An inequality for experimentally testing the violation of Savage's sure-thing principle

Taiki Takahashi[a]

[a]*Hokkaido University, Department of Behavioral Science, Sapporo, Japan*
**e-mail: ttakahashi@let.hokudai.ac.jp*

(21 Oct 2025 version)

Abstract—Recent advances in behavioral economics and quantum cognition and decision elucidated a number of deviations of actual human decisions and choices from mathematical principles of normative decision theory, which are referred to as "anomalies". One of the prominent anomalies is that the violations of Savage's sure-thing principle, which is the fundamental axiom of the rational theory of decision under uncertainty. It states that if prospect x is preferred to y knowing that Event A occurred, and if x is preferred to y knowing that A did not occur, then x should be preferred to y even when itis not known whether A occurred. I explicitly derive an equality for testing the violations of Savage's principle in behavioral experiments on decision under uncertainty. Future applications for behavioral and neuro- economics and quantum cognition and decision theory are discussed.

Keywords: behavioral economics, risk, uncertainty, Savage's sure-thing principle

1. INTRODUCTION

Behavioral economic studies have explored various "anomalies" in human judgment and decision making. For instance, people's decision under risk and uncertainty often violates von Neuman and Morgenstern (1944)'s expected utility theory based on the independence axiom (see Allais, 1953, for an early examination of this axiom), and people usually discount future outcomes (gains and losses) in a hyperbolic manners (Thaler, 1981) rather than rational (dynamically consistent), exponential manner. As a result of extensive studies in behavioral economics, Daniel Kahneman, who mainly studied judgement and decision-making under uncertainty, and Richard Thaler, who are well-known for studies of self-control problems in intertemporal choice and nudge, a new type of behavioral interventions, won Nobel prize for economics. Kahneman's colleague Amos Tversky has also performed great contributions to behavioral economics in studies of intransitive preferences, decision under risk and uncertainty, and the violations of Savage's sure-thing principle which is explained in the next part.

1.1 Behavioral studies of Savage's sure-thing principle

Notably, among fundamental axioms in rational decision theory, Savage's sure-thing principle has also been experimentally examined. The sure-thing principle was originally introduced by L.T. Savage (1954) using the following story:

A businessman contemplates buying a certain piece of property. He

> considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew that the Democratic candidate were going to win, and decides that he would. Similarly, he considers whether he would buy if he knew that the Republican candidate were going to win, and again finds that he would. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say. [p. 21]

As is demonstrated by this story, the sure-thing principle appears innocent and compelling. Savage legitimately recognized its general validity and noted: I know of no other extralogical principle governing decisions that finds such ready acceptance.

Slovic and Tversky(1974)'s study is one of the important early attempts to test this principle. They characterized Savage's sure-thing principle in the following manner:

"Savage' (1954) independence principle (SIP), also called the sure-thing principle. This axiom asserts that if two alternatives have a common outcome under a particular state of nature, then the ordering of the alternatives should be independent of the value of that common outcome. Objections to SIP, in the form of counter- examples have been proposed by Allais (1955) and Ellsberg (1961)".

In Slovic and Tversky's formulation, the distinct characteristics of Savage's sure-thing principle from von Neumann and Morgenstern's independence axiom:

> If a person prefers option A over option B, they should also prefer a lottery that gives A with probability p and some third option C with probability $(1-p)$ over a lottery that gives B with probability p and C with probability $(1-p)$.
> In other words, preferences between options should remain consistent when they are part of probabilistic mixtures with a common third option.

is somewhat obscure. At a later stage in the emergent of behavioral economics, Tversky and Shafir (1992) stated the principle in this manner:

"One of the basic axioms of the rational theory of decision under uncertainty is Savage's (1954) sure-thing principle (STP). It states that if prospect x is preferred to y knowing that Event A occurred, and if x is preferred to y knowing that A did not occur, then x should be preferred to y even when it is not known whether A occurred "

This formulation is distinctive in that it explicitly incorporates an assumption concerning the decision maker's knowledge about the state of the world. However, in their studies (Shafir and Tversky (1992) and Tversky and Shafir (1993) on the Prisoner's Dilemma, the Newcomb's problem, Hawaii problem etc., they examined whether the following proposition, rather than Savage's sure-thing *per se,* holds:

When the three probabilities are compared — namely, the probability that prospect x is preferred to y given that Event A occurred, the probability that x is preferred to y given that A did not occur, and the probability that x is preferred to y when it is not known whether A occurred — the latter tends to take an intermediate value between the two conditional probabilities.

