WORKING PAPERS

N° 1624

March 2025

"The disjunction effect does not violate the Law of Total Probability"

Alexandros Gelastopoulos & Gaël Le Mens

Toulouse School of Economics

THE DISJUNCTION EFFECT DOES NOT VIOLATE THE LAW OF TOTAL PROBABILITY

A PREPRINT

Alexandros Gelastopoulos
Department of Business and Management
University of Southern Denmark
alex@sam.sdu.dk

Gaël Le Mens
Department of Economics and Business
Pompeu Fabra University
gael.le-mens@upf.edu

July 16, 2024

ABSTRACT

The disjunction effect (DE) refers to an empirical violation of the Sure-Thing Principle (STP), which states that if a person is willing to take an action independently of the outcome of some event, then they must be willing to do so even when the outcome of the event is unknown. A standard practice for inferring a DE, especially in between-subjects experiments, consists of showing a population-level version of this phenomenon, specifically that fewer people are willing to take the proposed action when the outcome of the event is unknown than for any possible known outcome. Although this does not prove a violation of the STP, this population-level condition has received a lot of attention, because it presumably violates the Law of Total Probability, and it is sometimes used as the definition of the DE itself. Here we show that this condition is in fact unrelated to the Law of Total Probability and thus entirely irrelevant for the study of the DE and decision making in general. This calls for a reevaluation of experimental results that have been interpreted as showing a DE based on the above condition. We derive a new disjunction law that can be used to check for violations of the STP in between-subjects data.

Keywords disjunction effect · sure-thing principle · law of total probability · prisoner's dilemma · two-stage gamble

1 Introduction

The term disjunction effect was coined by Tversky and Shafir (1992) to describe an empirical pattern that appeared to violate the Sure-Thing Principle (STP). The latter states that if an individual is willing to take an action when an event A occurs *and* when A does not occur, then they must also be willing to take the same action when they do not know whether A has occurred or not (Savage, 1954).

A closer look at the relevant literature reveals that authors have actually been using two non-equivalent definitions of the term disjunction effect. According to the first, which is the most straightforward one, the disjunction effect is simply defined as *a violation of the STP* (e.g., Shafir and Tversky, 1992). Because the STP refers to the preferences of single individuals, this definition is also about single individuals: if an individual expresses the will to take an action in both possible outcomes of an event, but not when the outcome is unknown, then this individual exhibits the disjunction effect. We will call this the *Within-Subjects Disjunction Effect* (WSDE).

The second definition of the disjunction effect found in the literature (e.g., Broekaert et al., 2020) concerns preferences at the population level. It is defined to mean that the proportion of people in a population who are willing to take an action when it is unknown whether an event A has occurred or not is smaller than both the proportion of people who would take it if A occurred *and* the proportion of people who would take it if A didn't occur. This can be concisely written as

$$x_U < \min\{x_A, x_{A^c}\}, \tag{1}$$

where x_A, x_{A^c}, and x_U denote the proportions of people in the population who would take the action if A occurred, if A didn't occur, and if the outcome was unknown, respectively.[1] We will call this definition the *Population-Level Disjunction Effect* (PLDE).

The behavioral relevance of the WSDE is clear, as long as one accepts the relevance of the Sure-Thing Principle. It is less clear what makes the PLDE a phenomenon worth studying. Starting with Tversky and Shafir (1992), it was initially believed that the PLDE implied the WSDE,[2] but Lambdin and Burdsal (2007) showed that this is not the case (see also Broekaert et al., 2020). More recently, many authors have claimed that the PLDE implies a violation of the Law of Total Probability (Blutner and beim Graben, 2016; Moreira and Wichert, 2016b, 2018; Pisano and Sozzo, 2020; Tesař, 2020b; Broekaert et al., 2020; Maruyama, 2020; Aerts et al., 2021; Wang et al., 2022; Waddup et al., 2021; Pothos and Busemeyer, 2022; Shan, 2022; Widdows et al., 2023).[3] This fact has in turn served as motivation for the introduction of several models of cognition that depart from classical decision theory, especially quantum-probabilistic models for which the standard Law of Total Probability does not hold (see Pothos and Busemeyer, 2022, for a review).

Here we show that eq. (1) does *not* imply a violation of the Law of Total Probability. The reason for the above widely held misconception appears to be poor use of mathematical notation when identifying the proportions that appear in eq. (1) with probabilistic quantities. Specifically, assuming that we randomly sample a population, the probability of observing a certain characteristic is equal to the proportion of the people who have that characteristic. This has led authors to rewrite eq. (1) as

$$\mathbb{P}\left(B\right) < \min\left\{\mathbb{P}\left(B \mid A\right),\, \mathbb{P}\left(B \mid A^c\right)\right\}, \tag{2}$$

where B is the event that the (randomly sampled) individual is willing to take the action in question, and the conditional probabilities on the right hand side refer to the conditions that A has or hasn't occurred, respectively. As we show, however, the correct transcription of eq. (1) is

$$\mathbb{P}_u\left(B\right) < \min\left\{\mathbb{P}_r\left(B \mid A\right),\, \mathbb{P}_r\left(B \mid A^c\right)\right\}. \tag{3}$$

where $\mathbb{P}_u\left(\cdot\right)$ and $\mathbb{P}_r\left(\cdot\right)$ are two different probability distributions. We argue that eq. (3) is in no way related to the Law of Total Probability, whose violation cannot be possibly shown in these experiments.

Given that the PLDE does not contradict either probability-theoretic laws or the STP, its study has no theoretical grounds whatsoever; it does not constitute a violation of classical decision theory. This result has important consequences both for the empirical study of the disjunction effect and for theoretical models of cognition (see our Discussion).

Having established that the Population-Level Disjunction Effect is behaviorally irrelevant, in the second part of the paper we are concerned with the question of whether the *Within-Subjects* Disjunction Effect can be inferred from *between-subjects* data, a task that Lambdin and Burdsal (2007) have argued to be impossible. We instead show that sometimes a WSDE *can* be inferred from between-subjects data alone. We do so by deriving a new relation involving population-level quantities, which follows from the Sure-Thing Principle and states that

$$x_A + x_{A^c} - 1 \leq x_U \leq x_A + x_{A^c}. \tag{4}$$

This relation, which we call the *disjunction law*, can be checked in between-subjects data and, because it follows from the STP, its violation implies the Within-Subjects Disjunction Effect. Moreover, we show that in between-subjects studies, this relation is both *necessary and sufficient* for the data to be consistent with the STP. In other words, if the relation we identify holds, although the WSDE cannot be excluded, *it cannot be concluded from the data*. More generally, even if the data comes from a within-subjects study, as long as eq. (4) holds, then the WSDE cannot be shown in an analysis that doesn't take into account the links between answers of the same subject.

