they would be indifferent between lotteries like those in the above assumption.

However, given transitivity of indifference for lotteries (including trivial ones, that is, outcomes), downward trade-off implies something that, although not quite continuity, violates the anti-continuity judgement. (In the next section I will add an assumption similar to downward trade-off which suffices to derive continuity.)

Take the three best outcomes, $O_{n-2} \prec O_{n-1} \prec O_n$. Continuity for adjacent elements entails that there is some q such that:

$$O_{n-1} \sim [O_n, q, O_{n-2}]$$

Downward trade-off then implies that there is some q' such that:

$$[O_n, q, O_{n-2}] \sim [O_n, q', O_{n-3}]$$

Repeatedly applying downward trade-off, we eventually find that there are some probabilities q'' and q''', both in $(0, 1)$, such that:

$$[O_n, q'', O_2] \sim [O_n, q''', O_1]$$

Applying transitivity of indifference then gives us:

$$O_{n-1} \sim [O_n, q''', O_1]$$

In other words, there is some $q''' \in (0, 1)$ such that the decision-maker would be indifferent between the second best outcome and a lottery that with probability $q''' < 1$ results in the best outcome and with probability $1 - q'''$ results in the worst outcome.

More generally, from downward trade-off, continuity for adjacent elements and transitivity of indifference we have derived:[7]

[7] Note that if we in addition assume *Stochastic dominance*, then we can drop the "adja-

Implication. *For any adjacent O_i and O_{i+1}, and for any $j < i$, there is some $p \in (0, 1)$ such that:*

$$O_i \sim [O_{i+1}, p, O_j]$$

So, for *any* pair of adjacent outcomes, the decision-maker is willing to give up the certainty of the less preferred one in the pair for a lottery between the more preferred one in the pair and the worst outcome in **O**. In particular, the decision-maker is willing to give up any fantastic social outcome for a gamble between, on the one hand, the smallest possible improvement on the fantastic outcome, and, on the other hand, the worst outcome—in violation of the anti-continuity judgement.

In section 5, I shall weaken both continuity for adjacent elements and downward trade-off by appealing to a threshold (intuitively, a threshold below which an outcome becomes catastrophic) that blocks the above implication. But first, in the next section, I shall examine how we can get the full continuity axiom from continuity for adjacent elements and downward trade-off. For note that while the above implication violates the anti-continuity judgement, it does fall short of the fully general continuity axiom. But as we shall see, an assumption very similar to downward trade-off suffices to derive continuity.

4 Deriving continuity

Consider the following assumption, which, although logically independent of downward trade-off, seems at least as hard to deny as that assumption and for similar reasons:

Assumption 2 (Upward trade-off)**.** *For any triple $O_j, O_i, O_{i+1} \in \mathbf{O}$, where $j < i$, and for any $p \in (0, 1)$, there is a $q \in (0, 1)$ such that*

$$[O_i, p, O_j] \sim [O_{i+1}, q, O_j]$$

To see that this added assumption suffices to derive continuity, take now any adjacent $O_{j-1}, O_j, O_{j+1} \in \mathbf{O}$. By continuity for adjacent elements, there is a

cent" qualification, but the implication will then be in terms of weak preference rather than indifference. (Thanks to Christian Tarsney for pointing this out.)

$q \in (0,1)$ such that:

$$O_j \sim [O_{j+1}, q, O_{j-1}]$$

By downward trade-off, there is a $q' \in (0,1)$ such that:

$$[O_{j+1}, q, O_{j-1}] \sim [O_{j+1}, q', O_{j-2}]$$

By upward trade-off, there is a $q'' \in (0,1)$ such that:

$$[O_{j+1}, q', O_{j-2}] \sim [O_{j+2}, q'', O_{j-2}]$$

Repeating this type of reasoning we find that for any $O_i, O_k \in \mathbf{O}$, where, for O_j from the first line in the argument, $O_i \prec O_j \prec O_k$, there is a $p \in (0,1)$ such that:

$$[O_{j+1}, q, O_{j-1}] \sim [O_k, p, O_i]$$

So, by transitivity of indifference:

$$O_j \sim [O_k, p, O_i]$$

In other words, we have established:

Observation. *Continuity for adjacent elements, transitivity of indifference, downward trade-off and upward trade-off together imply continuity.*

5 Further weakening

Perhaps some even want to deny continuity for adjacent elements.[8] So let's weaken it even further. To that end, let $\mathbf{C}$ be the (possibly fuzzy) set of catastrophic outcomes and assume there to be an η such that for any $O_k \in \mathbf{O}$, if $k \geq \eta$ then $O_k \notin \mathbf{C}$.

