Theory and Decision (2020) 88:595–607
https://doi.org/10.1007/s11238-019-09740-5

Resolving Zeckhauser's paradox

Yudi Pawitan[1] · **Gabriel Isheden**[1]

Published online: 17 December 2019

Abstract
Zeckhauser's paradox has puzzled and entertained many rationality enthusiasts for almost half a century. You are forced to play a Russian Roulette with a 6-chamber revolver containing either (A) two bullets, or (B) four bullets. Would you pay more to remove the two bullets in (A) than you would to remove one in (B)? Most would say yes, but rational considerations based on the classical utility theory suggest you should not. We discuss a possible solution within the classical framework, by explicitly stating and accounting for more detailed preferences in terms of fewer bullets and smaller debt. To a large extent, the paradox arises due to a surreptitious trespassing of Savage's Small-World utilities implied by a limited set of preferences to govern a larger world containing potentially conflicting preferences. To avoid logical issues associated with death in the roulette, we also describe a non-fatal game-show version, where you choose one box out of six that could be either empty or contain prize money. Here, the paradox arises when you pay from the prize money, but not when you pay from your own money. In summary, the paradox provides a useful lesson about the normative role of the utility function as a rational guide for our decisions and preferences.

Keywords Preference graph · Probability · Rationality · Russian Roulette · Utility · Von Neumann and Morgentern

1 Introduction

A terrorist group puts you in a Russian Roulette: a gun to your head with some bullets inside. Assume it is a standard-issue revolver with 6-bullet chambers. As part of their entertainment, they tell you there are two possibilities: (A) there are two bullets in the gun, or (B) there are four bullets. They ask, are you willing to pay more for removing the two bullets in (A) than for removing one in (B)? If you are like other normal human beings, you would say an emphatic yes. However, a

✉ Yudi Pawitan
yudi.pawitan@ki.se

[1] Karolinska Institutet, Stockholm, Sweden

calculation based on the classical utility theory implies that you are being inconsistent about money, which is close enough to being irrational. That is called Zeckhauser's paradox. Even mathematicians, economists and game theorists report that they react like normal humans. But, perhaps unlike normal humans, they admit that they make a mistake. See, for example, a quote from Binmore (2009, Chap. 3):

> I made this mistake myself when the paradox was first put to me, but I changed my mind when I learned that my decision was inconsistent with my preferring being alive to being dead and having more money rather than less.

And another from Landsburg (2011)'s blog:

> Did you pass the test? I for one did not. That leaves me (and also you, if you failed along with me) two options. Either we can maintain that there's some flaw in the [...] argument - some way in which it fails to capture the "right" meaning of rationality - or we can conclude that we don't always make good decisions, and that meditating on our failures can help us make better decisions in the future (including in situations more likely to arise than forced Russian Roulette). I am mostly in the latter camp.

The philosopher Jeffrey (1988) 'lingered by the rail for some time,' but eventually 'bit the Bayesian bullet,' i.e. followed the direction prescribed by the utility. He proposed that a good part of rationality is strength of character: fortitude or steadfastness or nerve. So for him the 'deviant preference' is irrational, though in this case 'the irrationality is no intellectual flaw, but a characterological one, i.e. not stupidity but funk.' In summary, from all these reactions, if you do not follow the action prescribed by the utility, you are either irrational or lacking in character. Neither is admirable.

We aim to show that the normal human reaction is rational in accordance with the classical utility theory, as long as we state our preferences explicitly and account for them in deriving the utilities. We also aim to extract from the paradox a general lesson about the normative role of the utility function as a rational guide of our decisions and preferences.