[1] One can alternatively use the relation $x_U > \max\{x_A, x_{A^c}\}$ in place of eq. (1). One form can be transformed into the other by considering the opposite action.

[2] The exact wording in Tversky and Shafir (1992) is "The data show that a majority of subjects accepted the second gamble after having won the first gamble, and a majority accepted the second gamble after having lost the first gamble. Most subjects, however, rejected the second gamble when the outcome of the first was not known. This pattern of preference clearly violates Savage's STP, we call it the disjunction effect".

[3] Early versions of this statement appear in Busemeyer et al. (2006), but without explicit mention of the Law of Total Probability, and in Pothos and Busemeyer (2009). Some researchers have also claimed that the STP itself follows from the laws of probability theory, or equivalently that the WSDE implies their violation (Khrennikov, 2010, 2015, 2022; Pothos et al., 2011; Moreira and Wichert, 2016b; Snow et al., 2024). The only alleged proofs of this claim that we are aware of appear in Pothos et al. (2011) and possibly in Khrennikov, 2010, p. 94, but they rely on the same argument as the corresponding claim for the PLDE, which here we show is incorrect. The example given in table 2 below also serves as a counterexample.

2 Background

The typical experimental setup for the disjunction effect is as follows. A situation is described to the participant, including an event that has two possible outcomes and over which the participant has no control. In one experimental condition, the outcome of that event is revealed to the participant, while in another it is not. The participant then is given a choice between two actions. Starting with Tversky and Shafir (1992) and Shafir and Tversky (1992), the two most common scenarios used in these experiments are variations of the following two types.

- **Prisoner's Dilemma** (Shafir and Tversky, 1992; Li and Taplin, 2002; Busemeyer et al., 2006; Li et al., 2010; Hristova and Grinberg, 2010; Tesař, 2020b; Waddup et al., 2021): Participants are paired with human or artificial agents and play the well-known Prisoner's Dilemma game. The available options for the participant are to cooperate (C) or defect (D). In one condition, the participant is told what their opponent has played, while in the other they are not.
- **Two-stage gamble** (Tversky and Shafir, 1992; Kühberger et al., 2001; Bagassi and Macchi, 2006; Lambdin and Burdsal, 2007; Li et al., 2012; Broekaert et al., 2020; Ziano et al., 2021): Participants are told that they have just taken a bet in which they had 50% chance of winning 200$ and 50% chance of losing 100$. In one condition, the outcome of the bet is revealed to them, while in the other it is not. They are then asked whether they would take the same bet once more.

The Sure-Thing Principle asserts that if the participant chooses the same action for both possible outcomes of the event when this outcome is known, then they must choose the same action also in the unknown-outcome variant. Violations of this principle can be easily inferred from *within-subjects* experiments, by comparing the answers of the same participant to all three variants of the question (known outcome 1, known outcome 2, or unknown outcome). Such a conclusion does not require checking any population-level quantities, and the current paper does not challenge this approach.

However, researchers have also tried to infer a disjunction effect from aggregate data, either from within-subjects or from between-subjects studies (in a between-subjects design each participant answers only one variant of the question). The idea, which goes back to Tversky and Shafir (1992), is the following: if the proportion of people who choose an action in the unknown-outcome variant is lower than both proportions of people who choose this action in the first and second known-outcome variant, then this implies a disjunction effect. Although Lambdin and Burdsal (2007) showed that this is not necessarily the case, the above analysis based on proportions has continued to be used widely, presumably because it can show a violation of the Law of Total Probability and consequently of classical decision theory. Below we are going to argue that this is not at all the case, but let us first review the argument, repeated in numerous studies on the disjunction effect (Busemeyer et al., 2006; Blutner and beim Graben, 2016; Broekaert et al., 2020; Pisano and Sozzo, 2020; Tesař, 2020b; Aerts et al., 2021; Pothos and Busemeyer, 2022; Shan, 2022; Widdows et al., 2023; Wang et al., 2022; Mahalli and Pusuluk, 2024). We are going to use the terminology of the two-stage gamble, but the argument is completely analogous for the Prisoner's Dilemma and other experimental scenarios.

We denote by W and L the events of winning, respectively losing, the first bet. We also denote by G, standing for "gamble", the event that the participant takes the second bet. The law of total probability states that

$$\mathbb{P}\left(G\right) = \mathbb{P}\left(W\right) \cdot \mathbb{P}\left(G \mid W\right) + \mathbb{P}\left(L\right) \cdot \mathbb{P}\left(G \mid L\right). \tag{5}$$

Because $\mathbb{P}\left(W\right)$ and $\mathbb{P}\left(L\right)$ are both positive and add up to 1,[4] the right hand side is a weighted average of the quantities $\mathbb{P}\left(G \mid W\right)$ and $\mathbb{P}\left(G \mid L\right)$. As a result, $\mathbb{P}\left(G\right)$, which is equal to this weighted average, must be between those two values. Therefore, if we find empirically that $\mathbb{P}\left(G\right)$ is smaller than both quantities, in other words

$$\mathbb{P}\left(G\right) < \min\left\{\mathbb{P}\left(G \mid W\right), \mathbb{P}\left(G \mid L\right)\right\}, \tag{6}$$

then eq. (5) cannot possibly hold.

So far there is nothing wrong with the argument. The problem arises in the next step, when one relates eq. (6) with eq. (1), which in this context can be written in the more suggestive notation

$$x_U < \min\left\{x_W, x_L\right\}. \tag{7}$$

To go from eq. (6) to eq. (7), the quantity $\mathbb{P}\left(G\right)$ is substituted by the proportion of people in the population who would be willing to take the second gamble when the outcome of the first gamble is unknown. Similarly, the quantities $\mathbb{P}\left(G \mid W\right)$ and $\mathbb{P}\left(G \mid L\right)$ are substituted by the proportions of people who are willing to take the second bet when the outcome of the first is revealed to be W or L, respectively. Thus, by showing empirically eq. (7) (taking into account possible statistical error), we presumably derive a violation of the Law of Total Probability.