It is here reasonable to question why this condition on three types of the probabilities (i.e., the two conditional probabilities and the unconditional probability) is equivalent to the Savage's sure-thing principle, which incorporates an assumption concerning the decision maker's knowledge about the state of the world, but not probabilistic concepts. This issue has been extensively studied in subsequent mathematical modelling studies by several physicists, psychologists, and behavioral economists who have started to utilize non-standard probability theories (e.g., quantum probability theory, Pothos and Busemeyer, 2009) to model human violations of Savage's sure-thing principle.

1.2 Relations between Savage's sure-thing principle and the law of total probability

According to the law of total probability, if Events A and B form a partition of the sample space, the overall probability of Y can be expressed as a weighted sum of its conditional probabilities:

$$P(Y) = P(A)P(Y|A) + P(B)P(Y|B).$$

This law states that the total probability of Y, the unconditional probability $P(Y)$, which is equivalent to $P(Y|A\ or\ B) = P(Y|unknown)$, is obtained by summing the conditional probabilities of Y occurring under each possible condition, weighted by the unconditional probability of each condition. When we set $P(Y|A) = P(Y|B) = 1$, the law of total probability reduces to $P(Y) = 1$, by using that $P(A) + P(B) = 1$. Here we must note that $P(Y|A) = P(Y|B) = 1$ corresponds to " if prospect X is preferred to Y knowing that Event A occurred, and if x is preferred to y knowing that A did not occur (Event B occurred)" and $P(Y) = 1$ corresponds to " x is preferred to y even when it is not known whether A occurred". Therefore, it can be stated that Savage's sure-thing principle in Tversky and Shafir' (1992)s formulation is a special case of the law of total probability when $P(Y|A) = P(Y|B) = 1$. Tversky and Shafir 1992) and Shafir and Tversky (1992) exploited the law of total probability instead of Savage's sure-thing principle *per se*. Although these Tversky and Shafir's studies and subsequent studies in quantum models of cognition and decision (Khrennikov, 2010) have utilized the law of total probability, no studies to date explicitly address the inequality on behavioral-experimentally measurable values (i.e., $P(Y),\ P(Y|A), P(Y|B)$) which holds if and only if the Savage's sure-thing principle holds. This point important for experimental psychologists and economist to test whether Savage's sure-thing principle is violated under experimental conditions in laboratory settings.

2. MATHEMATICAL ANALYSIS OF VIOLATIONS OF SAVAGE'S SURE-THING PRINCIPLE

As stated earlier, Savage's sure-thing principle can be regarded as a special case for the law of total probability. By exploiting this, we can claim that the violations of Savage's sure-thing principle is, under behavioral experimental conditions that participant has two mutually exclusive responses (Y or N) and s/he is allocated to two mutually distinct known experimental conditions (A or B) and an unknown condition, equivalent to the violations of the following inequality (see a proof below) with standard definition of conditional probabilities of which values are obtained from experimental data (provided $P(Y|A) \neq P(Y|B)$) :

$$0 < \frac{P(Y) - P(Y|B)}{P(Y|A) - P(Y|B)} < 1. \quad (1)$$

When this inequality is not satisfied, i.e., one of the following two inequalities hold, we can say that Savage's sure thing principle is violated, which suggest that the participants are irrational in the sense of normative principles of consequential rationality and/or standard probability theory:

$$P(Y|B) < P(Y) < P(Y|C), \quad (2)$$

or alternatively,

$$P(Y|A) < P(Y) < P(Y|D). \quad (3)$$

It is to be noted that inequality 1 can only be utilized when $P(Y|A) \neq P(Y|B)$, in other words, the probability that participant's response is Y under condition A is different from that under condition B.