2 Standard calculation

For clarity and to establish notations, let us review the standard calculation that leads to the paradox. Let C_k be the condition of being under the gun with k bullets inside, for $k = 0, \ldots, 5$. To reduce complexity later, we exclude $k = 6$. By definition C_0 just means 'Alive'. Let us set the base utilities $u(\text{ Alive }) \equiv 1$ and $u(\text{ Dead }) \equiv 0$. We can then compute

$$u(C_k) = \frac{k}{6} u(\text{ Dead }) + \frac{6-k}{6} u(\text{ Alive }). \tag{1}$$

$$= \frac{6-k}{6}, \quad \text{for } k = 1, \ldots, 5. \tag{2}$$

Now, agreeing to pay x to remove two bullets in (A) creates a new condition C_0^*, which means 'Alive but with a debt x', with an equal utility to C_2. That is, we must have

$$u(C_0^*) = u(C_2) = 4/6.$$

It is reasonable that $u(C_0^*) < u(C_0) = 1$, since in C_0^* you are alive with a debt x. (We prefer the shorter word 'debt' to the perhaps more appropriate 'liability'.) Agreeing to pay y to remove one bullet from a four-bullet gun means you're now in C_3^*, which is a three-bullet state plus a debt of y. The indifference relative to C_4 means

$$u(C_3^*) = u(C_4) = 2/6, \tag{3}$$

which is again reasonable to be less than $u(C_3) = 3/6$, since you are in condition C_3 *plus* having a debt y. So far there is no paradox.

Let A^*d mean 'Alive with debt d'. From the above

$$u(A^*x) = u(C_0^*) = 4/6,$$

and from (1) and (3) we have

$$u(C_3^*) = \frac{3}{6}u(A^*y) = 2/6,$$

giving the final reckoning:

$$u(A^*y) = 2/3 = u(A^*x). \tag{4}$$

Now, this is the paradox: You intuitively prefer $x > y$, but (4) implies you are then indifferent between being alive with a larger debt x and alive with a smaller one y! Any sensible person with a mortgage would think that is pretty irrational. If you interpret the utility normatively, the rational choice is to pay the same amount in (A) and in (B). We can also derive other surprising propositions. Since removing 2 out of 2 bullets should be worth more than removing 1 out of 1, this means you should pay more to remove 1 out of 4 bullets than to remove 1 out of 1 bullet. With similar calculations, you should pay the same amount to remove 1 out of 5 bullets as to remove 3 out of 3 bullets. 'Resolving the paradox' must include resolving all these propositions.

3 Resolution

The key weakness in the standard calculation is that, without any warning, it fixes *for everyone* the utilities of the different conditions C_k's in (2). We know, however, that different people have different levels of tolerance for danger—some may even invite danger comparable to playing the roulette, but we seem to have no way of adjusting the utilities. Formula (2) becomes set because the calculation makes no distinction between having survived from different roulette conditions, i.e. by setting $u(\text{ Alive }) \equiv 1$. In real life, you have a greater sense of relief from surviving

a high-risk event, such as a serious car accident, compared to surviving a low-risk event, such as a sprained ankle.

So, let us define the survival status more carefully, keeping track of what situation you have survived from: A_k means 'being alive or surviving the Roulette with k-bullet gun.' Immediately, instead of (1), we have a new formula

$$u(C_k) = \frac{k}{6}u(\text{ Dead }) + \frac{6-k}{6}u(A_k). \tag{5}$$

$$= \frac{6-k}{6}u(A_k), \quad \text{for } k = 1, \ldots, 5. \tag{6}$$

Following the same reasoning, let C_k^* mean 'under the gun with k bullets plus having a debt', and A^*d_k 'surviving the k-bullet gun plus having a debt d.' The equivalent relationships are now

$$u(C_0^*) = u(C_2) = \frac{4}{6}u(A_2) = u(A^*x_0) \tag{7}$$

$$u(C_3^*) = u(C_4) = \frac{2}{6}u(A_4) = \frac{3}{6}u(A^*y_3). \tag{8}$$

Without any further specifications on $u(A_k)$'s, we have no obvious paradoxical relationship between $u(A^*x_0)$ and $u(A^*y_3)$ as implied by (4).

To be specific, let us assume the following preferences. Other preferences are possible and they lead to different utilities.