In the next section we show why the identification of eq. (6) with eq. (7) is problematic.

[4] In the standard two-stage gamble example, $\mathbb{P}\left(W\right)$ and $\mathbb{P}\left(L\right)$ are both equal to 0.5, but in general they don't have to be.

	W	L	U
Type 1 individuals	1	0	0
Type 2 individuals	0	1	0
Entire population	0.5	0.5	0

Table 1: We suppose that half of the population is type 1 individuals, who are willing to take the second gamble if they know they won the first one, but not if they lost it or if they don't know. The other half is type 2 individuals, who are willing to take the second gamble if they know they lost the first one, but not if they won it or if they don't know. The last row shows the frequency of the corresponding behavior in the entire population.

3 Main argument

3.1 Conditioning vs revealing the outcome

We will continue using the two-stage gamble to make our case, but the same argument applies to other types of experiments.

We use the following example for a concrete demonstration of our argument. Suppose that among all individuals in the population, half of them would take the second bet if they had won the first one (W), but they wouldn't take it if they had lost the first bet (L) or if they didn't know the outcome (U). The other half of the participants would take the second bet only if they had lost the first bet (L), but not if they had won it (W) or if they didn't know the outcome (U). See table 1. Such preferences are entirely inline with a classical model of decision making (e.g., expected utility theory) and, in particular, do not violate the Sure-Thing Principle (see also Lambdin and Burdsal, 2007).

As table 1 shows, in the entire population, the proportions of people who would take the second bet in each of the three possible variants are $x_W = 0.5$, $x_L = 0.5$, and $x_U = 0$. We thus conclude that in this example we have a strong PLDE, because x_U is much lower than both x_W and x_L (see eq. (7)). According to the commonly accepted view in the literature, this translates into eq. (6) and thus violates the Law of Total Probability. But how can such a natural example, well within the context of classical decision theory, violate such a fundamental relation as the Law of Total Probability? We will now argue that it doesn't.

The main idea is that the act of telling the participant the outcome of the first experiment does not correspond to *conditioning* on the event W or L, but to *altering* the probability distribution itself. To see this, note that the event W, which appears for example in the expression $\mathbb{P}(W)$, is the event that the first bet is won, which is not the same as the first bet being won *and* telling the participant. After all, the latter fact cannot be captured by the probabilistic model, since it is an external manipulation that we make. Consistent with this interpretation of W, $\mathbb{P}(G \mid W)$ is the probability that the participant takes the second bet, given that the first bet is won. Again, there is no reference to whether we reveal this outcome to the participant. But unlike $\mathbb{P}(W)$, the value of $\mathbb{P}(G \mid W)$ depends on the external manipulation we employ: if we do not reveal the outcome, then the events G and W are independent, therefore $\mathbb{P}(G \mid W) = \mathbb{P}(G) = 0$ (bottom right cell of table 1); but if we do reveal the outcome, then $\mathbb{P}(G \mid W) = 0.5$ (last row, W column of table 1).

What is then the correct value of $\mathbb{P}(G \mid W)$? The answer is that they are both correct values, but for different probability distributions. By changing the conditions of the experiment (revealed vs unrevealed outcome), we are changing the underlying probability distribution. In our calculations, we must always specify to which of the two distributions we are referring.

We will distinguish the two probability distributions with the subscripts r and u, standing for *revealed* and *unrevealed* outcome. The previous paragraph shows that $\mathbb{P}_u(G \mid W) = 0$, while $\mathbb{P}_r(G \mid W) = 0.5$. The probability of all other events can also be considered under both contexts, for example $\mathbb{P}_u(G)$ is the probability that the participant takes the second bet when they don't know the outcome of the first bet, while $\mathbb{P}_r(G)$ is the probability of taking the second bet if they are told the outcome of the first bet (but without conditioning on what that outcome is).

3.2 Non-violation of the Law of Total Probability

Based on these observations, let us now revisit eq. (7) and express the quantities that appear there in terms of probabilities more precisely.

The quantity x_U refers to the proportion of people who would take the second bet in the unknown outcome condition, which (assuming we are sampling randomly) is equal to $\mathbb{P}_u(G)$. On the other hand, the quantities x_W and x_L refer to the proportion of people who would take the second bet in the known outcome condition, when the outcome is W or L, respectively, that is, $\mathbb{P}_r(G \mid W)$ and $\mathbb{P}_r(G \mid L)$. Therefore, eq. (7) can be rewritten as

$$\mathbb{P}_u(G) < \min\{\mathbb{P}_r(G \mid W), \mathbb{P}_r(G \mid L)\}. \tag{8}$$

Does eq. (8) imply a violation of the Law of Total Probability? The answer is no. Our argument consists of two parts. First, we show that the strictly mathematical interpretation of the Law of Total Probability is not violated when eq. (8) holds. We then argue that an alternative relation, which appears to be related to the Law of Total Probability and is inconsistent with eq. (8), has no theoretical grounds.

To argue for the first part, note that the axioms of probability theory, from where the Law of Total Probability follows, only deal with single probability distributions, thus they lead to no predictions about how the probability distributions $\mathbb{P}_u(\cdot)$ and $\mathbb{P}_r(\cdot)$ relate to each other. In particular, the Law of Total Probability can be written for either of the two probability distributions, taking the form

$$\mathbb{P}_u(G) = \mathbb{P}(W) \cdot \mathbb{P}_u(G \mid W) + \mathbb{P}(L) \cdot \mathbb{P}_u(G \mid L) \quad \text{or} \tag{9}$$

$$\mathbb{P}_r(G) = \mathbb{P}(W) \cdot \mathbb{P}_r(G \mid W) + \mathbb{P}(L) \cdot \mathbb{P}_r(G \mid L), \tag{10}$$

where in the case of $\mathbb{P}(W)$ and $\mathbb{P}(L)$ we have suppressed the subscripts, because these probabilities are the same in both cases.