Axiom (Weakened continuity for adjacent elements)**.** *For any triple $O_{i-1}, O_i, O_{i+1} \in \mathbf{O}$ such that $O_{i-1} \notin \mathbf{C}$, there is a $p \in (0,1)$ such that:*

$$O_i \sim [O_{i+1}, p, O_{i-1}]$$

[8] In fact, Arrhenius and Rabinowicz (2005) deny continuity for adjacent elements, and end their article on remarks about borderline cases that may support my next weakening.

Those who hold the anti-continuity judgement due to perceived discontinuity between catastrophic and non-catastrophic outcomes will presumably find weakened continuity for adjacent elements to be an improvement over continuity for adjacent elements. However, it is straightforward to see (from the above argument) that weakened continuity for adjacent elements is inconsistent with the anti-continuity judgement in the presence of transitivity of indifference and downward trade-off.

There is though a simple weakening of downward trade-off that will allow us to accommodate the anti-continuity judgement while accepting weakened continuity for adjacent elements and transitivity of indifference:

Assumption 3 (Weakened downward trade-off)**.** *For any triple $O_{j-1}, O_j, O_i \in \boldsymbol{O}$, where $j < i$ and $O_{j-1} \notin \boldsymbol{C}$, and for any $p \in (0,1)$, there is a $q \in (0,1)$ such that:*

$$[O_i, p, O_j] \sim [O_i, q, O_{j-1}]$$

The improvement of weakened downward trade-off over downward trade-off, from the perspective of those who hold the anti-continuity intuition, is that the weakened trade-off principle stops the sequences in the derivations in the last two sections before reaching catastrophic outcomes, meaning that the general and universally quantified continuity condition ("For any $A, B, C \in$ **O**...") cannot be derived.

But even those who hold the anti-continuity judgement might accept that there is some sort of continuity condition that applies even to catastrophic outcomes. To formulate a contender for such a condition, I will suppose that there is some value $\gamma \in [0,1]$ such that:

Definition. *Any O_j, $O_k \in \{O_1, ..., O_n\}$, where $O_j \precsim O_k$, are "sufficiently distant" iff:*

$$\frac{k-j}{n} \geq \gamma$$

Now, there may be problems with an *ordinal* definition of "sufficiently distant", in particular, if the difference in value between different pairs of adjacent elements is not constant. But however one defines "sufficiently distant", even those who hold the anti-continuity judgement might be tempted to accept:

Axiom (Distance continuity)**.** *For any $O_i, O_j, O_k \in$ **O** such that $O_i \prec O_j \prec O_k$, and*

*$O_i \in$ **C**, $O_j \notin$ **C**, there is a $p \in (0,1)$ such that:*

$$O_j \sim [O_k, p, O_i]$$

if and only if O_j, O_k are sufficiently distant.

Even the "if" part of distance continuity generally seems plausible given a high value of γ (i.e., close to 1).[9] Suppose, for instance, that $\gamma = (n-2)/n$. Then distance continuity only states that the decision-maker is indifferent between the second worst outcome and some lottery between the best and the worst outcome. (The higher γ is, the closer O_i is to O_j.) For this instance of distance continuity to to fail to hold, the difference in value between the second worst outcome and the best outcome would have to be surprisingly small in comparison to the difference in value between the second worst outcome and the worst outcome.