(P1) Fewer bullets in the gun is always preferable to more bullets. As a corollary, as long as $m > n$, you would pay more (i.e. prefer) to reduce the number of bullets from any number to n bullets, than from any number to m bullets. When $m = n$, you would of course pay more to remove more bullets. For example, in the extreme, reducing from 1 to 0 bullet is preferable to reducing from 5 to 1 bullet. But, reducing from 5 to 1 bullet is preferable to reducing from 2 to 1 bullet.

(P2) Being alive with a smaller debt is better than being alive with a bigger debt.

As in the standard calculation, assume that when you are dead you have no further preferences. Let us translate these preferences in terms of utilities; it is clear that if there exists a utility function that agrees with these preferences, then the utility is free of Zeckhauser's paradox. Suppose there are j bullets in your gun, and you agree to pay x_{ji} to reduce the number of bullets to i. Keeping track of the number of bullets, denote by C_{ji}^* the condition after the payment agreement. Then, the following equation holds, for $i \leq j$ and $i = 0, \ldots, 5$, $j = 0, \ldots, 5$,

$$u(C_{ji}^*) = u(C_j) = \frac{6-j}{6}u(A_j) = \frac{6-i}{6}u(A^*x_{ji}), \tag{9}$$

which means that the utility of each indebted survival is worth

$$u(A^*x_{ji}) = \frac{6-j}{6-i}u(A_j). \tag{10}$$

All preferences implied by P1 and P2 can be expressed in terms of A^*x_{ji}'s. There are $6 \times 7/2 = 21$ of such terms, generating $21 \times 20/2 = 210$ preferences. In Sect. 5, we describe a general graph-theoretic approach to analyse a large number of preferences for (i) testing the existence of Zeckhauser-like paradox, and (ii) computing a specific utility function. In the current example, we can show that the set of preferences is equivalent to the following system of inequalities:

$$\begin{aligned}
&u(C_0) > u(C_1) > \cdots > u(C_5)\\
&u(A^*x_{10}) > u(A^*x_{51})\\
&u(A^*x_{21}) > u(A^*x_{52})\\
&u(A^*x_{32}) > u(A^*x_{53})\\
&u(A^*x_{43}) > u(A^*x_{54}).
\end{aligned}$$

We show by way of an example the existence of a utility function that satisfies those inequalities:

$u(C_0)$	$u(C_1)$	$u(C_2)$	$u(C_3)$	$u(C_4)$	$u(C_5)$
1	0.86	0.84	0.81	0.77	0.72

(11)

They correspond to the following utilities:

$u(A_0)$	$u(A_1)$	$u(A_2)$	$u(A_3)$	$u(A_4)$	$u(A_5)$
1	1.032	1.26	1.62	2.31	4.32

If you have previously seen Zeckhauser's paradox and the standard calculation, you may feel there is something fishy here. In Sect. 6, we state the formal decision theory framework to make it clear that we do stay within the classical utility theory. Interestingly, the preference assumptions P1 and P2 lead to an increasing $u(A_k)$ over k. Theoretically, there is no problem with having $u(A_k) > 1$ for $k > 0$, since we can just re-scale all values to get a maximum 1. We have $u(A_k) > u(A_{k-1})$; in contrast, in the standard calculation, $u(A_k) = 1$ for all k. For $k = 2$, say, the inequality $u(A_2) > u(A_0)$ means that having survived a two-bullet Roulette you are now valuing your life more than if you have never gone through it. This seems like an extra-ordinary solution—in the literal sense of being out of the ordinary—but it is a fitting solution to an extraordinary paradox that has tricked so many people.

4 Money versions

On reading numerous blogs about the paradox, including the even more numerous commentaries in them, we found some common themes:

- Virtually everyone accepts that they 'fail' the test, i.e. choosing to pay more for the two bullets from the 2-bullet gun than for one bullet from the 4-bullet gun.

