Journal de la Société Française de Statistique
Vol. 161 No. 1 (2020)

Simpson's paradox, a tale of causality

Titre: Le paradoxe de Simpson, un conte causal

Antoine Chambaz[1] , Isabelle Drouet[2] and Sonia Memetea[3]

Abstract: For the mathematically wary and unwary alike, Simpson's paradox may well function as a permanent invitation to error. We present Simpson's paradox and discuss its nature based on three examples. It appears that to run afoul of Simpson's paradox it suffices to *(a)* conflate an invalid probabilistic reasoning with a valid instance of unassailable causal reasoning, or *(b)* confuse the evidential concept of learning from observation, which for rational agents proceeds by conditioning on the evidence, with the causal concept of acting, represented in causal analysis by the operation of intervening in a causal graph.

Résumé : Le paradoxe de Simpson peut induire en erreur jusqu'au mathématicien prudent. Nous présentons le paradoxe de Simpson et discutons sa nature en nous appuyant sur trois exemples. Il apparaît que pour se faire prendre à son piège, il suffit *(a)* de combiner un raisonnement probabiliste hasardeux avec un raisonnement causal inattaquable, ou bien *(b)* de confondre le concept évidentiel d'apprentissage à partir de l'observation, qui pour des agents rationnels procède par conditionnement selon les données, avec le concept causal d'action tel qu'il est représenté en analyse causale par une intervention dans un graphe.

Keywords: causality, Simpson's paradox
Mots-clés : causalité, paradoxe de Simpson
AMS 2000 subject classifications: 60-01, 62A01

[1] MAP5 (UMR CNRS 8145), Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris cedex 06.
E-mail: antoine.chambaz@parisdescartes.fr
[2] Sciences, Normes, Décision (SND, FRE CNRS 3593), Sorbonne Université.
E-mail: isabelle.drouet@paris-sorbonne.fr
[3] E-mail: sonia.memetea@gmail.com

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

1. Preamble

En attendant Godot, *Waiting for Godot* in English, is a famous play written by Samuel Beckett (1952). Vladimir and Estragon, two vagrants, wait for the arrival of someone named Godot. While waiting they engage in a variety of discussions. The play has two acts. We imagine a third one. The French version of the first part of the third act is presented in Section 1.1 and its English translation in Section 1.2. Presented in Sections 8.1 and 8.2, the French and English versions of the second part will close the article.

1.1. Acte troisième, première partie

Lendemain. Même heure. Même endroit.

L'arbre porte quelques fruits.

Estragon est agenouillé sous l'arbre, pieds nus. Il compte sur ses doigts, ramasse un petit caillou, le place dans une de ses chaussures, recommence.

Entre Vladimir.

VLADIMIR. – Encore toi ! (*Estragon s'arrête mais ne lève pas la tête. Vladimir va vers lui.*) Viens que je t'étreigne !

ESTRAGON. – Ne me touche pas !

Vladimir suspend son vol, peiné. Silence.

VLADIMIR. – Veux-tu que je m'en aille ? (*Un temps*). Gogo ! (*Un temps. Vladimir le regarde avec attention.*) On t'a battu ? (*Un temps.*) Gogo ! (*Estragon se tait toujours, la tête basse.*) Où as-tu passé la nuit ? (*Silence. Vladimir avance.*)

ESTRAGON. – Ne me touche pas ! Ne me demande rien ! Ne me dis rien ! Reste avec moi !

VLADIMIR. – Est-ce que je t'ai jamais quitté ?

ESTRAGON. – Tu m'as laissé partir.

VLADIMIR. – Regarde-moi ! (*Estragon ne bouge pas. D'une voix tonitruante.*) Regarde-moi, je te dis !

Estragon lève la tête. Ils se regardent longuement, puis soudain s'embrassent. Fin de l'étreinte.

ESTRAGON. – Quelle journée !

VLADIMIR. – Qui t'a esquinté ? Toujours les mêmes ?

ESTRAGON. – Les mêmes quoi ?

VLADIMIR. – Toi aussi, tu dois être content, au fond, avoue-le.

ESTRAGON. – Content de quoi ?

VLADIMIR. – De m'avoir retrouvé.

ESTRAGON. – Tu crois ? Et toi ?

VLADIMIR. – Je ne me suis pas levé de la nuit, pas une seule fois.

ESTRAGON *(tristement)*. – Tu vois Didi, tu pisses mieux quand je ne suis pas là.

VLADIMIR. – Dis, Je suis content.

ESTRAGON. – Je suis content.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

VLADIMIR. – Moi aussi.

ESTRAGON. – Moi aussi.

VLADIMIR. – Nous sommes contents.

ESTRAGON. – Nous sommes contents. (*Silence.*) Qu'est-ce qu'on fait, maintenant qu'on est content ?

VLADIMIR. – On attend Simpson.

ESTRAGON. – C'est vrai. (*Silence.*) Et on joue.

VLADIMIR. – A se pendre ?

ESTRAGON. – De quoi tu parles ?

VLADIMIR. – Est-ce possile que tu aies oublié déjà ?

ESTRAGON *(agacé)*. – Bon, on joue ou quoi ?

VLADIMIR *(vexé)*. – On joue.

Ils s'assoient au pied de l'arbre. Estragon rapproche ses chaussures.

ESTRAGON. – Dans mes chaussures. Celle de gauche, trois petits fruits du saule et sept petits caillous. Et encore douze gros fruits et dix-huit gros caillous.

VLADIMIR. – Et dans celle de droite ?

ESTRAGON *(il regarde sur le côté, inquiet)*. – Quoi ? Je ne vois rien.

VLADIMIR *(excédé)*. – Mais non, dans ta chaussure.

ESTRAGON. – Ah. Vingt-et-un petits fruits, neuf petits caillous. Et puis huit gros fruits et deux gros caillous.

VLADIMIR. – Ça promet...

ESTRAGON *(l'œil brillant)*. – Maintenant écoute. Imagine que je vide mes chaussures en tas là devant nous. Pour tirer un fruit les yeux fermés, tu préfèrerais piocher parmi les petites choses ou les grosses ?

VLADIMIR *(sa curiosité piquée, il réfléchit l'espace d'un instant)*. – Les petites.

ESTRAGON *(malicieusement)*. – Bien. Et pour tirer un fruit de ma chaussure gauche, tu préfèrerais piocher parmi ?

VLADIMIR *(le front plissé, puis sentencieusement)*. – Les grosses.

ESTRAGON *(d'un air absent)*. – Et de ma chaussure droite ?

VLADIMIR *(troublé)*. – Encore les grosses.

ESTRAGON. – Es-tu d'accord que cette chose que tu tirerais du tas sortirait forcément de l'une de mes chaussures.

VLADIMIR *(dédaigneusement)*. – Bien entendu.

ESTRAGON *(doctement)*. – Et pourtant tu préfèrerais piocher parmi les petites choses quand bien même celle que tu tirerais sortirait de l'une de mes chaussures et que si tu pouvais piocher directement dans chacune d'elles tu préfèrerais tirer une grosse chose.

VLADIMIR *(avec irritation)*. – Tu ne m'amuses pas Gogo. Je croyais qu'on allait jouer.

Silence.

ESTRAGON *(triste)*. – Et maintenant, Didi, on fait quoi ?

VLADIMIR. – Attendons de voir ce qu'il va nous dire.

ESTRAGON. – Qui ?

VLADIMIR. – Simpson.

ESTRAGON. – Et s'il ne vient pas ?

VLADIMIR. – Nous aviserons.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

1.2. Third act, first part

Next day. Same time. Same place.

The tree bears some fruits.

Estragon kneels under the three, bare foot. He counts on his fingers, takes a small pebble, puts it in one of his shoes, and again.

Vladimir comes in.

VLADIMIR. – You again! (*Estragon stops but doesn't look up. Vladimir goes to him.*) Come here that I embrace you!

ESTRAGON. – Don't touch me!

Vladimir holds back, pained. Silence.

VLADIMIR. – Do you want me to go away? *(Pause.)* Gogo! *(Pause. Vladimir observes him attentively.)* Did they beat you? *(Pause.)* Gogo! *(Estragon remains silent, head bowed.)* Where did you spend the night? *Silence. Vladimir advances.*

ESTRAGON. – Don't touch me! Don't question me! Don't speak to me! Stay with me!