Equation (8) places no restrictions between quantities that appear either both in the first or both in the second of these equations, so it cannot possibly imply a violation of either of those separately.[5] Could it prevent them from being true simultaneously? The answer is still no. To see this, note that eq. (8) places no restrictions on the values of $\mathbb{P}_u(G \mid W)$, $\mathbb{P}_u(G \mid L)$, and $\mathbb{P}_r(G)$. Assuming that all other variables are given, we may set $\mathbb{P}_u(G \mid W)$ and $\mathbb{P}_u(G \mid L)$ equal to $\mathbb{P}_u(G)$ and let $\mathbb{P}_r(G)$ be defined by eq. (10), to get a simultaneous solution for eqs. (9) and (10) (recall that $\mathbb{P}(W) + \mathbb{P}(L) = 1$). Thus, in a strict mathematical sense, eq. (8) does not imply a violation of the Law of Total Probability.

We now turn to the question of whether a different interpretation of the terminology can lead to a violation of *some* "law of total probability". The only reasonable candidate here seems to be the following relation:

$$\mathbb{P}_u(G) = \mathbb{P}(W) \cdot \mathbb{P}_r(G \mid W) + \mathbb{P}(L) \cdot \mathbb{P}_r(G \mid L). \tag{11}$$

Indeed, by an argument similar to the one given for eq. (6), it follows from eq. (11) that the value of $\mathbb{P}_u(G)$ must be *between* the values of $\mathbb{P}_r(G \mid W)$ and $\mathbb{P}_r(G \mid L)$. Thus, if eq. (8) holds, then eq. (11) would be violated.

Does eq. (11) deserve to be called a "law"? As we have seen, this relation does not follow from probability-theoretic considerations. We have also seen that it doesn't follow from the Sure-Thing Principle, demonstrated by the example at the beginning of the previous section, where we had $\mathbb{P}_r(G \mid W) = \mathbb{P}_r(G \mid L) = 0.5$ and $\mathbb{P}_u(G) = 0$, while the Sure-Thing Principle was satisfied. An alternative justification would be to argue that $\mathbb{P}_u(G)$ and $\mathbb{P}_r(G)$ should be equal, for then eq. (11) would follow from eq. (10). But for any given individual, the additional information provided in the revealed outcome condition may lead them to change their preference (e.g., to take the second bet), and this does not have to violate the Sure-Thing Principle (see table 1). And since individual participants may justifiably choose differently in the revealed vs the unrevealed condition, why should we require that the population proportion taking a given action be the same in the two conditions (i.e., $\mathbb{P}_u(G) = \mathbb{P}_r(G)$)?

We are not aware of any argument in the literature for this equality, neither for any alternative argument for eq. (11) to hold. And given that this relation is emphatically violated by such a natural example as the one given earlier (table 1), we conclude that eq. (11) can have no theoretical grounds. In light of this observation, eq. (8), which could only have been motivated by the fact that it violates eq. (11), appears entirely irrelevant to decision theory.

4 Inferring violations of the Sure-Thing Principle from between-subjects data

In studies of the disjunction effect, eq. (1) (the Population-Level Disjunction Effect) has been the standard method to argue for violations of classical decision theory. This has been especially the case for between-subjects studies, where there is no available data regarding preferences of the same individual to different experimental scenarios. But given the behavioral irrelevance of the PLDE that we showed in the previous section, a question that naturally arises is the following: is it possible to deduce a violation of the Sure-Thing Principle (i.e., a *Within-Subjects* Disjunction Effect) from between-subjects data? For if not, then between-subjects experiments would be deemed entirely unsuitable for studying the disjunction effect.

[5] In fact, because the Law of Total Probability is a theorem of probability theory which follows from its axioms, it cannot be violated without violating some of the axioms. To the best of our knowledge, no study has claimed a direct violation of any particular axiom of probability theory.

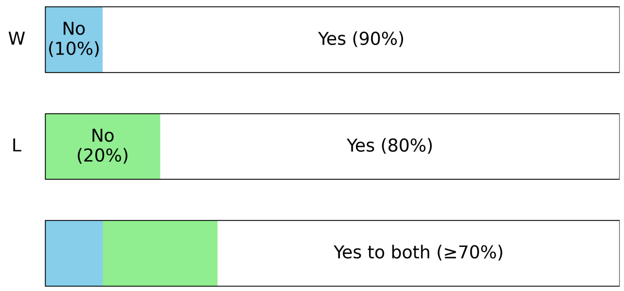

Figure 1: Assuming that 90% of the population would take the second bet in the W known outcome variant (top bar), and 80% would take it in the L known outcome variant (middle bar), then at least 70% of the people would take it in both cases (bottom bar). This last percentage can be higher, because the blue and green areas may overlap. The Sure-Thing Principle asserts that people in the white area of the bottom bar will also take the second bet in the unknown outcome variant (U).

Lambdin and Burdsal (2007) have argued that between-subjects studies are indeed not appropriate for showing violations of the Sure-Thing Principle. Here we instead show that violation of the STP *can* in fact, in some cases, be inferred from population-level quantities that are available in between-subjects. Specifically, we derive a relation that follows from the STP and set-theoretic arguments, thus its violation implies the Within-Subjects Disjunction Effect. We further show that this condition is also the strictest condition one can ask for; as long as it is satisfied, there is no way to infer a violation of the Sure-Thing Principle in a between-subjects study. This does not mean that the STP is necessarily satisfied by each individual in the population, but simply that it is theoretically impossible to infer the contrary with the available data.

4.1 The disjunction law: a population-level implication of the Sure-Thing Principle

Let us also start this section with an example. Suppose that we perform a between-subjects experiment and we find that, in the known outcome condition, the proportion of people who choose to take the second bet if they know the first bet is won is 90%, and the proportion of people who choose to take the second bet if they know the first bet is lost is 80%. We assume that our sample is very large, so that these percentages reflect the actual proportions in the population. Then, there are at most $10\% + 20\% = 30\%$ people who would *not* take the second bet under some known outcome variant (W or L), meaning that the remaining 70% (or more) would take it under both W and L (see fig. 1). The Sure-Thing Principle asserts that these individuals (comprising $\geq 70\%$) must be willing to take the second bet in the unknown outcome variant (U) as well. Therefore, if we found that fewer than 70% of the people are willing to take the bet in the unknown outcome variant, then the STP must be violated (by at least a subset of the population).