- Those familiar with the expected utility theory typically trust the theory more than they trust their own intuition.
- Many people find it hard to put the utility value on being alive or dead, let alone on being dead after paying for the bullet.

The last point is not surprising. So, let us take the gun out of the picture and imagine instead that you're a contestant in a game show. You are shown 6 opaque boxes and told that k of them are empty, and each of the others contains a big prize money $\$M$. You are to pick one box. Losing a big prize is painful but not fatal, so it should not affect your critical thinking. Here are the corresponding scenarios parallel to the Russian Roulette:

(A) There are $k = 2$ empty boxes. How much of the prize money would you be willing to pay to turn the two empty boxes into winning ones (so there is no empty boxes, and you are guaranteed to win)?
(B) There are $k = 4$ empty boxes. How much of the prize money—if you win—would you be willing to pay to turn one out of the four empty boxes into a winning one (so there would be 3 empty boxes)? There is no need to pay if you lose.

In both cases, you pay to avoid or reduce the disaster of losing the prize. Would you pay more in (A) than in (B)? If you intuitively say yes, you're in an equivalent situation as in Zeckhauser's paradox.

Because how you feel about money depends on your financial health, let us assume that you have an initial wealth W. Let's agree also that more wealth is preferable, so the utility function is increasing. In this specific example there is no need to assume concavity for risk aversion. Suppose you consider paying x in (A), and paying y in (B), so you have the equalities:

$$\frac{4}{6}u(W+M)+\frac{2}{6}u(W) = u(W+M-x)$$
$$\frac{2}{6}u(W+M)+\frac{4}{6}u(W) = \frac{3}{6}u(W+M-y)+\frac{3}{6}u(W).$$

After some simple algebra, we end up with

$$u(W+M-y)-u(W+M-x)=0.$$

which implies $x = y$, or that you should pay the same amount in (A) and in (B). This is exactly the same result as (4) for the Russian Roulette.

But now, let us suppose that you have to pay from *your own money*, not from the prize money. This leads to the following equalities:

$$\frac{4}{6}u(W+M)+\frac{2}{6}u(W) = u(W+M-x)$$
$$\frac{2}{6}u(W+M)+\frac{4}{6}u(W) = \frac{3}{6}u(W+M-y)+\frac{3}{6}u(W-y).$$

Again, after performing some simple algebra, since $u(\cdot)$ is increasing, we have

$$u(W + M - y) - u(W + M - x) = u(W) - u(W - y) > 0,$$

implying $x > y$, or that you should be willing to pay more in (A) than in (B), which now matches your intuition, hence no paradox. So, comparing the two versions, it is clear that the wrong intuition in the prize money version arises because you are greedily thinking of the prize money as your own. Admittedly this is encouraged by the scenario (A), in which you are guaranteed to win, but the thinking is wrong in (B).

5 General solution for Zeckhauser-like problems

The problem we have discussed so far is analytically tractable as the number of preferences and utilities is relatively small. But how do we deal with thousands of states or preferences? (Technically, in formal decision theory terms, the 'states' are the consequences of an act; see Section 6.) How can we be sure that a Zeckhauser-like problem does not arise? And how do we compute specific utility functions? We describe a graph-theoretic approach to these problems, based on the *preference graph*. Bouyssou and Vincke (2010) give a general account of preference matrices and preference graphs, but they do not discuss the specific results we need for the Zeckhauser problem.

Let $S = \{S_1, S_2, \ldots, S_n\}$ be the set of n states, and $\mathcal{P}$ be the set of k preferences between these states

$$\mathcal{P} = \{(S_{a_1} > S_{b_1}), \ldots, (S_{a_k} > S_{b_k})\},$$

where $\{a_i, i = 1, \ldots, k\}$ and $\{b_i, i = 1, \ldots, k\}$ are the index sets. We want to:

(A) check if the system of states and preferences is free of a Zeckhauser-like problem, i.e. there is no pair of states (i, j) where both $(S_i > S_j)$ and $(S_i < S_j)$ are in $\mathcal{P}$.
(B) find a utility function $u(\cdot)$ that agrees with this set of preferences, if it exists.