VLADIMIR. – Did I ever leave you?

ESTRAGON. – You let me go.

VLADIMIR. – Look at me! *(Estragon doesn't move. Violently.)* Will you look at me!

Estragon raises his head. They look long at each other, then suddenly embrace. End of the embrace.

ESTRAGON. – What a day!

VLADIMIR. – Who beat you? The same lot as usual?

ESTRAGON. – The same what?

VLADIMIR. – You must be happy too, deep down, if you only knew it.

ESTRAGON. – Happy about what?

VLADIMIR. – To be back with me again.

ESTRAGON. – Would you say so? And you?

VLADIMIR. – I didn't get up in the night, not once!

ESTRAGON *(sadly)*. – You see, Didi, you piss better when I'm not there.

VLADIMIR. – Say, I am happy.

ESTRAGON. – I am happy.

VLADIMIR. – So am I.

ESTRAGON. – So am I.

VLADIMIR. – We are happy.

ESTRAGON. – We are happy. *(Silence.)* What do we do now, now that we are happy?

VLADIMIR. – Wait for Simpson.

ESTRAGON. – It's true. *(Silence.)* And we play.

VLADIMIR. – Hanging ourselves?

ESTRAGON. – What are you talking about?

VLADIMIR. – Is it possible you've forgotten already?

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

ESTRAGON *(angrily)*. – So, do we play or not?

VLADIMIR *(vexed)*. – We play.

They sit under the tree. Estragon brings his shoes closer.

ESTRAGON. – In my shoes. The left one, three little fruits from the willow and seven little pebbles. Moreover, twelve big fruits and eighteen big pebbles.

VLADIMIR. – And to the right?

ESTRAGON *(he looks on the side, worried)*. – What? I see nothing.

VLADIMIR *(exasperated)*. – No, in your shoe.

ESTRAGON. – Ah. twenty-one small fruits, nine small pebbles. Then eight big fruits and two big pebbles.

VLADIMIR. – I can't wait...

ESTRAGON *(bright-eyed)*. – Now listen. Imagine that I empty my shoes in a heap here in front of us. To draw a fruit your eyes closed, would you prefer to pick from the small things or the big ones?

VLADIMIR *(his curiosity piqued, he thinks for an instant)*. – The small ones.

ESTRAGON *(mischievously)*. – Good. And to draw a fruit from my left shoe, you would prefer to pick from?

VLADIMIR *(forehead wrinkled, then sententious)*. – The big ones.

ESTRAGON *(acting distant)*. – And from my right shoe?

VLADIMIR *(troubled)*. – From the big ones again.

ESTRAGON. – Do you agree that this thing that you would draw from the heap would necessarily come from one of my shoes.

VLADIMIR *(disdainfully)*. – Of course.

ESTRAGON *(learnedly)*. – And yet you would prefer to draw among the small things even though the thing you would pick would come from one of my shoes and that if you could draw directly from either shoe you would prefer to pick a big thing.

VLADIMIR *(irritated)*. – You don't amuse me Gogo. I thought we were going to play.

Silence.

ESTRAGON *(sadly)*. – And now, Didi, what do we do?

VLADIMIR. – Let's wait and see what he says.

ESTRAGON. – Who?

VLADIMIR. – Simpson.

ESTRAGON. – And if he doesn't come?

VLADIMIR. – We'll see when the time comes.

1.3. Overview

The literature on Simpson's paradox is vast. Early accounts date back to (Yule, 1900; Pearson, 1900; Cohen and Nagel, 1934) well before the phrase "Simpson's paradox" was coined by Blyth in 1972. We do not aim to provide an original solution or an extensive review of the literature, but rather to offer a teaser to the present volume.

The reason why this introductory article, in a volume on causal analysis, is devoted to Simpson's paradox is that we endorse Pearl (2011, 2014)'s view that this paradox is essentially a causal

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

one. Following Quine (1976) we will construe a paradox as "a conclusion that at first sounds absurd but has an argument to sustain it". Our claim is that in the case of Simpson's paradox the air of absurdity originates in a conflicting (often covert) causal interpretation of the premises of the sustaining argument. More precisely, it stems from a conflict between the probabilistic characterization of those situations that count as instances of the paradox and the naive causal reading we more or less naturally make of these situations. The main aim of the article, built upon (Memetea, 2015), is to substantiate this claim and to explain how unreflective reasoning based on an undue causal interpretation can elicit the surprising elements evinced bySimpson's paradox on a first encounter.

To pursue this argument, we turn to a close examination of three concrete examples. Our investigation aims to bring out into the open the causal nature of the psychological surprise elicited by Simpson's paradox, and thus to throw light on the nature of the paradox.

2. Simpson's paradox illustrated

The time-honored probabilistic characterization of Simpson's paradox is silent on the question of causality. Narrowly construed, Simpson's paradox is simply a pattern of probabilistic reversal, which involves the reversal or cancellation of a global association between two variables, when conditioned upon a third. Yet to bring out its specific interest and challenge, it is necessary to supplement the probabilistic account with narrative information, to characterize *scenarios*, that is, interpreted instances of the mathematical pattern of reversal adumbrated above. In this section we present three plausible scenarios, both to exemplify the phenomenon and to structure the ensuing discussion.

The first example is taken from (Julious and Mullee, 1994). The actual data are reported in Table 1.

Example A (Comparing kidney stone removal modus operandi)**.**

> Charig et al. (1986) undertook a historical comparison of success rates in removing kidney stones. Overall, open surgery (method m_1) had a smaller success rate than percutaneous nephrolithotomy (method m_2). However, the success rates looked rather different when stone diameter was taken into account. For stones of diameter smaller than 2cm, open surgery was more successful compared with percutaneous nephrolithotomy. Likewise, for stones of diameter larger than 2cm, open surgery was more successful than percutaneous nephrolithotomy.

Discussed at greater length in Section 3, Example A is a straightforward instance of Simpson's paradox, insofar as the association between method and success is quite evidently reversed when conditioning on the stone diameter. Yet even a cursory glance at Example A reveals obvious causal underpinnings. The puzzling probabilistic reversal baffles our causal intuition to the effect that if the best way to cure a disease in subpopulations is to apply a certain method, then that method ought to prevail in the aggregated population as a whole. But the probabilistic pattern *appears* to tell a different story; to wit, that contrary to our intuitive causal projection, the outperformed method seems to be the optimal choice at the aggregated level (later on we will argue that it is not, and that it is a mere probabilistic quirk, an artifact of a specific kind of causal model).

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

Example A is an easy case for us in the sense that its causal dimension is quite explicit. For this same reason, it is insufficient to establish that Simpson's paradox in general is causal in nature. This motivates the introduction and discussion of Example B. Taken from (Bandyoapdhyay et al., 2011), it is attributed to J. G. Bennett. Table 2 gives a synthetic numerical example.

TABLE 1. ***Comparing kidney stone removal modus operandi.*** *Success in removing kidney stones. Method m_1 stands for open surgery, method m_2 for percutaneous nephrolithotomy, and d for the stone's diameter.*

	d <2cm		d ≥2cm		all ds	
	success	failure	success	failure	success	failure
m_1	81	6	192	71	273	77
m_2	234	36	55	25	289	61

Example B (Bags of marbles)**.**

Suppose we have two bags of marbles, all of which are either big or small, and red or blue. Suppose that in each bag the proportion of red to blue is higher among the large than among the small marbles. Now suppose we pour all the marbles from both bags into one box. Would we expect it to be more probable to be red for large marbles of the box than for small marbles of the box? Most of us would indeed expect it and therefore would be surprised to find out that this may fail to be the case. Table 2 exhibits such a failure.

TABLE 2. ***Bags of marbles.*** *A numerical example where polarity of the dependence between two variables (color and size) is reversed when disaggregating with respect to a third variable (bag), see Example B.*

	bag 1		bag 2		bags 1, 2	
	red	blue	red	blue	red	blue
large	12	18	8	2	20	20
small	3	7	21	9	24	16

In stark contrast with Example A, no causality is at play in Example B. Does our uneasiness merely stem from what we would like to call an "arithmetical surprise", that is, the revelation that seemingly natural but in fact reckless arithmetical derivations are invalid? Yes, but only partly. Section 4 mainly aims at showing that there is more to Example A than the mere arithmetical surprise. We argue that our uneasiness is completely explained only once we realize that we cannot help but reading the situation causally. To sustain this claim, in presence of marbles, we argue by appealing to several betting games that involve marbles.