To generalize the above idea, let x_W, x_L, and x_U denote the proportions of people who are willing to take the second bet when the outcome of the first is W and revealed, L and revealed, or unrevealed, respectively. We also denote by x'_W, x'_L, and x'_U the complements of these proportions, i.e., $x'_W = 1 - x_W$ and similarly for L and U. The proportion of people *not* willing to take the bet under either W or L is *at most* $x'_W + x'_L$ (it can be smaller, if the corresponding sets of people overlap). Equivalently, the proportion of people who would take the bet in both cases is *at least*

$$1 - (x'_W + x'_L) = 1 - [(1 - x_W) + (1 - x_L)] = x_W + x_L - 1 \tag{12}$$

The Sure-Thing Principle asserts that all these people (possibly among others) must take the bet also in the unknown outcome variant, therefore

$$x_U \geq x_W + x_L - 1. \tag{13}$$

For the example given above, where $x_W = 0.9$ and $x_L = 0.8$, we get

$$x_U \geq 0.9 + 0.8 - 1 = 0.7, \tag{14}$$

as found earlier.

Applying the same logic to the action of *not* taking the second bet, assuming that the Sure-Thing Principle applies to it as well, we similarly find that $x'_U \geq x'_W + x'_L - 1$, and by substituting $x'_i = 1 - x_i$ for each of the three terms we get $x_U \leq x_W + x_L$.

	W	L	U
Type 1 individuals	1	1	0
Type 2 individuals	0	0	1
Entire population	0.5	0.5	0.5

Table 2: As in table 1, half of the population is assumed to be type 1 individuals and the other half type 2, with preferences as shown. Here both type 1 and type 2 individuals violate the STP. However, eq. (15) is satisfied, because $x_W = x_L = x_U = 0.5$ (last row).

The above reasoning applies to any situation where an event has two possible outcomes and individuals have to take one of two possible actions (or choose between action and inaction). We have thus shown the following:

Theorem 4.1 (Disjunction Law)**.** *Let x_A, x_{A^c}, x_U be the proportions of people who would take an action if they knew that event A occurred, if they knew that A didn't occur, and if they didn't know the outcome, respectively. If everyone in the population abides by the Sure-Thing Principle, then*

$$x_A + x_{A^c} - 1 \leq x_U \leq x_A + x_{A^c}. \tag{15}$$

Because the only assumption that theorem 4.1 makes is the Sure-Thing Principle, a violation of eq. (15) implies that there are individuals in the population that exhibit a (within-subjects) disjunction effect.

4.2 Sufficiency of the disjunction law

We now turn to the following two questions: Suppose that eq. (15) is satisfied. Is it possible that some (or all) individuals violate the Sure-Thing Principle (i.e., exhibit the WSDE)? If yes, is there some other relation (perhaps stricter than eq. (15)) that can be checked and whose violation would allow us to infer the WSDE?

The answer to the first question is positive. Table 2 gives an example where every single person violates the Sure-Thing Principle, yet eq. (15) is satisfied.[6] In a within-subjects study, individual-level data can be used to demonstrate this. In between-subjects experiments, however, such data is not available, hence we have to rely on population-level quantities like x_U, x_A, and x_{A^c}. It is thus reasonable to ask whether a violation of the Sure-Thing Principle can be possibly inferred from observing x_U, x_A, and x_{A^c} while eq. (15) holds. Here we answer this question in the negative: if we have access only to the population-level quantities x_U, x_A, and x_{A^c} (in particular, if the data was obtained in a between-subjects study), eq. (15) is not only necessary, but also a sufficient condition for the data to be *consistent* with the Sure-Thing Principle. In other words, if eq. (15) is satisfied, then it is impossible to infer a violation of the STP using such data only.

To make the above claims precise, we assume a classical decision-theoretic setting, where each individual is characterized by their preference in the three variants of the question. We use a convention similar to Broekaert et al. (2020), denoting these preferences with triplets of letters of the form 'ynn', where the first two letters describe the preference under the two known-outcome variants (A and A^c, in this order) and the third letter describes the preference in the unknown outcome variant (U). Preference 'y' (yes) means that the participant would take the available action (e.g., take the second bet) in the corresponding variant, while 'n' (no) means that they would not.

The various preference combinations split the population into $2^3 = 8$ groups. Let $x_{yyy}, x_{yyn}, \ldots, x_{nnn}$ denote the proportions of the population that belong to these 8 groups. These proportions must clearly satisfy

$$x_{ijk} \geq 0, \quad \text{for all } \ i, j, k \in \{y, n\} \tag{16}$$

$$\text{and} \quad \sum_{i,j,k \in \{y,n\}} x_{ijk} = 1. \tag{17}$$

[6] The example given in Table 2 also disproves the claim that the STP is a consequence of the Law of Total Probability. Indeed, the STP is violated in this example while eqs. (9) and (10) hold. To check eqs. (9) and (10), note that by the last row of table 2, $\mathbb{P}_u(G) = \mathbb{P}_r(G \mid W) = \mathbb{P}_r(G \mid L) = 0.5$. Moreover, because G and W, as well as G and L, are independent when the outcome of the first bet is unrevealed, we also have $\mathbb{P}_u(G \mid W) = \mathbb{P}_u(G \mid L) = \mathbb{P}_u(G) = 0.5$. Finally, $\mathbb{P}_r(G)$ is equal to the probability that the participant is of type 1, hence equal to 0.5 as well. Recalling that $\mathbb{P}(W) + \mathbb{P}(L) = 1$, we find that eqs. (9) and (10) are indeed satisfied.

In a between-subjects setting, these proportions are not directly observed. Instead, we observe the quantities x_U, x_A and x_{A^c},[7] which can be expressed in terms of the previous proportions as (see also Broekaert et al., 2020)

$$x_U = x_{yyy} + x_{yny} + x_{nyy} + x_{nny}, \tag{18}$$
$$x_A = x_{yyy} + x_{yyn} + x_{yny} + x_{ynn}, \tag{19}$$
$$x_{A^c} = x_{yyy} + x_{yyn} + x_{nyy} + x_{nyn}. \tag{20}$$

To check these equations, observe that the first one expresses the proportion of people who would take a bet (‘y’) in the unknown outcome condition, thus we have to sum all those variables with a ‘y’ in the third index position. The other two equations can be checked similarly.