Definition Given S and $\mathcal{P}$, the *preference graph* is a directed graph $G = (V, E)$ with a vertex v for every state in S and a directed edge e between the two vertices corresponding to each preference in the preference set, where the direction agrees with the preference.

Assuming the preferences are transitive, that is: $S_1 > S_2$ and $S_2 > S_3 \Rightarrow S_1 > S_3$, then the following theorem holds:

Theorem *Given a set of states S, then the set of state preferences $\mathcal{P}$ is free of the Zeckhauser-like problem if and only if the corresponding preference graph $G = (V, E)$ is free of directed loops.*

Proof We prove this via contradiction. First, if there is a directed loop and no Zeckhauser-like problem, then this loop is equivalent to a set of preferences that contains

$$\{(S_1 > S_2), \ldots, (S_{k-1} > S_k), (S_k > S_1)\}.$$

By transitivity, we have $(S_1 > S_k)$ and by assumption we have $(S_k > S_1)$, which is a contradiction.

Secondly, if there is no loop but the Zeckhauser-like problem is present, then the set of preferences contains

$$\{(S_1 > S_2), (S_2 > S_1)\}.$$

But these preferences form a loop in the preference graph, which is a contradiction.

Thus in practice we can detect the Zeckhauser-like problem by assessing whether the preference graph has a directed loop. A well-known algorithm for finding loops in a directed graph is Tarjan's (1972) strongly connected components algorithm.

Let us now examine the problem of finding a utility function $u(\cdot)$ that satisfies the set of inequalities

$$u(\mathcal{P}) = \{(u(S_{a_1}) > u(S_{b_1})), \ldots, (u(S_{a_k}) > u(S_{b_k}))\}.$$

We say that the utility function agrees with $\mathcal{P}$, and by definition it also agrees with the corresponding preference graph G. One way to solve the inequalities is by linear programming, adding a small $\epsilon > 0$ to the right-hand side of each inequality

$$u(S_{a_i}) \geq u(S_{b_i}) + \epsilon_i$$

and solving the linear system $A\bar{u} \geq \bar{\epsilon}$, where $\bar{u} = (u(S_1), \ldots, u(S_n))$, $\bar{\epsilon} = (\epsilon_1, \ldots, \epsilon_k)$, and A is the matrix determining the inequalities.

If the number of preferences is very large and potentially over-specified, we might want to reduce the problem to a smaller set of inequalities. We can do that using the concept minimal spanning tree from graph theory.

Definition A minimal directed spanning tree is any subgraph $G' = (V, E')$ of $G = (V, E)$ with the same set of vertices V, but a minimal number of edges E', such that, for every pair of vertices (v_i, v_j) in G, if there is a directed path from v_i to v_j, then G' also has a directed path from v_i to v_j.

We now reduce the number of inequalities for the utilities using the following theorem.

Theorem *Let $G' = (V, E')$ be a minimal directed spanning tree of the preference graph G. If there exists a utility function $u(\cdot)$ that agrees with G', then this utility function also agrees with G.*

Proof Assume that there is a utility function $u(\cdot)$ that agrees with G'. Now take any preference $S_i > S_j$ from $\mathcal{P}$. Since this preference is in the preference set, there is a directed edge (v_i, v_j) in G. Now, from the definition of a spanning directed tree, since there is a path from v_i to v_j in G there must also be a path from v_i to v_j in G'. Denote this path $(v_i, v'_1, v'_2, \ldots, v'_{x-1}, v'_x, v_j)$. Since $u(\cdot)$ satisfies the inequalities implied by G', we have

$$u(S_i) > u(S'_1), u(S'_1) > u(S'_2), \ldots, u(S'_{x-1}) > u(S'_x), u(S'_x) > u(S_j),$$

and by transitivity we get $u(S_i) > u(S_j)$, which proves that any inequality in $u(\mathcal{P})$ is satisfied by $u(\cdot)$.