Elaborating on Sections 3 and 4, Section 5 finally develops our analysis of the nature of Simpson's paradox. We characterize it as a graded problem. Starting again from its probabilistic facet, we explore its evidential and, most importantly, causal dimensions. As the so called causal sure thing principle is at the core of the discussion, we devote Section 6 to sure thing principles and their relationship to Simpson's paradox.

All of this leaves open the question of the reality of Simpson's paradox. It does happen. However, one should not jump to conclusions too quickly. Consider for instance the following third and last example (which does not naturally lend itself to being accompanied by contingency tables because, contrary to Examples A and B, it does not reduce to counts).

Example C (Synthetic microbial system)**.**

The evolution of cooperation offers a puzzle to solve. In the presence of producers and non-producers of a "common good", where the latter benefit from the shared resource without bearing its cost of pro-

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

duction, why does natural selection allow the former to survive? Chuang et al. (2009) shed light on the problem by analyzing a series of synthetic microbial systems where the proportions of producers at the start of the experiment, controlled by design, range across 0%, 10%, ..., 90%, 100%. They find that, in each non-pure system, the proportion of producers decreases during the fixed duration of the experiment, revealing that non-producers are advantaged indeed and grow faster than producers. However, the authors also find that the overall proportion of producers in a mixture of equal volumes of each non-pure system is larger at the end of the experiment than at its start (when it was 45%).

Chuang et al. (2009) claim that Example C is another real-life example of Simpson's paradox. Is it, really? Section 7 answers this question and closes the article. It suggests guidelines to decide whether or not putative instances of Simpson's paradox are genuine.

3. Example A: a causal surprise

3.1. Association reversal

We make the statistical assumption that the empirical probabilities below estimate their theoretical counterparts. Being sure that they do would require knowing more about how the data set was built. To emphasize that the probabilities are empirical, we use the symbol P_n in lieu of P.

On the one hand, considering the stone diameters separately, we observe that

$$P_n(\text{success}|d<2,m_1)=\frac{81}{87}\approx 93\% \quad > \quad P_n(\text{success}|d<2,m_2)=\frac{234}{270}\approx 87\%, \tag{1}$$

$$P_n(\text{success}|d\geq 2,m_1)=\frac{192}{263}\approx 73\% \quad > \quad P_n(\text{success}|d\geq 2,m_2)=\frac{55}{80}\approx 69\% \tag{2}$$

(in probability, open surgery is more successful than percutaneous nephrolithomy to remove stones of both small and large diameters). On the other hand, neglecting the stone diameters, we observe that

$$P_n(\text{success}|m_1)=\frac{273}{350}\approx 78\% < P_n(\text{success}|m_2)=\frac{289}{350}\approx 83\% \tag{3}$$

(in probability, percutaneous nephrolithomy is more successful than open surgery). The data in Table 1 may be deemed surprising because they suggest that one method performs better than the other when stone diameter is taken into account, but performs worse when stone diameter is neglected.

3.2. Causal construal of the case

The misleading suggestion originates in an unwary causal construal of the *three* probabilistic statements

$$P(\text{success}|d<2,m_1) \quad > \quad P(\text{success}|d<2,m_2), \tag{4}$$

$$P(\text{success}|d\geq 2,m_1) \quad > \quad P(\text{success}|d\geq 2,m_2), \tag{5}$$

$$P(\text{success}|m_1) \quad < \quad P(\text{success}|m_2). \tag{6}$$

Interpreting causally (4), (5) and (6) would commit us to a medical model where the presumed causal mechanism responsible for the successful removal of stones both of small and of large

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

diameters somehow ceases to operate when stone diameter is neglected. Nothing in our understanding of causality can license this implication. This is where the paradox lies.

But perhaps the reader finds this statement a little peremptory, so let us try to better articulate it. Suppose that we are asked which surgical method to perform to remove a kidney stone. If (6) were to express a genuine causal statement, then we ought to recommend performing percutaneous nephrolithomy. However, if (4) and (5) were to be interpreted as genuine causal statements too, then we ought to recommend performing open surgery, for we *know* that the stone diameter is necessarily either smaller or larger than 2cm. This is unsettling because we end up with two contradictory answers to the same question.

To elaborate, this situation arises from the underdetermination of the causal structure underlying (4), (5), (6). It is simply impossible to go by the sole basis of these equations. To make a decision, it is necessary to supplement them with a causal model, for instance, a causal graph.

We assume that no other variable influences at least two among stone diameter, method and success. Therefore, there are 27 *possible* (as opposed to *plausible*) such graphs, of which we exclude the two cyclic ones. Some of the remaining graphs are not compatible with the dependence structure encapsulated in the data presented in Table 1. For instance, it cannot be the case that the graph features no arrow. However, the data alone do not single out a unique causal graph.

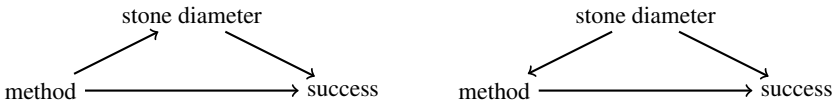

FIGURE 1. *Two causal graphs that are compatible with the data arising from the experiment of Example A. Both graphs assume (i) that stone diameter and method influence causally success and (ii) that stone diameter and method are correlated. In the LHS graph, stone diameter is a mediating variable between method and success. In the RHS graph, stone diameter has a causal influence on method.*

In particular, the two causal graphs of Figure 1 are undistinguishable based on the sole data. The LHS graph assumes that method influences causally success directly and indirectly, through stone diameter. Stone diameter is a mediating variable between method and success. Consequently, (6) can be interpreted causally, as opposed to (4) and (5). On the contrary, the RHS graph assumes that stone diameter influences causally both method and success. Then stone diameter is a confounder and confounding is at play. Consequently, (4) and (5) can be interpreted causally, as opposed to (6). In this light, Simpson's paradox can be viewed as an attempt to reconcile contradictory causal graphs such as those of Figure 1.

In Example A, the choice between graphs facilitates the solution to the puzzle. Intuitively given prior knowledge of cause and effect in the etiology of disease, it seems fair to assume *(i)* that both stone diameter and method do influence success causally, and *(ii)* that neither the method nor the success of the removal can causally influence stone diameter. The contrary would blatantly violate the chronology of events, which in general coincides with the causal ordering. Consequently, the choice of the RHS as the *correct* or most plausible causal representation of the scenario mandates the choice of open surgery.

The argument developed in the three previous paragraphs hinges on the assumption that no

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

other variable influences at least two among stone diameter, method and success. If, on the contrary, age (for instance) is identified as a factor susceptible to causally influence all of stone diameter, choice of surgical method and success of the removal, then (4) and (5) cannot be statements about causality anymore. Disaggregating the two LHS contingency tables in Table 1 by conditioning on age may, or may not, suggest a different rule to choose which surgical method to use on a case-by-case basis. To understand how conditioning on age may in turn reverse, say, (4), it suffices to understand how conditioning on stone diameter reverses (6).

3.3. Lessons for the representation of causality

All of this is quite difficult to express and still more to handle correctly. This establishes the usefulness of Pearl's *do*-operator in causal analysis (Pearl, 2000). In the remainder of the article, we will resort to *do*-calculus without further ado. We refer the unfamiliar reader to the above monograph or to (Shpitser, 2019) in the present special issue.

4. Example B: an arithmetical surprise?

Our analysis of Example A makes it a causal paradox, which baffles our causal intuitions or, more precisely, our intuitive causal interpretations of probabilistic associations. It seems, however, that such an analysis cannot be given for all instances of Simpson's paradox.

Attributed to J. G. Bennett by Bandyoapdhyay et al. (2011), the conception of Example B was driven by the intention to exclude causality. For this reason, there ought to be no causality involved in our discomfort with Example B. In this case it looks as if the surprise were arithmetical.