I here denote the proof of inequality 1. The law of total probability states:

$$P(Y) = P(A)P(Y|A) + P(B)P(Y|B). \quad (4)$$

Because condition A and B are mutually exclusive, it is satisfied from the axiom of Kolmogorovian probability theory:

$$P(B) = 1 - P(A). \quad (5)$$

If we solve equation 4 in terms of $P(A)$, by utilizing equation 5, we obtain:

$$P(A) = \frac{P(Y) - P(Y|B)}{P(Y|A) - P(Y|B)}. \quad (6)$$

Now from $0 < P(A) < 1$, we obtain equation 1. (QED)

5. DISCUSSIONS

The present analysis indicates that after determining(measuring) three probabilities (two conditional probabilities of participants' behavioral choices $P(Y|A)$, $P(Y|B)$ and one unconditional probability $P(Y)$: the probability of participants' behavioral choices under the unknown condition) , we are able to distinguish between violations and non-violation of Savage sure-thing principle by checking whether the inequality 1 holds (non-violation) or not (violation). As far as I know, this is the first study to derive the inequality on experimentally measurable quantities to test the violations of Savage's sure thing principle.

5.1 Relationship to previous studies

Here discussed the relationship of the present study to the previous ones examining the violations of Savage's sure-thing principle and human probability judgement errors.

After Tversky and Shafir's studies, several studies in quantum cognition and decision models (see Khrennikov 2010, for the rich applications of this approach) utilized the formula of probabilities in quantum theory, in which so-called "interference term" I appears, as is the usual case in quantum theory in physics, in the right-hand side of the law of total probability:

$$P(Y) = P(A)P(Y|A) + P(B)P(Y|B) + I \quad (7)$$

which can capture the violation of Savage's sure-thing principle in a quantitative manner. Originally, for modelling probability judgement errors in humans, Franco quantified, by utilizing quantum settings, the interference term as

$$I = 2\sqrt{P(A)P(B)P(Y|A)P(Y|B)}\cos\,\theta,$$

where θ is a parameter called "quantum phase" in mathematical formalism in quantum physics. This expression has been widely utilized to model the violation of Savage's sure-thing principle and other human probability judgement errors, and extended in later studies. However, in Tversky and Shafir's studies (and most subsequent ones), $P(A)$ and $P(B)$ were not measured and only $P(Y), P(Y|A), P(Y|B)$ are experimentally measured. This makes it difficult to distinguish between violation and non-violation of the Savage's sure-thing principle from the experimental data in most quantum theoretical model studies. In contrast, the present approach based on the inequality only containing experimentally observable quantities $(P(Y), P(Y|B), P(Y|A))$ which is an advantage when researchers are only interested in whether Savage's axiom is violated or not.

The present approach also indicates a psychological origin of the Savage's sure-thing principle, other than that proposed in Tversky and Shafir (and quantum models of cognition and decision). Shafir and Tversky (1992) attributes the violation of the sure-thing principle (STP, in the citation below) to people's "non-consequential" reasoning (i.e., lacking a clear reason for choice under unknown conditions):

"people do not always choose in a consequentialist manner, then STP may sometimes be violated. For example, we have shown else- where that many people who chose to purchase a vacation to Hawaii if they were to pass an exam and if they were to fail, decided to postpone buying the vacation in the disjunctive case, when the exam's outcome was not known (Tversky & Shafir, 1992). Having passed the exam, the vacation is presumably seen as a time of celebration following a success- fail semester; having failed the exam, the vacation becomes a consolation and time of recovery. Not knowing the outcome of the exam, we suggest, the decision maker lacks a clear reason for going and, as a result, may prefer to wait and learn the outcome before deciding to go, contrary to STP."

This interpretation of the violation of Savage's sure-thing principle attributes "irrationality" in decision making to the wrong value of $P(Y)$: probabilities of choice under unknown conditions, rather than wrong values of $P(Y|A)$ and $P(Y|B)$. In quantum theorical modelling

studies, additional "interference" term I alone is problematic, that is, the values of $P(A), P(Y|A), P(B), P(Y|B)$ are legitimate, resulting in "irrationality" of the left-handed side of equation 1, P(Y). It can together be said that both conventional (Tversky and Shafir) and quantum models focus on the origin of psychological deviation from correct values of choice probabilities $P(Y)$ under unknown conditions. In contrast, the present analysis indicates another interpretation that equation 5, the additivity of probabilities, the sum of the probabilities of A and B (not A) is equal to one, has a problem in human mind, since in deriving the inequality of the Savage's sure-thing principle, equation 5 is utilized. There are several studies demonstrating non-additivity of probabilities in human mind. For instance, in Kahneman=Tversky's prospect theory (1979), the psychological probabilities in decision under risk ("decision weights", also referred to as probability weighting) is not additive: $w(p) + w(1-p) < 1$ (subadditivity), as shown by Allais(1953)' classical paradox.