We have now shown that to find a utility function $u(\cdot)$ that agrees with $\mathcal{P}$, it is sufficient to (i) create the preference graph G related to $\mathcal{P}$; (ii) find a minimal directed spanning tree G' of G, for example using Edmonds' algorithm (Chu 1965); (iii) solve the reduced system of linear inequalities, for example using linear programming. This is essentially the approach we took when solving Zeckhauser's paradox in Sect. 3.

6 Discussion and conclusion

If you consider Zeckhauser's problem as a rationality test and admit to 'failing' it, then it means that for you the utility function has primacy over your preferences. The problem arises in this case as an implied preference derived from the utility contradicts your—and most people's—seemingly natural preference. Should you adjust the utility or your preference? In general, this is indeed tricky: part of being rational is to accept that you are fallible, so it is tempting to use the utility as a rationality calculator and to follow its exact-looking answers. This mirrors what happens in arithmetic, where you simply do not trust your intuition to get 68×73 and just follow a calculator. But how do you know that you have not gone too far in accepting the primacy of the utility over your 'deviant preferences'?

Where does the normative power of the utility come from? Theoretically, as famously established axiomatically by Von Neumann and Morgenstern (VNM) in 1947, our preferences are rational—i.e. obey VNM's rationality axioms—if and only if there is a utility function that agrees with those preferences. This means that once we settle on a utility function, its implied preferences are certified 'Rational', so 'being rational' can be achieved by following the utility function. Who likes to be accused of being 'ìrrational'? But we also know that a person may behave rationally without ever being aware of his utility. From his perspective, it is not that he prefers A over B because $u(\text{ A }) > u(\text{ B })$, but the other way around: $u(\text{ A }) > u(\text{ B })$ because he prefers A over B. This means the preferences come first, then the utility. This is what we have done by first listing all the preferences in P1 and P2, then deriving the specific utility function that satisfies them. The issue of primacy between preference and utility is also discussed in Easwaran (2014).

But surely the value of having a utility function is to derive implied preferences numerically from it. How do we explain the potential mismatch between the iron law of VNM rationality with our natural preferences? To suggest an answer from the lesson of Zeckhauser's paradox, it is instructive to first put it in Savage's formal decision theory framework by explicitly specifying the acts, the states of the world and the consequences. See Table 1.

The base preference 'Alive is better than Dead', implying $u(\text{ Alive }) = 1$ and $u(\text{ Dead }) = 0$, and the probability 1/6 for each state of the world, lead to uncontroversial utilities $u(C_k)$'s in (2). But *paying to reduce the number of bullets is*

Table 1 Decision table for the basic Russian Roulette

	Chamber number					
	1	2	3	4	5	6
C_0	Alive	Alive	Alive	Alive	Alive	Alive
C_1	Dead	Alive	Alive	Alive	Alive	Alive
C_2	Dead	Dead	Alive	Alive	Alive	Alive
C_3	Dead	Dead	Dead	Alive	Alive	Alive
C_4	Dead	Dead	Dead	Dead	Alive	Alive
C_5	Dead	Dead	Dead	Dead	Dead	Alive

C_k is the 'act' (lottery) that corresponds to having a gun with k bullets. The true state of the world is the identity of the fired chamber. The consequences are 'Dead' or 'Alive'

Table 2 Part of the decision table for an extended Russian Roulette that allows new acts of paying to reduce the number of bullets. C_0^*, C_3^*, A^*x_0 and A^*y_3 are defined in Sect. 2

	Chamber number					
	1	2	3	4	5	6
C_0^*	A^*x_0	A^*x_0	A^*x_0	A^*x_0	A^*x_0	A^*x_0
C_3^*	Dead	Dead	Dead	A^*y_3	A^*y_3	A^*y_3

a new act not represented in the basic Table 1. Part of an extended Russian Roulette to cover this new act and its consequences is given in Table 2. This formal decision table shows explicitly the different consequences (thus corresponding utilities) of different acts, so we are fully within the classical decision/utility theory.