4.1. Arithmetical construal of the case

Consider the data in Table 2, which summarizes Example B. We observe that

$$P(\text{red}|\text{large}, \text{bag } 1) = \frac{12}{30} = 40\% \quad > \quad P(\text{red}|\text{small}, \text{bag } 1) = \frac{3}{10} = 30\%, \tag{7}$$

$$P(\text{red}|\text{large}, \text{bag } 2) = \frac{8}{10} = 80\% \quad > \quad P(\text{red}|\text{small}, \text{bag } 2) = \frac{21}{30} = 70\% \tag{8}$$

("in each bag, the proportion of large marbles that are red is bigger than the proportion of small marbles that are red") whereas

$$P(\text{red}|\text{large}) = \frac{20}{40} = 50\% < P(\text{red}|\text{small}) = \frac{24}{40} = 60\% \tag{9}$$

("overall, the proportion of large marbles that are red is not bigger than the proportion of small marbles that are red").

Here, probabilities are ratios or proportions that are computed out of contingency tables. Correspondingly, it can be argued that example B traces the following elementary mathematical fact: the inequalities

$$\frac{a}{a+b} > \frac{c}{c+d} \quad \text{and} \quad \frac{e}{e+f} > \frac{g}{g+h} \tag{10}$$

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

do not imply the inequality

$$\frac{a+e}{a+b+e+f} > \frac{c+g}{c+d+g+h}. \tag{11}$$

Likewise, the equality

$$\frac{1}{a}+\frac{1}{b}=\frac{1+1}{a+b}$$

holds if and only if $a=b\neq 0$. Now, a pupil may be surprised at first when they learn it. Yet, this peculiarity of elementary arithmetic does not qualify as a paradox... What makes Simpson's peculiarity a paradox? Our answer to this question is that the specificity of Simpson's paradox considered as an arithmetical oddity is precisely that it can be given a causal interpretation.

4.2. Causal reasoning in disguise

It must be conceded that facts about elementary arithmetic, however peculiar, are neither causal in nature, nor do they entail any particular relationship to a specific causal structure. It would therefore seem that *if* the explanation of Example B is satisfactorily given in terms of a priori mathematical facts about ratios, *then* it is otiose to pursue the characterization of the paradox in causal terms (at least in this specific instance; the case at large is moot). More precisely, the danger is that it is either plainly superfluous to supplement the a priori explanation causally, or it is plainly incoherent to appeal to the wrong sort of *explanans* in the attempt to shed light on the *explanandum*. Nevertheless, in our search for a unified causal explanation of Simpson's paradoxical phenomenon, we will explore this seemingly purely a-causal arithmetical scenario from a causal perspective. We hope to persuade the reader that this is neither perverse nor fruitless; on the contrary, it shows that causal reasoning is deeply ingrained and inescapable. In short, whilst we grant that the a priori explanation in terms of ratios satisfactorily explains the superficial arithmetical surprise, to account for the deep pressure of the paradox in full we cannot avoid causal reasoning, albeit in disguise.

The critical reader may already baulk at the introduction of a distinction between surface and depth in the nature of paradox, no less than at the new-fangled notion of causal reasoning in disguise. To dispel the obscurity of these terms, we will say at the outset just what we take the crucial terms to be. With respect to the former concern, we do not invidiously mean that arithmetical peculiarities count as surface peculiarities merely because they are arithmetical. Rather, we treat the arithmetical facts as surface because they are *easily* exhausted (by means of elementary algebra). What is harder to explain is (as we make clear below) the type of game-playing (optimal betting behavior) that we might still be puzzled by, and that only makes sense in light of an assumed causal story (which we term causal reasoning in disguise). Thus, as regards the latter concern, we take causal reasoning in disguise to mean acting *as if* there is a substantive causal story to tell, so as to make reasonable a certain type of strategy in a game-playing context.

We take it that recent advances in cognitive science make it plausible that we act in such a way. Indeed, as pinpointed by an anonymous referee, causal reasoning seems to be the most fundamental and easy to understand from among our different types of reasoning (Sloman and Fernbach, 2017, for instance). In our view, this also explains why a causal construal of Simpson's cases provides a good account for their paradoxicality. The reason why these cases keep

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

disturbing us even after we have clarified the underlying arithmetical facts is precisely that they touch upon entrenched, causal convictions.

Why games? We invoke games for two reasons. First, appealing to games serves a dialectical purpose in the presentation of our argument. But more importantly, the deeper reason is that we want to capitalize on the insatiable appetite for play that lies at the core of personhood (Caillois, 1958). Thus we believe that our strategy to uncover causal reasoning in disguise is both widely applicable and natural in the context of our example.

When we, for instance, are presented with Example B, we may be tempted to reconstruct it in imagination as a game the aim of which is to draw a ball of a particular color, say red, from the bags of marbles. In fact, several games can be considered along these lines — see Table 3 for a summary of the five games we focus on. In the simplest one, game I, we are merely given the opportunity to pick the first marble that we touch in the box where all marbles have been poured ("the box" for short). Since

$$P(\text{red}) = \frac{44}{80} = 55\%$$

the probability that we win, that is, draw a red marble, is 55%. In the second game, game II, we can choose the bag from which we pick the first marble that we touch. Because

$$P(\text{red}|\text{bag } 1) = \frac{15}{40} = 37.5\% < P(\text{red}|\text{bag } 2) = \frac{29}{40} = 72.5\%, \tag{12}$$

the optimal strategy for game II is to draw a marble from bag 2. Its winning probability equals 72.5% (versus 37.5% for a player who would pick the first marble that they touch in bag 1).

Games I and II are elementary because they neglect the size of the marbles. This remark prompts the introduction of two additional, more exciting games. Game II$^+$ extends game II: we can choose the bag from which we pick a marble **and**, *then*, we are allowed to select the marble based on its size.

Judging by (7) and (8), the optimal strategy for game II$^+$ is to select a large marble from bag 2. Its winning probability is a glorious 80%. Note that if we opt for selecting a marble from bag 1, then it is optimal to select a large one too, and the winning probability is the more modest 40%.

Note that the order of choice does not matter to the solution in game II$^+$ if played with the full description of the game in view. Suppose we are offered a variant in terms of first choosing size, and then choosing bag. Since one knows one is to choose some bag (even before making the choice of specific bag), and in each bag the proportion of red to blue marbles predominates among the large relative to the small, one might as well *first* choose size (large), and *then* choose bag (bag 2). Observe moreover that the independence of order of choice in game II $^+$ is premised on being able to *anticipate* that one will be offered the choice of bag at stage two; if one does not know one gets to choose bag as well, then the justification for choosing large falls through.

This marks an important difference between the two order-specific variants of game II$^+$. No knowledge of the further step of choosing size is necessary to justify choosing bag 2 in the original game II$^+$. To make the point even clearer, consider the games, choose bag, then *surprise!* choose size; versus, choose size, then *surprise!* choose bag. (The *surprise!* indicates the step at which it is revealed to the player that she can make a further choice.) The solutions in these

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

games come apart. The first player will rationally choose bag 2, then large. The second player will rationally choose small, then bag 2. It would seem that the justification for choosing bag 2 is order independent in the surprise games, whereas the justification for size is not.

Game I^+ extends game I in the same way as game II^+ extends game II. Concretely, in game I^+, we are allowed to select our marble from the box based on its size. For this game, (9) implies that the optimal strategy is to select a small marble, with a winning probability equal to 60%, to be compared to a winning probability of 50% if we selected a large marble from the box. Now, this seems peculiar, or even perhaps paradoxical, since selecting a large marble is always a better option than selecting a small marble when picking from either bag 1 or bag 2 in game II^+, with winning probabilities equal to 40% or 80%, respectively... The apparent incompatibility is striking.

In order to discredit the weak strategy consisting in selecting a large marble (not a small one) in game I^+, the simplest argument is to emphasize that in this a-causal story where none of bag, size and color can be viewed intrinsically either as (being part of) a cause or an effect, pouring all the marbles in the box dissolves the bag feature. With no causation behind the arithmetic, reasoning from the fact that it would be more advantageous to pick a large marble than a small one whichever bag the marble originates from is irrelevant.

A more informed argument hinges on causal graphs and *do*-calculus, in the spirit of the conclusions drawn in Section 3.3. The key causal graph representing Example B is given in the LHS graph of Figure 2. Naturally, the a-causal story results in a degenerate causal graph, where none of the bag, size and color features is a cause of another feature. The probabilistic dependence structure stems from the unobserved (hence the circle around it) common cause denoted by U.