5.2 Possible future directions

As suggested above, the violation of Savage's sure-thing principle arises from the violation of complementarity rule in probability theory: $P(A) + P(not\ A) = 1$ where $P(A)$ is the subjective belief that Event A occurs. This violation of the probability rule has been extensively studied in behavioral economics, particularly in Kahneman and Tversky's prospect theory as sub-additivity of probability weighting functions. Therefore, it may be a promising future directions to examine whether the violations of Savage's sure-thing principle is related to nonlinearity in decision-maker's probability weighting functions. Behavioral economist Drazen Prelec axiomatically derived the probability weighting function: $w(p) = \exp(-(-\ln p)^a)$, $0 < a < 1$. from, what is known as compound invariance. Al-Nowaihi and Dhami (2006), who derived the probability weighting function from "power invariance", states the importance of the axiom of compound invariance:

"The importance of this axiom is as follows. In expected utility theory, the product rule for probabilities allows us to reduce a compound lottery to a simple lottery of the same expected utility. Once we depart from expected utility theory, we need a rule that plays an analogous role. Compound invariance is a candidate for such a rule."

Furthermore, Takahashi (2011) derived the probability weighting function form psychophysical laws of time perception from which hyperbolic temporal discounting (Takahashi, 2005) and the characteristics of prospect theory's probability weighting (Takahashi and Han 2013) are derived. Hence, it is an interesting possibility that the violations of Savage's sure-thing principle originates from nonlinear distortion of time-perception. This should be examined in future studies in behavioral economics.

Several studies have attempted neural information processing underlying the violations of Savage's sure-thing principle. Takahashi and Cheon (2012) explored connection of quantum-like models to nonlinear neural population coding (bridge from cognitive model to neural mechanism). This study is a useful first attempt to investigate a neurally-plausible account of quantum probabilistic interference effects on decision through uncertainty resulting in the violation of Savage's sure-thing principle. Recently, Khrennikov and colleagues (2025) also attempts to embed quantum-like cognitive models inside neuronal/graphical network frameworks; explicitly targeted at reproducing disjunction and order effects (including the

violations of Savage's sure-thing principle) while connecting to neuronal architectures. Nevertheless, more work is needed for theoretical links between neural information processing and the violation of Savage's sure-thing principle.

In the field of neuroeconomics, Berns et al. (2007) was an early fMRI conceptual/empirical approach to estimating probability weighting from neural data. Tobler et al. (2008) conducted Single-unit work showing that neural signals represent probability in a non-linear fashion consistent with a weighting function (important neuronal evidence for distortions analogous to prospect theory weighting). Hsu et al. (2009) shows fMRI evidence that BOLD signals respond nonlinearly to objective probability; authors fit Prelec-style weighting forms to neural responses. This is central for linking the violations of Savage's sure-thing principle to distorted subjective probability representation. Ojala et al. (2018) Pharmacological manipulation (dopamine) changes estimated probability-weighting. This study elucidated neurotransmitter mechanisms underlying nonlinear probability weighting in decision under risk and uncertainty. Future studies should examine which types of neurotransmitters modulate both the violations of Savage's sure-thing principle and nonlinearity in probability weighting, to clarify common neuronal mechanisms underlying the violations of the complementarity probability law and Savage's sure-thing principle.

6. CONCLUSIONS

The present analysis suggests that by examining three experimentally observable probabilities—two conditional probabilities of participants' behavioral choices, $P(Y \mid A)$ and $P(Y \mid B)$, and one unconditional probability $P(Y)$ representing choices under the unknown condition—we can determine whether Savage's sure-thing principle is violated or not by testing whether inequality (1) holds (non-violation) or fails (violation). Theoretically, the present analysis shows a violation of the sure-thing principle may imply a corresponding breakdown of the rule of complementarity in probability theory. Future research employing neuroimaging and pharmacological approaches, together with theoretical investigations of the relationship between the sure-thing principle and neural information processing, will be important for advancing our understanding of this phenomenon.