How does the paradox appear? Savage (1954), as highlighted recently by Binmore (2009, 2017), was careful to emphasize the idea of 'Small World', a timeless self-contained world equipped once and for all with a complete set of acts, states of the world and consequences. This is a crucial concept, because the normative force of the theory applies only in this Small World. Defining a new act means going out of the small world. The old utilities may extend to the larger world, and by the VNM Theorem, the implied preferences *will* satisfy VNM's rationality axioms. But such set of preferences is just *one* of infinitely many rational sets in the larger world, each with its own utility function. There is no guarantee at all that the implied preferences would agree with your unstated—and unconsulted—preferences involving the new acts and consequences. Thus, the paradox appears when there is a mismatch. But your own set of preferences can still be rational; it just needs to correspond to another utility function.

An adopted Small World is always a simplification of a larger world. Being 'Alive' is a sufficient consequence in the Small World of basic roulette (Table 1), but it needs to be extended to cover the new act of buying out the bullets. As shown in Table 2, the standard calculation—which leads to the paradox—has to define new consequences A^*x_0 and A^*y_3. In Sect. 3, partly shown in Table 3, we have made explicit (i) all the new acts C_{ji}^* and consequences A^*x_{ji} that could appear from buying out bullets, and (ii) the preferences they should satisfy. In effect we have

Table 3 Part of the decision table for an extended Russian Roulette that allows new acts of paying to reduce the number of bullets, while keeping track of the original and final number of bullets

	Chamber number					
	1	2	3	4	5	6
C^*_{20}	A^*x_{20}	A^*x_{20}	A^*x_{20}	A^*x_{20}	A^*x_{20}	A^*x_{20}
C^*_{43}	Dead	Dead	Dead	A^*x_{43}	A^*x_{43}	A^*x_{43}

constructed a larger self-contained Small World. However, we can easily come up with other new acts—for example, buying and selling, auctioning or futures trading, etc., of bullets between multiple prisoners—that may yet generate other paradoxes if we again just allow the utilities from this new Small World to govern the larger world. Zeckhauser's paradox shows how easy it is for the old utility to seep out to the larger world and produce surprising preferences.

Our analysis also shows that Zeckhauser's paradox depends on a strong but unstated preference assumption that being alive is an absolute state with a static utility value regardless of any history or events. In contrast, the paradox does not occur if we describe the state of being alive more finely, as given in the preferences P1 and P2, to reflect the complexity of the situation. Interestingly, these preferences implicitly allow a dynamic utility, giving more value to life after a more serious threat. This looks rather like the state-dependent utility, but as stated above, we are within the framework of classical utility theory. This can also be seen by taking the life-and-death situation out of the story, by turning it into a game-show with prize money. In this version, the paradox arises because of the confusion between the use of prize money or own money to pay for the extra winning box.

Stefánsson and Bradley (2015) criticized the chance-neutrality of the classical utility and introduced the concept of utility of chance. To avoid Zeckhauser's paradox, Stefánsson (2017) used intuitive arguments to arrive at the following chance-dependent utilities . For $i = 1, \ldots, 6$, the utility of i / 6 chance of being alive is

$u(6/6)$	$u(5/6)$	$u(4/6)$	$u(3/6)$	$u(2/6)$	$u(1/6)$
1	0.67	0.48	0.41	0.34	0.24

These utilities correspond to ours in (11) above. He then assumed that the utility of the chance of being alive with a debt has the same shape as the function described by the table above, and showed a result that avoids the original paradox. But, unlike our derivation, it is not clear what preferences these chance-dependent utilities correspond to. This means these utilities might introduce another Zeckhauser-like paradox.