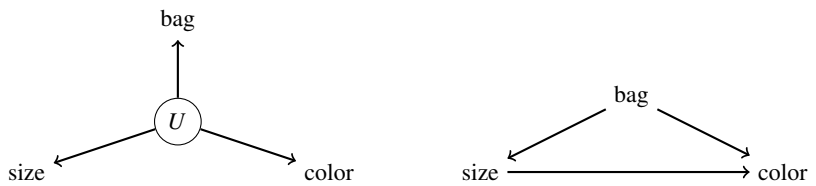

FIGURE 2. *Left: Degenerate causal graph representing the a-causal story of Example B. The probabilistic dependence structure between bag, size and color stems from the unobserved (hence the circle around it) common cause denoted by U. Right: Wrongly postulated causal graph representing the a-causal story of Example B.*

In Game I, drawing a marble boils down to sampling the latent U from its distribution, hence the observed features. In Game I^+, we are allowed to sample U from its conditional distribution given the resulting size. Symmetrically, in game II (in game II^+, respectively), we are allowed to sample U from its conditional distribution given the resulting bag (given the resulting bag and size couple, respectively). We recover the earlier conclusions about the optimal strategies from the relevant comparisons of conditional probabilities, see (7), (8), (9), (12).

We are now well equipped to explain why the optimal strategies in games I^+ and II^+ may seem contradictory. As already stated, the problem comes from an undue causal interpretation of the games. First we claim that, maybe because of the narrative of game II^+, we think of it in

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

terms of interventions in the causal graph represented in the RHS of Figure 2. Determining the optimal strategy within the postulated causal model gives the right answer, because

$$P'(\text{red}|do(\text{bag},\text{size})) = P(\text{red}|\text{bag},\text{size})$$

is maximized at (bag 2, large). The prime symbol is meant to emphasize that the *do* operator in the above equation refers to the causal graph represented in the RHS of Figure 2. Second, we believe that the causal story we tacitly (or covertly) appeal to in justifying game II$^+$ is surreptitiously extended to give the *wrong* answer to game I$^+$. In this case, determining the optimal strategy within the postulated causal model gives the wrong answer, because

$$\begin{aligned} P'(\text{red}|do(\text{large})) &= P(\text{bag 1}) \times P(\text{red}|\text{large},\text{bag 1}) + P(\text{bag 2}) \times P(\text{red}|\text{large},\text{bag 2}) \\ &= \frac{40}{80} \times \frac{12}{30} + \frac{40}{80} \times \frac{8}{10} = 60\%, \qquad (13) \\ P'(\text{red}|do(\text{small})) &= P(\text{bag 1}) \times P(\text{red}|\text{small},\text{bag 1}) + P(\text{bag 2}) \times P(\text{red}|\text{small},\text{bag 2}) \\ &= \frac{40}{80} \times \frac{3}{10} + \frac{40}{80} \times \frac{21}{30} = 50\%. \qquad (14) \end{aligned}$$

To sum up, we suggest that the paradox presented by simple, apparently purely arithmetical instances of Simpsons paradox in fact goes far deeper than the surface surprise involving ratios, notwithstanding the evident sufficiency of an a-causal a priori explanation of the latter. In choosing the optimal strategy for our series of simple games, we often act as though there were a causal story in the offing. In some cases this is appropriate and underwrites the optimal answer; in others it is not. Ultimately the paradox devolves upon the clash of causal intuition, however "purely arithmetical" the original set up we started off with. [1]

In conclusion, our uneasiness stems from an unduly causal construal of standard conditional probabilities. A causal reasoning is in disguise. The two salient features of our argument are the introduction of games on the one hand, and the assessment of whether it is adequate or not to reason in terms of interventions.

[1] Under what circumstances would a causal reasoning based on the causal graph in the RHS of Figure 2 be adequate? Answering this question complements our argument. It requires to make a causal story out of the originally a-causal story by changing the narrative considerably. We can for instance assert that we have handy six pouches of different textures and that, instead of pouring all the marbles in the box, first we gather them in four pouches consisting of marbles similar in size and originating from the same bag; second, we gather in two outer pouches the two inner pouches consisting of marbles from bag 1 on the one hand and from bag 2 on the other; third, we finally place those two outer pouches in the box. Thus, sampling a marble from the box decomposes now as the successive random selection of an outer pouch, then an inner pouch within the first one, then a marble from the second one. In order to reflect the overall distribution of marbles in the original story, we postulate that the two outer pouches are both sampled with probability one half and that within the pouch consisting of marbles from bag b ($b \in \{1,2\}$), the conditional probability to sample the pouch of large marbles equals $P(\text{large}|\text{bag } b)$. These possibly unequal probabilities may represent an unconscious ordering of textures by preference. In this more intricate story, picking a marble based on its size is a full-fledged causal intervention. Pearl's *do*-calculus yields (13) and (14), revealing that the optimal strategy in game I$^+$ is to pick a large marble. To implement this strategy, we would pick randomly one of the two outer pouches (with equal probabilities), extract the two inner pouches, assert which one contains large marbles, then finally pick one marble from it.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

TABLE 3. ***Irresistible games.*** *These games can be played in the context of Example B. In each of them, the aim is to draw a red marble.*

game	rule	optimal strategy	winning prob.
I	Pick the first marble that one touches from the box where all marbles have been poured.	no strategy	55.0%
I^+	Pick a marble from the box where all marbles have been poured knowing whether it is large or small.	pick a small marble (see (9))	60.0%
II	Choose a bag then pick from it the first marble that one touches.	pick from bag 2 (see (12))	72.5%
II^+	Choose a bag then pick a marble from it knowing whether it is large or small.	pick a large marble from bag 2 (see (7) and (8))	80.0%
III	Pick a bag (equiprobably), then pick a marble from it knowing whether it is large or small.	pick a large marble (see (7) and (8))	60%

5. Grades of paradoxical involvement

It is now time to take a step back and reflect on the grades of paradoxical involvement at play in Simpson's paradox. For the mathematically wary and unwary alike, Simpson's paradox may well function as a permanent invitation to error. To run afoul of Simpson's paradox it suffices to *(a)* conflate a probabilistically invalid inference with a valid instance of unassailable causal reasoning, or *(b)* confuse the evidential concept of learning from observation, which for rational agents proceeds by conditioning on the evidence, with the causal concept of acting, represented in causal analysis by the operation of intervening in a causal graph.

To make these points we present in order of ascending importance three glosses on the Simpson paradoxical reversal inequalities, in terms of comparative probability, evidential relevance, and causal bearing.

5.1. Comparative probability

Consider the following naive expectation concerning the ranking of contrastive effects:

Tentative principle: comparative contrast. If X is more likely given Y rather than $\neg Y$, both if Z and $\neg Z$, then X is more likely given Y rather than $\neg Y$, *tout court*.

In point of fact, any numerical instance of Simpson's paradox reformulated as a statement about comparative probability is a conclusive counterexample to the above tentative principle, see for instance (1), (2) and (3) from the discussion of Example A in Section 3. It shows that failure to attend to the possibility of paradox results in a tempting but disastrous pattern of probabilistically invalid reasoning.

5.2. Evidential relevance

In the context of the Bayesian framework, the gloss on Simpson's paradox in terms of contrastive probability connects seamlessly with the hallowed but not uncontested construal of evidential rel-

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

evance as probabilistic relevance, relative to the agent's background knowledge. Usually, the evidential Bayesian formalizes an agent's antecedent beliefs as the conditioning of some maximally uninformed probability distribution pr on her background knowledge K, an informed probability distribution denoted by pr_K. In the wake of Carnap (1962) it is customary to posit a principle of evidential relevance along the following lines, for e, f, g three objects in the underlying algebra representing propositions or events.

Bayesian evidence principle. It occurs that f is neutral with respect to g in relation to e if $pr_K(g|f,e) = pr_K(g|e)$. Otherwise, f is evidence for (respectively, against) g in relation to e if

$$pr_K(g|f,e) > pr_K(g|e)$$

(respectively, if $pr_K(g|f,e) < pr_K(g|e)$).

Note that f is evidence for g in relation to e if

$$(1 - pr_K(f|e)) \times (pr_K(g|f,e) - pr_K(g|\neg f,e)) > 0.$$

On the face of it, the Bayesian evidential principle licenses an epistemic interpretation of Simpson's inequalities. Under this reading, any paradoxical reversal directly belies the following *theoretically* naive Bayesian view of evidential relevance as cumulative and uniform:

Tentative Bayesian evidence annihilation principle (proved wrong by any occurrence of Simpson's paradox). If f is evidence for g, both given e and $\neg e$, then f is evidence for g, *tout court*.