FUNDING

This research was not funded.

CONFLICT OF INTEREST

The author declares that they have no conflicts of interest.

REFERENCES

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica, 21*(4), 503–546. https://doi.org/10.2307/1907921

Al-Nowaihi, A., & Dhami, S. (2006). A simple derivation of Prelec's probability weighting function. *Journal of Mathematical Psychology, 50*(6), 521–524. https://doi.org/10.1016/j.jmp.2006.03.003

Berns, G. S., Capra, C. M., Chappelow, J., Moore, S., & Noussair, C. (2008). Nonlinear neurobiological probability weighting functions for aversive outcomes. *NeuroImage, 39*(4), 2047–2057. https://doi.org/10.1016/j.neuroimage.2007.10.028

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics, 75*(4), 643–669. https://doi.org/10.2307/1884324

Hsu, M., Krajbich, I., Zhao, C., & Camerer, C. F. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *Journal of Neuroscience, 29*(7), 2231–2237. https://doi.org/10.1523/JNEUROSCI.5296-08.2009

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291. https://doi.org/10.2307/1914185

Khrennikov, A. (2010). *Ubiquitous quantum structure: From psychology to finance.* Springer. https://doi.org/10.1007/978-3-642-05101-2

Khrennikov, A., Ozawa, M., Benninger, F., & Shor, O. (2025). Coupling quantum-like cognition with the neuronal networks within generalized probability theory. [Manuscript in preparation].

Ojala, K. E., Janssen, L. K., Hashemi, M. M., Timmer, M. H. M., Geurts, D. E. M., ter Huurne, N. P., Cools, R., & Sescousse, G. (2018). Dopaminergic drug effects on probability weighting during risky decision making. *eNeuro, 5*(2), ENEURO.0330-18.2018. https://doi.org/10.1523/ENEURO.0330-18.2018

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of "rational" decision theory. *Proceedings of the Royal Society B: Biological Sciences, 276*(1665), 2171–2178. https://doi.org/10.1098/rspb.2009.0121

Prelec, D. (1998). The probability weighting function. *Econometrica, 66*(3), 497–528. https://doi.org/10.2307/2998573

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). Dover Publications. (Original work published 1954)

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology, 24*(4), 449–474. https://doi.org/10.1016/0010-0285(92)90015-T

Slovic, P., & Tversky, A. (1974). Who accepts Savage's axiom? *Behavioral Science, 19*(6), 368–373. https://doi.org/10.1002/bs.3830190603

Takahashi, T. (2005). Loss of self-control in intertemporal choice may be attributable to logarithmic time-perception. *Medical Hypotheses, 65*(4), 691–693. https://doi.org/10.1016/j.mehy.2005.04.040

Takahashi, T. (2011). Psychophysics of the probability weighting function. *Physica A: Statistical Mechanics and Its Applications, 390*(5), 902–905. https://doi.org/10.1016/j.physa.2010.10.013

Takahashi, T., & Cheon, T. (2012). A nonlinear neural population coding theory of quantum cognition and decision making. *World Journal of Neuroscience, 2*(3), 183–186. https://doi.org/10.4236/wjns.2012.23025

Takahashi, T., & Han, R. (2013). Psychophysical neuroeconomics of decision making: Nonlinear time perception commonly explains anomalies in temporal and probability discounting. *Applied Mathematics, 4*(11), 1520–1525. https://doi.org/10.4236/am.2013.411207

Thaler, R. H. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters, 8*(3), 201–207. https://doi.org/10.1016/0165-1765(81)90067-7

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2008). Neuronal distortions of reward probability without choice. *Journal of Neuroscience, 28*(44), 11703–11711. https://doi.org/10.1523/JNEUROSCI.2870-08.2008

Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science, 3*(5), 305–309. https://doi.org/10.1111/j.1467-9280.1992.tb00678.x

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior* (3rd ed., 1953). Princeton University Press.