Using the same steps as in Stefánsson (2017), and assuming that by 'the same shape' means 'proportional', we can compute the following. The utility of being alive with debt incurred from reducing the number of bullets from 5 to 4 is $u_{54} = 0.24/0.34 = 0.706$. While the utility being alive with debt incurred from reducing the number of bullets from 2 to 1 is $u_{21} = 0.48/0.67 = 0.716$. This means

that we should pay more for reducing the number of bullets from 5 to 4 than from 2 to 1. Other similar situations can be constructed, e.g. paying more for reducing from 4 to 3 bullets than from 3 to 2 bullets. Strictly, these are not Zeckhauser's paradoxes, but they are Zeckhauser-like in the sense that one may not agree with these implied preferences, even though they come from seemingly sensible utilities. We have avoided this problem by stating all the preferences explicitly; furthermore, we have shown that, at least in this problem, we do not need any non-standard concept of utility as long as we are careful with the details of the small world we settle in.

Some commentaries on the paradox question the relevance of utility after you are dead. This is somewhat related to the ergodicity argument from Taleb (2018): removing all the bullets in (A) guarantees your survival, so it is always more rational than removing one from (B), which does not guarantee survival. We briefly discuss his solution here.

A time series X_t is said to be *ergodic* if the time average computed from following one observed time series converges to the ensemble average computed from many realizations of the time series at *one single time point*. Mathematically, on observing the series $X_1, \ldots, X_T$, we have

$$\frac{1}{T}\sum_{i=1}^{T} X_i \to EX_0,$$

as $T \to \infty$, where the ensemble-based expected value EX_0 is taken at the origin. In practice, we can say ergodicity to hold if the convergence occurs within our time horizon. To apply the reasoning in the Russian Roulette story, you imagine there were a large number of people put under the gun, and you are one of those unlucky ensemble. So, the previous utility calculations leading to (4) apply to this ensemble.

However, any time series with an absorbing state—such as death or bankruptcy—is not ergodic. If the terrorists put you on the Russian Roulette repeatedly, the time average of the utility of any choice other than removing all bullets from the gun is zero in the long run. Taleb (2018)'s ergodicity argument states that when you are certain to get to the absorbing state, relying on the ensemble average of the utility for your personal decision is not rational. So it is rational to pay any amount to survive, and this justifies our normal human reaction. But what if the Roulette is a one-off event? There is no legal, moral or mathematical obligation for you to pay attention to the ensemble- nor the time-average argument. But in von Neumann and Morgentern's and Savage's frameworks, you can interpret the utilities subjectively for one-off decisions. So it is still useful to resolve the paradox within the classical utility theory.

Acknowledgements Open access funding provided by Karolinska Institute.

References

Binmore, K. (2009). *Rational decisions*. Princeton: Princeton University Press.
Binmore, K. (2017). On the foundations of decision theory. *Homo Oeconomicus*, *34*(4), 259–273.
Bouyssou, D., & Vincke, P. (2010). *Binary relations and preference modeling. Decision-making process: Concepts and methods* (pp. 49–84). New York: Wiley.
Chu, Y. J. (1965). On the shortest arborescence of a directed graph. *Scientia Sinica*, *14*, 1396–1400.
Easwaran, K. (2014). Decision theory without representation theorems. *Philosophers' Imprint*, *14*(27), 1–30.
Jeffrey, R. (1988). Biting the Bayesian bullet: Zeckhauser's problem. *Theory and Decision*, *25*(2), 117–122.
Landsburg, S. (2011). Another rationality test. http://www.thebigquestions.com.
Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.
Stefánsson, H. O., & Bradley, R. (2015). How valuable are chances? *Philosophy of Science*, *82*(4), 602–625.
Stefánsson, H. O. (2017). Gambling with death. *Topoi*,. https://doi.org/10.1007/s11245-017-9519-z.
Taleb, N. N. (2018). *Skin in the game: Hidden asymmetries in daily life*. New York: Penguin Random House.
Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, *1*(2), 146–160.
Von Neumann, J., & Morgenstern, O. (1947). *The theory of games and economic behavior* (2nd ed.). Princeton: Princeton University Press.