Contrary to the naive view of cumulative evidential import, assumptions about evidential relevance do not project systematically across epistemic partitions. At the very least, Simpson's paradox functions as a salutary and essential qualification for a system of unguarded evidential management.

In this connection, we remark that it is a moot point whether Simpson's paradox irreparably impugns the Bayesian construal of evidence, or whether the challenge can be obviated. Bayesians may argue that the problem does not concern the Bayesian account altogether, but only the specific concept of "confirmation as increase in firmness" (Carnap, 1962, corresponding to the above Bayesian evidence principle). Indeed, as shown by Fitelson (2017) after (Salmon, 1975), the rival notion of "confirmation as firmness" is immune to Simpson-like reversals. Nevertheless increase in firmness is indeniably an important concept of evidential relevance and, arguably, the most important one in the Bayesian perspective. Therefore, one cannot dismiss the paradoxical phenomenon as applied to evidence on the grounds that it is merely one more probabilistic vagary in a catalog replete with similar counterintuitive consequences, stemming from a too-close identification of evidential and probabilistic relevance.

5.3. Causal bearing

Consider now the third and paramount gloss on the Simpson Paradox inequalities, with causal concepts explicitly brought to bear on the interpretation. In this context, let us state a naive causal counterpart to the tentative comparative contrast and Bayesian evidence annihilation principles of Sections 5.1 and 5.2:

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

Tentative causal sure thing principle (proved wrong by some occurrences of Simpson's paradox). If Y causally promotes or inhibits X, with respect to a population partitioned by Z, then Y promotes or inhibits X, in the population as a whole.

Example A discussed in Section 3 is in accordance with the tentative principle. As Pearl has convincingly argued (Pearl, 2011, 2014, 2016), there is no mistaking the causal nature of the inviolable common-sense intuition that there can be no magical drug (Y) that is beneficial (X) to women (Z) and men ($\neg Z$) separately, but detrimental ($\neg X$) to the population as a whole.

However, a *do*-calculus Pearl (2000) analysis reveals that there is something wrong with the tentative principle. Indeed,

$$P(X|do(Y),Z) > P(X|do(\neg Y),Z) \quad \text{and} \quad P(X|do(Y),\neg Z) > P(X|do(\neg Y),\neg Z)$$

do not necessarily yield

$$P(X|do(Y)) > P(X|do(\neg Y)),$$

contrary to what the tentative principle claims. They do whenever the condition $P(Z|do(Y)) = P(Z|do(\neg Y))$ is met. This fact entails a valid counterpart to the above tentative principle:

Causal sure thing principle. If Y causally promotes or inhibits X, with respect to a population partitioned by Z, then Y promotes or inhibits X, in the population as a whole, provided that Y has no causal bearing on Z, according to our background knowledge.

Going back to the drug example, what guides the paradigmatically sensible reasoning is that we assume that drug has no effect on gender. Causal intervention to hand, Simpson's Paradox loses its sting.

In the next section, we reframe the causal sure thing principle in the context of decision theory, where the original sure thing principle was coined.

6. Sure thing principles

6.1. Savage's sure thing principle

Since the publication of Savage's *Foundations of Statistics* (Savage, 1972), the sure thing principle has attained to an exalted status. By its staunch admirers it has been deemed severally to be the cornerstone of Savage's theory (Joyce, 1999), a fundamental principle of human cognition (Pearl, 2000, 2016), and a unanimity-gathering "extralogical axiom" of rationality (Savage himself). Yet it has not gone entirely unchallenged, in both content and validity (Gibbard and Harper, 1978).

The first to raise the suspicion that Simpson's paradox might put pressure on the principle was the mathematician Colin Blyth. The ensuing discussion takes its cue from his classic presentation in (Blyth, 1972).

Savage states the principle as follows:

> If [a rational agent] would not prefer A to B, either knowing that the event E obtained, or knowing that the event $\neg E$ obtained, then he [ought] not to prefer A to B. Moreover (provided that he does not regard E as virtually impossible), if he would definitely prefer A to B knowing that E obtained, and if he would not prefer B to A, knowing that E did not obtain, then he will definitely prefer A to B.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

Savage's structural axioms on preference ranking guarantee that any action f can be decomposed over a logical partition into a mixture of conditional actions of the form *"do f if condition C is met"*. In the simplest case, where Z is an event of interest to the rational agent, a given action f is equivalent to an extended mixed action *"do f if Z and do f if $\neg Z$"*. This decomposition comes down to

$$f = f_Z \& f_{\neg Z}$$

where f_Z and $f_{\neg Z}$ are the actions *"do f if Z"* and *"do f if $\neg Z$"*. Leaving aside certain internal complications pertaining to virtually null events, it is helpful (and customary) to take a simpler restatement as the core idea:

Sure thing principle. If an agent would prefer both an option f_Z to g_Z, and an option $f_{\neg Z}$ to $g_{\neg Z}$, then she ought to prefer $f = f_Z \& f_{\neg Z}$ to $g = g_Z \& g_{\neg Z}$.

How does Simpson's paradox impinge on this eminently reasonable principle? For dialectical reasons we will continue to adapt the present considerations to Example B. Recall game I^+ introduced and discussed in Section 4.2. It consists in drawing a marble based on its size from the box where all marbles have been poured. Consider the following pair of *Blyth-actions* defined as follows:

— f, draw a large marble;
— g, draw a small marble.

Now apply Savage's conditional decomposition referencing the condition Z, *the marble comes from bag 1*, to both f and g respectively, to generate a fourfold proliferation of actions. Recapitulating the presentation of the data, it is straightforward to see that

1. f_Z dominates g_Z;
2. $f_{\neg Z}$ dominates $g_{\neg Z}$;
3. g dominates f.

But by the sure thing principle, 1, 2, together with the pair of conditional decompositions $f = f_Z \& f_{\neg Z}$, and $g = g_Z \& g_{\neg Z}$, the agent should find herself expecting that, in contradiction with 3,

4. f dominates g.

It is worth observing that the method by which the actions are constructed is entirely general, indeed *a-causal*, rendering the threat of invalidity ubiquitous in reach and range. Accepting the argument at face value, it would seem that Simpson's paradox invalidates the sure thing principle (systematically and a-causally), and consequently, that the sure thing principle must either be rejected or modified in such a way as to render it proof against Blyth's ingeniously constructed counterexamples. It is no surprise that systematic rejection, or else causal modification of the sure thing principle, is the most often encountered theoretical solution to the conundrum of the sure thing. In either case, a more or less radical departure from the original form of Savage's "extralogical axiom" is apparently enforced by Simpson's paradox.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

6.2. Pearl's sure thing principle

Pearl's causal modification of the sure thing principle is in part a concession to the force of Blyth's strategy for invalidating the principle. Let us rephrase the principle in decision-theoretic terms:

Causal sure thing principle (bis). Let f and g be two actions, and Z any event that is equally probable under f and under g, that is, $P(Z|do(f)) = P(Z|do(g))$. If a person prefers f to g, either knowing that the event Z obtains, or knowing that it does not, then she ought to prefer f to g if she knows nothing about Z.

As remarked, Blyth's method is uniformly general, exploiting the pattern of probabilistic dependencies characterizing Simpson's paradox. Interestingly, Pearl blocks the generation of counterexamples equally uniformly by exploiting the very same pattern. Specifically, the proposed causal restriction on Savage's sure thing principle allows for stochastic dependence between states (or events) and actions but requires that states be equiprobable under acting or intervening. The pair of actions defined by Blyth do not meet this condition, hence do not enter into the causally qualified form of sure thing reasoning. Likewise, in Example B where the agent engages in game I^+, since $P(\text{bag } 1|do(f)) = P(\text{bag } 1|\text{large})$ differs from $P(\text{bag } 1|do(g)) = P(\text{bag } 1|\text{small})$, actions f and g do not yield *equiprobable partitions under intervention* hence cannot be related by sure thing reasoning.

When, on the contrary, the equiprobability condition holds, the causal sure thing principle applies. A typical example of this situation is when Z stands for gender and f and g are two possible ways to treat a medical condition. If an agent prefers f over g both in the subpopulation of men and in that of women, then the causal version of the extended sure thing principle leads to the conclusion that she ought to prefer f to g when gender is not know. This clearly is perfectly right, whether or not the observational data exhibit a reversal of the Simpson kind. Moving from the initial, Savagian sure thing principle to the causal version supported by Pearl amounts to restricting the principle precisely to the cases where it is supported by causal analysis.