Let us now consider a toy example that further illustrates how distance continuity works. The example will moreover serve as an illustration of why weakened continuity for adjacent elements, weakened downward trade-off, upward trade-off, transitivity, distance continuity, and the anti-continuity judgement can be satisfied in a "non-trivial" way, that is, in an example where all of these conditions *sometimes* apply.[10]

Let $\mathbf{O} = \{O_1, ..., O_{10}\}$, $\mathbf{C} = \{O_1\}$ and $\gamma = 0.7$. Then the conditions in distance continuity hold for the following triples: O_1, O_2, O_{10}, and O_1, O_2, O_9, and O_1, O_3, O_{10}. For instance, the axiom ensures that there is a $p \in (0,1)$ such that:

$$O_3 \sim [O_{10}, p, O_1]$$

Applied to this example, weakened continuity for adjacent elements ensures that for any triple $O_{i-1}, O_i, O_{i+1} \in \mathbf{O} \setminus \{O_1\}$, there is a $q \in (0,1)$ such that:

$$O_i \sim [O_{i+1}, q, O_{i-1}]$$

The implication of upward trade-off and weakened downward trade-off is then that, given transitivity of indifference, we can derive continuity for all

[9] For distance continuity to ever apply, we however must assume that: $\gamma \leq (n-2)/n$.

[10] I thank a referee for encouraging me to include an example like this in the paper.

elements in $\mathbf{O} \setminus \{O_1\}$, that is, not only for adjacent ones. However, we cannot derive this for the full set $\mathbf{O}$; the weakening of downward trade-off prevents the sequence from entering $\mathbf{C} = \{O_1\}$, as it were. Another way to put this, is that the weakening of downward trade-off puts a floor on the sequence of continuity-type judgements that we can derive from continuity for adjacent elements. So, we cannot derive the fully general continuity condition, and thus the anti-continuity judgement is satisfied.

There may nevertheless be reasons to think that those who hold the anti-continuity judgement would not accept distance continuity.[11] Perhaps the judgment is based on the intuition that any lottery that confers a positive probability on a catastrophe is worse than a sufficiently good status quo *even if that lottery provides a chance of an arbitrarily large improvement*. I personally find that too extreme. Nevertheless, it could be that although formally, distance continuity is consistent with the anti-continuity judgement (even assuming the aforementioned assumptions and axioms), the intuition underlying the anti-continuity judgement conflicts with distance continuity.

6 Concluding remarks

To summarise, the following axioms and assumptions form a consistent set and can together accommodate the anti-continuity judgement:

- Distance continuity
- Weakened continuity for adjacent elements
- Upward trade-off
- Weakened downward trade-off
- Transitivity of indifference

In contrast, we have seen that, for decision-makers whose aim is to maximise some continuous quantity under certainty, one cannot plausibly replace the traditional continuity axiom with continuity for adjacent elements without violating the anti-continuity judgement.

[11] I am grateful to Wlodek Rabinowicz for making me see this.

ACKNOWLEDGMENTS
This paper has been presented at the Work in Progress seminar at the Global Priorities Institute (June 2020). I am very grateful to the participants for their helpful comments and suggestions. I also received very useful comments on the paper from John Broome, Daniel Ramöller, and Wlodek Rabinowicz. Financial support from Riksbankens Jubileumsfond is gratefully acknowledged.

BIOGRAPHICAL INFORMATION
H. Orri Stefánsson is an associate professor of practical philosophy at Stockholm University, Pro Futura Scientia fellow at the Swedish Collegium for Advanced Study, and advisor at the Institute for Futures Studies. He currently works on decision-making under extreme uncertainty, population ethics, and climate ethics.

References

Arrhenius, G. and Rabinowicz, W. (2005). Value and unacceptable risk. *Economics and Philosophy*, 21(2):177–197.

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

Jensen, K. K. (2012). Unacceptable risks and the continuity axiom. *Economics and Philosophy*, 28(1):31–42.

Luce, R. D. and Raiffa, H. (1967). *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.

Stefánsson, H. O. and Bradley, R. (2015). How valuable are chances? *Philosophy of Science*, 82(4):602–625.

Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. Stanford: CSLI publications.

Temkin, L. (2001). Worries about continuity, expected utility theory, and practical reasoning. In D. Egonsson et.al, editor, *Exploring Practical Philosophy: From Action to Values*, pages 95–108. London: Ashgate Publishers.