Recall that we are interested in inferring a violation of the Sure-Thing Principle. The STP states that if an individual prefers to take the suggested action in both known-outcome variants, they must also prefer to take it in the unknown outcome variant. In other words, if the STP holds, we must have $x_{yyn} = 0$, i.e., there must be no individuals who prefer to take the action in the first two variants but not in the third one. Applying a similar logic to not taking the action, we conclude that $x_{nny} = 0$ as well, again as a consequence of the STP. On the other hand, the STP makes no prediction about individuals who take the action in one known-outcome variant and not in the other, implying that it imposes no restrictions on x_{yny}, x_{ynn}, x_{nyy}, and x_{nyn}. Also, it does not say anything about how many people are willing to take the action in all variants or in none of the variants, that is, it imposes no restrictions on x_{yyy} or x_{nnn} either. To summarize, the Sure-Thing Principle is equivalent to $x_{yyn} = x_{nny} = 0$.[8]

The following theorem says that, as long as the disjunction law (eq. (15)) is satisfied, no violation of the Sure-Thing Principle can be inferred.

Theorem 4.2 (Sufficiency of the Disjunction Law)**.** *Suppose that $x_U, x_A, x_{A^c} \in [0, 1]$ satisfy eq.* (15)*. Then, based only on eqs.* (16) *to* (20)*, the possibility that $x_{yyn} = x_{nny} = 0$ cannot be excluded.*

The proof is given in the appendix.

5 Discussion

We have shown that a widely held view among scholars who study the disjunction effect, namely the fact that eq. (1) is inconsistent with the Law of Total Probability, is not correct. Our results show that eq. (1) is behaviorally irrelevant and thus impact a large literature that has studied this relation, as we explain below. We have also derived a new *disjunction law* (eq. (15)), which can be used to check for a within-subjects disjunction effect (a violation of the Sure-Thing Principle) even in between-subjects studies. We have further shown that for between-subjects studies this disjunction law is the strictest test one can obtain; as long as it is satisfied, it is impossible to infer a violation of the STP.

There appears to be substantial confusion in the literature with regards to what the term disjunction effect refers to. The term was introduced by Tversky and Shafir (1992), apparently with the intended meaning of *a violation of the Sure-Thing Principle* (see also Shafir and Tversky, 1992), to which in the current manuscript we refer as the Within-Subjects Disjunction Effect (WSDE). Already in the same study though, the authors appear to assume that the WSDE follows from a relation of the form of eq. (1), which here we have termed the Population-Level Disjunction Effect (PLDE). Later authors have used the term disjunction effect for both, sometimes even using the WSDE as the theoretical definition but checking for the PLDE in their empirical analyses in the same study (Kühberger et al., 2001; Bagassi and Macchi, 2006; Busemeyer et al., 2006; Moreira and Wichert, 2016b). This would have been perfectly valid if the PLDE indeed implied the WSDE, but this is not true, as was shown by Lambdin and Burdsal (2007) (see also the counterexample in table 1 of the current paper). Unfortunately, this result has not received enough attention, and the belief that the WSDE (i.e., a violation of the Sure-Thing Principle) follows from the PLDE (eq. (1)) still seems to be held by many scholars (Li et al., 2010; Moreira and Wichert, 2016b; Pisano and Sozzo, 2020; Shan, 2022; Xin et al., 2022; Widdows et al., 2023; Mahalli and Pusuluk, 2024).

The consequences of the above confusion would have been moderate, had the PLDE held a value of its own. Until now this was believed to be the case, because the PLDE presumably implied a violation of the Law of Total Probability and consequently of classical decision theory. The first such claim seems to appear in Busemeyer et al. (2006) (although without explicit mention of the Law of Total Probability), but many later studies have repeated the same argument

[7] Here again we assume that we have a very large sample, so that estimation errors can be ignored.

[8] Using the fact that the STP amounts to $x_{yyn} = x_{nny} = 0$, we can get an alternative proof of theorem 4.1. Indeed, subtracting eqs. (19) and (20) from eq. (18) and setting $x_{yyn} = x_{nny} = 0$, we get $x_U - x_A - x_{A^c} = -x_{yyy} - x_{ynn} - x_{nyn}$, which is bounded between 0 and -1 (see eq. (16) and eq. (17)). That is, $-1 \leq x_U - x_A - x_{A^c} \leq 0$, from where eq. (15) immediately follows.

(Blutner and beim Graben, 2016; Broekaert et al., 2020; Pisano and Sozzo, 2020; Tesař, 2020b; Aerts et al., 2021; Pothos and Busemeyer, 2022; Shan, 2022; Wang et al., 2022; Widdows et al., 2023; Mahalli and Pusuluk, 2024). However, as we have shown here, the argument contains a flaw, and the Law of Total Probability turns out to be perfectly consistent with eq. (1). Consequently, eq. (1) appears entirely irrelevant for the cognitive sciences and we suggest that the term disjunction effect be dissociated from it.

Our results call into question claims of a disjunction effect that are based (at least partially) on checking some version of eq. (1) (Tversky and Shafir, 1992; Shafir and Tversky, 1992; Li and Taplin, 2002; Bagassi and Macchi, 2006; Busemeyer et al., 2006; Hristova and Grinberg, 2010; Li et al., 2010, 2012; Tesař, 2020b; Broekaert et al., 2020; Ziano et al., 2021). Although it is possible that a violation of classical decision theory (e.g., a Within-Subjects Disjunction Effect) can still be inferred in some or all of these studies, revisiting their results seems necessary. One way to check for a WSDE is by checking whether the disjunction law that we have identified (eq. (15)) is satisfied. For between-subjects studies, this law is both necessary and sufficient for the data to be consistent with the Sure-Thing Principle (theorems 4.1 and 4.2). For example, one of the experiments reported in Tversky and Shafir (1992) is a between-subjects two-stage gamble experiment, where 69% of the people said they would take the second bet if they had won the first one, 57% would take it if they had lost the first one, and 38% would take it if the outcome of the first bet was unknown. These numbers satisfy eq. (15), hence no violation of the Sure-Thing Principle can possibly be inferred.