Finally, it is noteworthy that this line of analysis can be maintained when one considers not only preference but also indifference between actions (as in Bandyopadhayay, 1994, for instance). An unproblematic addition to the sure thing principle has it that if an agent is indifferent between f and g both knowing that Z obtains and knowing that it does not, then she ought to be indifferent between f and g if she knows nothing about Z. The causal version of this addition holds when Z is equiprobable under action f and action g, and this is unobjectionable.

7. Example C: Simpson's paradox in real life?

Simpson's paradox is not only a philosophical topic. It happens in reality with real data. Chuang et al. (2009) claim that they identify a real-life instance of Simpson's paradox in a series of synthetic microbial systems described in Example C. In this section, we analyze the case and argue that Simpson's paradox is in fact not at play here.

We do not pretend to do justice to the refinements of Chuang et al. (2009)'s biological experiment. For the sake of argumentation, it suffices to focus on a similar thought-experiment presented under the form of a causal graph in Figure 3.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

At time $t = 0$, for each π in, say, $\{20\%, 60\%\}$, a test-tube is prepared with $2N \times \pi$ producers and $2N \times (1-\pi)$ non-producers (see $(2N^p_{0a}, 2N^{np}_{0a})$ and $(2N^p_{0b}, 2N^{np}_{0b})$ in Figure 3). Immediately, at time $t = 0^+$, a third test-tube is prepared by mixing half of the above homogenized test-tubes (see (N^p_{0c}, N^{np}_{0c}) in Figure 3). By experimental design, the ratios $N^p_{0a}/(N^p_{0a}+N^{np}_{0a})$, $N^p_{0b}/(N^p_{0b}+N^{np}_{0b})$ and $N^p_{0c}/(N^p_{0c}+N^{np}_{0c})$ of producers in the three test-tubes equal 20%, 60% and 40%, respectively. The third test-tube is discarded. The contents of the two others are let to evolve independently but in similar experimental conditions from time $t = 0^+$ to time $t = 1$.

At time $t = 1$, the ratios of producers in the test-tubes are $N^p_{1a}/(N^p_{1a}+N^{np}_{1a})$ and $N^p_{1b}/(N^p_{1b}+N^{np}_{1b})$ (see Figure 3) and we observe that they are smaller than 20% and 60%, respectively, revealing that non-producers are advantaged and grow faster than producers in each test-tube. At time $t = 1^+$, a fourth test-tube is prepared by mixing the contents of the two others. Its ratio of producers is $N^p_{1d}/(N^p_{1d}+N^{np}_{1d})$ (see Figure 3).

It occurs that the ratio of producers in the fourth test-tube is larger than the ratio of producers in the third test-tube. The same conclusion holds when the experiment is replicated. The biological reason for this is that the content of the test-tube with the larger initial ratio of producers grows substantially more than the other one, sufficiently to counterbalance globally the local advantage of non-producers. This is nicely illustrated in (Chuang et al., 2009, Figure 1).

The observed reversal evokes the reversal of association between two variables when conditioning on a third one, which characterizes Simpson's paradox. However, the causal graph of Figure 3 suggests that this impression is misleading. Firstly, if there were a variable to condition upon, that would necessarily be a variable indicating which test-tube is considered. Whereas it is clear what means focusing on the first or second test-tube at times $t = 0^+$ and $t = 1$, it is unclear what means neglecting which test-tube is under consideration. In particular, the latter does not consist in focusing on either the third test-tube at time $t = 0^+$ or the fourth test-tube at time $t = 1^+$. Secondly, the fourth test-tube is not the by-product of the third test-tube in the same way as the first and second test-tube at time $t = 1$ are the by-products of themselves at time $t = 0^+$. Thirdly, contrary to the Examples A and B viewed as experiments, Example C viewed as an experiment is not the aggregation of two separate sub-experiments. In Example A, each realization of the experiment falls into one of two categories, depending on the stone diameter. In Example B viewed as a game, each realization of the experiment falls into one of two categories, depending on the bag from which the marble happens to be drawn. By contrast, in Example C, there is one single category of realization of the experiment, despite the fact that there is one sub-experiment for each initial ratio of producers and two additional mixing sub-experiments. Lastly, the objects of the reversed inequalities are fundamentally different in nature in Examples A and B on the one hand and in Example C on the other hand. In the former, the reversed inequalities concern measures of association, such as conditional probabilities like in (1), (2) and (3). They are features of the law of the experiment. In the latter, the reversed inequalities concern random variables, namely $N^p_{0a}/(N^p_{0a}+N^{np}_{0a})$, $N^p_{0b}/(N^p_{0b}+N^{np}_{0b})$, $N^p_{0c}/(N^p_{0c}+N^{np}_{0c})$ and $N^p_{1d}/(N^p_{1d}+N^{np}_{1d})$. They are by-products (realizations) of the law of the experiment, as opposed to features of it.

This shows that it can be difficult sometimes to decide whether a putative instance of Simpson's paradox is a real one or not. We think that the analysis of the paradox laid out in this article may be helpful in this regard.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

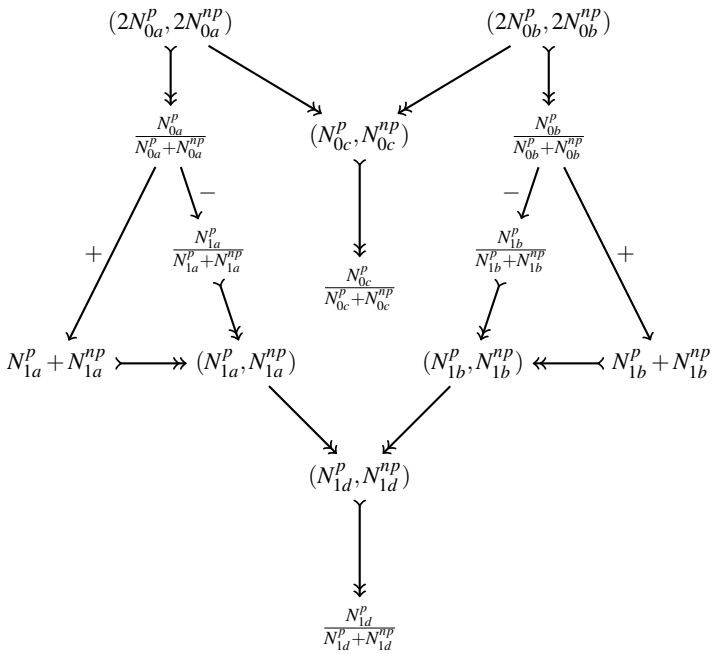

FIGURE 3. ***A causal graph representing a thought-experiment inspired by Chuang et al. (2009)'s experiment.*** *The variables of interest N^p_{tx} and N^{np}_{tx} ($(t,x) \in \{(0,a),(1,a),(0,b),(1,b),(0,c),(1,d)\}$) stand for numbers of producers and non-producers at the start ($t=0$) and end ($t=1$) of the experiment in four different settings. By experimental design, we have $(N^p_{0a}, N^{np}_{0a}) = 2N \times (20\%, 80\%)$, $(N^p_{0b}, N^{np}_{0b}) = 2N \times (60\%, 40\%)$ and $(N^p_{0c}, N^{np}_{0c}) = N \times (40\%, 60\%)$ where N is a reference number. There are several kinds of arrows. Two-headed arrows represent deterministic mechanisms, of which there are two kinds: two-headed arrows with tails are merely of algebraic nature (with no experimental counterpart) whereas two-headed arrows without tails model a controlled experimental process (such as mixing the content of two test-tubes). Single-headed arrows correspond to uncontrolled biological processes. The $+$ and $-$ signs distinguish between positive and negative associations.*

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

Acknowledgements.

The authors wish to thank warmly an anonymous reviewer for their careful reading and thoughtful comments.

8. Epilogue

Did Simpson come after all? Will he? Let us see how the third act unfolds.

8.1. Acte troisième, seconde partie

Estragon retourne ses chaussures, les secoue, en fait tomber caillous et fruits, petits et gros. Il se love, s'assoupit.