For within-subjects studies, it is possible to use individual-level data to demonstrate a violation of the STP, which some of the above studies have indeed done. For example, in the within-subjects version of the two-stage gamble in Tversky and Shafir (1992), although the population-level preference data satisfies eq. (15) and thus cannot demonstrate a violation of the STP, their individual-level data shows that 30 of the 98 participants violate it.[9] Other within-subjects studies, in contrast, ignore the individual-level data and base their analyses on population-level preferences only (Busemeyer et al., 2006; Li et al., 2010; Hristova and Grinberg, 2010; Surov et al., 2019; Tesař, 2020b),[10] which is evidence for how much this literature has relied on eq. (1) as a means to show a disjunction effect.

Our results also affect a large literature that uses the PLDE as a benchmark for comparing cognitive models (Moreira and Wichert, 2016b) or as motivation for the introduction of new ones (Moreira and Wichert, 2018; Xin et al., 2022), especially quantum cognition and other non-classical models for which the standard probability calculus does not hold (Busemeyer et al., 2006; Pothos and Busemeyer, 2009; Khrennikov, 2010; Moreira and Wichert, 2016a; Broekaert et al., 2020; Pisano and Sozzo, 2020; Tesař, 2020b,a; Waddup et al., 2021; Aerts et al., 2021; Shan, 2022; Huang et al., 2023; Mahalli and Pusuluk, 2024; Snow et al., 2024). Given that the Law of Total Probability is not violated in studies of the disjunction effect, a principal argument for using such models needs to be reconsidered. We are not claiming in any way that these models are not useful, but that the PLDE cannot be used to motivate them.

Part of our argument in showing that the Law of Total Probability is not violated by eq. (1) was that a participant may justifiably take a different action when a given outcome of an initial event is revealed to them vs. when the same outcome is not revealed. In several related experimental studies, the treatment does not consist in revealing the outcome, but rather in asking the participant to make a guess about it (Croson, 1999; Tesař, 2020a,b). The argument we have advanced in the present paper does not apply to these studies. Here we have only concerned ourselves with the standard context in which the term disjunction effect has been used, as introduced originally by Tversky and Shafir (1992), namely when contrasting revealed vs. unrevealed information, rather than guessing vs. not guessing.

Our paper makes the point that great care is needed when interpreting conditional probabilities; conditioning on an event must not be confused with applying an external manipulation that changes the underlying probability distribution itself. This distinction between conditioning and applying an external manipulation is a central theme in the study of causality (Pearl, 2009).[11] Further research is needed in order to check whether other apparent violations of coherence criteria can be better understood by making a similar distinction. More generally, our results illustrate that using precise mathematical language and notation is necessary in order to avoid incorrect claims about the statistical predictions that follow from assumptions on individuals' behavior (see also Gigerenzer, 1991).

References

Aerts, D., M. Sassoli de Bianchi, S. Sozzo, and T. Veloz (2021). Modeling human decision-making: An overview of the brussels quantum approach. *Foundations of Science 26*, 27–54.

[9] 26 individuals responded that they would take the gamble in both known-outcome variants, but not in the unknown-outcome variant; 4 people said they would not gamble in either known-outcome variant, but they would when the outcome was unknown (Tversky and Shafir, 1992, Table 1).

[10] Our reference to Tesař (2020b) concerns the experiment that replicates previous studies. The main experiment reported does not fit into the framework we are considering here.

[11] See Pearl (2016) for a treatment of the Sure-Thing Principle using the language of causal reasoning.

Bagassi, M. and L. Macchi (2006). Pragmatic approach to decision making under uncertainty: The case of the disjunction effect. *Thinking & reasoning 12*(3), 329–350.

Blutner, R. and P. beim Graben (2016). Quantum cognition and bounded rationality. *Synthese 193*, 3239–3291.

Broekaert, J., J. Busemeyer, and E. Pothos (2020). The disjunction effect in two-stage simulated gambles. an experimental study and comparison of a heuristic logistic, markov and quantum-like model. *Cognitive Psychology 117*, 101262.

Busemeyer, J. R., M. R. Matthew, and Z. Wang (2006). A quantum information processing explanation of disjunction effects. In *Proceedings of the annual meeting of the cognitive science society*, Volume 28.

Croson, R. T. (1999). The disjunction effect and reason-based choice in games. *Organizational Behavior and Human Decision Processes 80*(2), 118–133.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European review of social psychology 2*(1), 83–115.

Hristova, E. and M. Grinberg (2010). Testing two explanations for the disjunction effect in prisoner's dilemma games: Complexity and quasi-magical thinking. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 32.

Huang, J., J. Busemeyer, Z. Ebelt, and E. Pothos (2023). Quantum sequential sampler: a dynamical model for human probability reasoning and judgments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 45.

Khrennikov, A. (2010). *Ubiquitous quantum structure*. Springer.

Khrennikov, A. (2015). Quantum-like modeling of cognition. *Frontiers in Physics 3*, 77.

Khrennikov, A. (2022). Contextual probability in quantum physics, cognition, psychology, social science, and artificial intelligence. In *From Electrons to Elephants and Elections: Exploring the Role of Content and Context*, pp. 523–536. Springer.

Kühberger, A., D. Komunska, and J. Perner (2001). The disjunction effect: Does it exist for two-step gambles? *Organizational Behavior and Human Decision Processes 85*(2), 250–264.

Lambdin, C. and C. Burdsal (2007). The disjunction effect reexamined: Relevant methodological issues and the fallacy of unspecified percentage comparisons. *Organizational Behavior and Human Decision Processes 103*(2), 268–276.

Li, S., C.-M. Jiang, J. C. Dunn, and Z.-J. Wang (2012). A test of "reason-based" and "reluctance-to-think" accounts of the disjunction effect. *Information Sciences 184*(1), 166–175.

Li, S. and J. E. Taplin (2002). Examining whether there is a disjunction effect in prisoner's dilemma games. *Chinese Journal of Psychology*.

Li, S., Z.-J. Wang, L.-L. Rao, and Y.-M. Li (2010). Is there a violation of savage's sure-thing principle in the prisoner's dilemma game? *Adaptive Behavior 18*(3-4), 377–385.

Mahalli, N. F. and O. Pusuluk (2024). What is quantum in probabilistic explanations of the sure-thing principle violation? *BioSystems 238*, 105180.

Maruyama, Y. (2020). Rationality, cognitive bias, and artificial intelligence: a structural perspective on quantum cognitive science. In *Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pp. 172–188. Springer.

Moreira, C. and A. Wichert (2016a). Quantum-like bayesian networks for modeling decision making. *Frontiers in psychology 7*, 163811.

Moreira, C. and A. Wichert (2016b). Quantum probabilistic models revisited: The case of disjunction effects in cognition. *Frontiers in Physics 4*, 26.

Moreira, C. and A. Wichert (2018). Are quantum-like bayesian networks more powerful than classical bayesian networks? *Journal of Mathematical Psychology 82*, 73–83.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2016). The sure-thing principle. *Journal of Causal Inference 4*(1), 81–86.

Pisano, R. and S. Sozzo (2020). A unified theory of human judgements and decision-making under uncertainty. *Entropy 22*(7), 738.

Pothos, E. M. and J. R. Busemeyer (2009). A quantum probability explanation for violations of'rational'. *Proceedings of the Royal Society B 276*(1665).

Pothos, E. M. and J. R. Busemeyer (2022). Quantum cognition. *Annual review of psychology 73*, 749–778.

Pothos, E. M., G. Perry, P. J. Corr, M. R. Matthew, and J. R. Busemeyer (2011). Understanding cooperation in the prisoner's dilemma game. *Personality and Individual Differences 51*(3), 210–215.

Savage, L. J. (1954). The foundations of statistics.

Shafir, E. and A. Tversky (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive psychology 24*(4), 449–474.

Shan, Z.-H. (2022). Brainwave phase stability: Predictive modeling of irrational decision. *Frontiers in Psychology 13*, 617051.

Snow, L., V. Krishnamurthy, and B. M. Sadler (2024). Quickest detection for human-sensor systems using quantum decision theory. *IEEE Transactions on Signal Processing*.

Surov, I. A., S. V. Pilkevich, A. P. Alodjants, and S. V. Khmelevsky (2019). Quantum phase stability in human cognition. *Frontiers in Psychology 10*, 453145.

Tesař, J. (2020a). How do social norms and expectations about others influence individual behavior? a quantum model of self/other-perspective interaction in strategic decision-making. *Foundations of science 25*, 135–150.

Tesař, J. (2020b). A quantum model of strategic decision-making explains the disjunction effect in the prisoner's dilemma game. *Decision 7*(1), 43.

Tversky, A. and E. Shafir (1992). Choice under conflict: The dynamics of deferred decision. *Psychological science 3*(6), 358–361.

Waddup, O., P. Blasiak, J. M. Yearsley, B. W. Wojciechowski, and E. M. Pothos (2021). Sensitivity to context in human interactions. *Mathematics 9*(21), 2784.

Wang, Z., J. R. Busemeyer, and B. deBuys (2022). Beliefs, actions, and rationality in strategical decisions. *Topics in Cognitive Science 14*(3), 492–507.

Widdows, D., J. Rani, and E. M. Pothos (2023). Quantum circuit components for cognitive decision-making. *Entropy 25*(4), 548.

Xin, X., M. Sun, B. Liu, Y. Li, and X. Gao (2022). A more realistic markov process model for explaining the disjunction effect in one-shot prisoner's dilemma game. *Mathematics 10*(5), 834.

Ziano, I., M. F. Kong, H. J. Kim, C. Y. Liu, S. C. Wong, B. L. Cheng, and G. Feldman (2021). Replication: Revisiting tversky and shafir's (1992) disjunction effect with an extension comparing between and within subject designs. *Journal of Economic Psychology 83*, 102350.

Appendix

Proof of theorem 4.2. We need to show that for any possible values of x_U, x_A and x_{A^c} in $[0,1]$, as long as eq. (15) is satisfied, there exist values for the six variables x_{yyy}, x_{yny}. x_{nyy}, x_{ynn}, x_{nyn}, and x_{nnn}, such that they are all non-negative, their sum is 1, and the following equations are satisfied:

$$x_U = x_{yyy} + x_{yny} + x_{nyy}, \tag{21}$$
$$x_A = x_{yyy} + x_{yny} + x_{ynn}, \tag{22}$$
$$x_{A^c} = x_{yyy} + x_{nyy} + x_{nyn}. \tag{23}$$

(These equations follow from eqs. (18) to (20) by setting $x_{yyn} = x_{nny} = 0$.)

We have the following four cases:

- If $x_U \leq \min\{x_A, x_{A^c}\}$, we may set

$$\begin{aligned} x_{nyy} = x_{yny} &= 0 \\ x_{yyy} &= x_U \\ x_{ynn} &= x_A - x_U \\ x_{nyn} &= x_{A^c} - x_U \\ x_{nnn} &= 1 - x_A - x_{A^c} + x_U. \end{aligned} \tag{24}$$

- If $x_A \geq x_U \geq x_{A^c}$, we may set
$$\begin{aligned}x_{nyy} = x_{nyn} &= 0 \\ x_{yyy} &= x_{A^c} \\ x_{yny} &= x_U - x_{A^c} \\ x_{ynn} &= x_A - x_U \\ x_{nnn} &= 1 - x_A\end{aligned} \tag{25}$$
- The case $x_{A^c} \geq x_U \geq x_A$ is similar to the previous one with the roles of x_A and x_{A^c} reversed.
- If $x_U \geq \max\{x_A, x_{A^c}\}$, we may set
$$\begin{aligned}x_{ynn} = x_{nyn} &= 0 \\ x_{yyy} &= x_A + x_{A^c} - x_U \\ x_{nyy} &= x_U - x_A \\ x_{yny} &= x_U - x_{A^c} \\ x_{nnn} &= 1 - x_U.\end{aligned} \tag{26}$$

The only requirements that are non-trivial to check are eqs. (21) to (23) in the last case. By adding the equations for x_{yyy}, x_{nyy}, and x_{yny} in eq. (26), we get $x_U = x_{yyy} + x_{nyy} + x_{yny}$, which is eq. (21). Now substituting this value of x_U in the equations for x_{nyy} and x_{yny} and solving for x_A and x_{A^c}, respectively, we get

$$\begin{aligned}x_A &= x_{yyy} + x_{yny} \quad \text{and} \\ x_{A^c} &= x_{yyy} + x_{nyy},\end{aligned} \tag{27}$$

which are the same as eqs. (22) and (23), given that $x_{ynn} = x_{nyn} = 0$ in this case. □

Observe in the above proof that eq. (15) is needed only to show non-negativity of x_{nnn} in the first case and of x_{yyy} in the last case.