Vladimir fait le tour de l'arbre en agitant les bras et en murmurant indistinctement.

Entre à gauche le garçon de la veille. Il s'arrête. Silence.

GARÇON. – Monsieur... (*Vladimir se retourne.*) Monsieur Albert...
VLADIMIR. – Reprenons. (*Un temps. Au garçon.*) Tu ne me reconnais pas ?
GARÇON. – Non monsieur.
VLADIMIR. – C'est toi qui es venu hier ?
GARÇON. – Non monsieur.
VLADIMIR. – C'est la première fois que tu viens ?
GARÇON. – Oui monsieur.

Silence.

VLADIMIR. – C'est de la part de monsieur Simpson ?
GARÇON. – Oui monsieur.
VLADIMIR. – Il ne viendra pas ce soir ?
GARÇON. – Non monsieur.
VLADIMIR. – Mais il viendra demain.
GARÇON. – Oui monsieur.
VLADIMIR. – Sûrement.
GARÇON. – Oui monsieur.

Silence.

VLADIMIR. – Comment va ton frère ?
GARÇON. – Il va mieux, monsieur.
VLADIMIR. – Ah c'est bien ça.

Silence.

GARÇON. – Qu'est-ce que je dois dire à monsieur Simpson, monsieur ?
VLADIMIR. – Tu lui diras – (*il s'interrompt*) – tu lui diras que tu m'as vu et que – (*il réfléchit*) – que tu m'as vu et que je le remercie.
Silence. Soudain, le garçon se sauve comme une flèche. Silence. Le soleil se couche et la lune se lève. Estragon se réveille, se lève, va vers Vladimir, le regarde.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

ESTRAGON. – Qu'est-ce que tu as ?
VLADIMIR. – Je comprends mieux.
ESTRAGON. – C'est bien ?
VLADIMIR. – Je crois.

Silence.

ESTRAGON. – Il y avait longtemps que je dormais ?
VLADIMIR. – Je ne sais pas.

Silence.

ESTRAGON. – Moi je m'en vais.
VLADIMIR. – Moi aussi.

Silence.

ESTRAGON. – Où irons-nous ?
VLADIMIR. – Pas loin.
ESTRAGON. – Si si, allons-nous en loin d'ici !
VLADIMIR. – On ne peut pas.
ESTRAGON. – Pourquoi ?
VLADIMIR. – Il faut revenir demain.
ESTRAGON. – Pour quoi faire ?
VLADIMIR. – Attendre Simpson.
ESTRAGON. – C'est vrai. (*Un temps.*) Il n'est pas venu ?
VLADIMIR. – Non.
ESTRAGON. – Et maintenant il est trop tard.
VLADIMIR. – Oui, c'est la nuit.

Silence.

VLADIMIR. – Alors, on y va ?
ESTRAGON. – Allons-y.

Ils ne bougent pas.

RIDEAU

8.2. Third act, second part

Estragon turns over his shoes, shakes them, makes the pebbles and fruits, big and small, fall from them. He curls up, dozes off.

Vladimir walks around the tree, waving and mumbling.

Enter Boy left. He halts. Silence.

BOY. – Mister… (*Vladimir turns.*) Mister Albert…
VLADIMIR. – Off we go again. (*Pause.*) Do you not recognize me?
BOY. – No Sir.
VLADIMIR. – It wasn't you came yesterday.
BOY. – No Sir.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

VLADIMIR. – This is your first time?
BOY. – Yes Sir.

Silence.

VLADIMIR. – You have a message from Mister Simpson?
BOY. – Yes Sir.
VLADIMIR. – He won't come this evening.
BOY. – No Sir.
VLADIMIR. – But he'll come tomorrow.
BOY. – Yes Sir.
VLADIMIR. – Without fail.
BOY. – Yes Sir.

Silence.

VLADIMIR. – How is your brother?
BOY. – Better, Sir.
VLADIMIR. – Ah this is good.

Silence.

BOY. – What am I to tell Mister Simpson, Sir?
VLADIMIR. – Tell him – (*he hesitates*) – tell him you saw me – (*he ponders on*) – that you saw me and that I thank him.

Silence. Suddenly, the Boy exits running. Silence. The sun sets, the moon rises. Estragon wakes, gets up, goes towards Vladimir, looks at him.

ESTRAGON. – What's wrong with you?
VLADIMIR. – I understand better.
ESTRAGON. – Is this good?
VLADIMIR. – I think so.

Silence.

ESTRAGON. – Was I long asleep.
VLADIMIR. – I don't know.

Silence.

ESTRAGON. – I'm going.
VLADIMIR. – So am I.

Silence.

ESTRAGON. – Where shall we go?
VLADIMIR. – Not far.
ESTRAGON. – Oh yes, let's go far away from here.
VLADIMIR. – We can't.
ESTRAGON. – Why not?
VLADIMIR. – We have to come back tomorrow.
ESTRAGON. – What for?
VLADIMIR. – To wait for Simpson.
ESTRAGON. – Ah! (*Silence.*) He didn't come?
VLADIMIR. – No.
ESTRAGON. – And now it's too late.
VLADIMIR. – Yes, now it's night.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238

Silence.

VLADIMIR. – Well? Shall we go?
ESTRAGON. – Yes, let's go.

They do not move.

CURTAIN

References

Bandyoapdhyay, P. S., Nelson, D., Greenwood, M., Brittan, G., and Berwald, J. (2011). The logic of Simpson's paradox. *Synthese*, 181(2):185–208.

Bandyopadhayay, P. S. (1994). In search of a pointless decision principle. In *PSA: Proceedings of the biennial meeting of the philosophy of science association*, volume 1994, pages 260–269. Philosophy of Science Association.

Beckett, S. (1952). *En attendant Godot*. Editions de Minuit, Paris.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.*, 67:364–366, 373–381. With comments by D. V. Lindley, I. J. Good, Robert L. Winkler and John W. Pratt, and a rejoinder by Colin R. Blyth.

Caillois, R. (1958). *Les jeux et les hommes*. Editions Gallimard. Le masque et le vertige.

Carnap, R. (1962). *Logical foundations of probability*. University of Chicago press, Chicago.

Charig, C. R., Webb, D. R., Payne, S. R., and Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *BMJ (Clin Res Ed)*, 292(6524):879–882.

Chuang, J. S., Rivoire, O., and Leibler, S. (2009). Simpson's paradox in a synthetic microbial system. *Science*, 323(5911):272–275.

Cohen, M. R. and Nagel, E. (1934). An introduction to logic and the scientific method.

Fitelson, B. (2017). Confirmation, causation, and Simpson's paradox. *Episteme*, 14(3):297–309.

Gibbard, A. and Harper, W. L. (1978). Counterfactuals and two kinds of expected utility. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs. Conditionals, Belief, Decision, Chance and Time*, The Western Ontario Series in Philosophy of Science, pages 153–190. Springer.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge University Press, Cambridge.

Julious, S. A. and Mullee, M. A. (1994). Confounding and Simpson's paradox. *BMJ*, 309(6967):1480–1481.

Memetea, S. (2015). *Simpson's paradox in epistemology and decision theory*. PhD thesis, University of British Columbia.

Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge. Models, reasoning, and inference.

Pearl, J. (2011). Simpson's paradox: An anatomy. Technical report, Department of Statistics, UCLA.

Pearl, J. (2014). Comment: understanding Simpson's paradox. *The American Statistician*, 68(1):8–13.

Pearl, J. (2016). The Sure-Thing Principle. *Journal of Causal Inference*, 4(1):81–86.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195:1–405.

Quine, W. V. (1976). *The ways of paradox and other essays*, chapter The ways of paradox, pages 2–18. Harvard University Press, Cambridge, Massachusetts and London, England.

Salmon, W. C. (1975). Confirmation and relevance. *Minnesota Studies in the Philosophy of Science*, 6:3–36.

Savage, L. J. (1972). *The foundations of statistics*. Dover Publications, Inc., New York, revised edition.

Shpitser, I. (2019). Identification in causal models with hidden variables. *Journal de la Société Française de Statistique*.

Sloman, S. and Fernbach, P. (2017). *The knowledge illusion*. Riverhead Books, Penguin Random House, New-York. Why we never think alone.

Yule, G. U. (1900). On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194:257–319.

© Société Française de Statistique et Société Mathématique de France (2020) ISSN: 2102-6238