American Philosophical Quarterly
Volume 52, Number 1, January 2015

SIMPSON'S PARADOX AND CAUSALITY

Prasanta S. Bandyopadhyay, Mark Greenwood,
Don Wallace F. Dcruz, and Venkata Raghavan R.

> The skeptic about causality pushes the brake pedal to make his car slow, flips a switch to make a lamp glow, puts his money in the bank to collect interest.
>
> —Spirtes, Glymour, and Scheines 2001, p. 2

> [Physicists] continued to write equations in the office and talk cause-effect in the cafeteria. . . . [P]hysicists talk, write, and think one way and formulate physics in another.
>
> —Pearl 2009, pp. 407–408

ABSTRACT

There are three questions associated with Simpson's Paradox (SP): (i) Why is SP paradoxical? (ii) What conditions generate SP?, and (iii) What should be done about SP? By developing a logic-based account of SP, it is argued that (i) and (ii) must be divorced from (iii). This account shows that (i) and (ii) have nothing to do with causality, which plays a role only in addressing (iii). A counterexample is also presented against the causal account. Finally, the causal and logic-based approaches are compared by means of an experiment to show that SP is not basically causal.

OVERVIEW

Simpson's Paradox (SP) involves the reversal of the direction of a comparison or the cessation of an association when data from several sets are pooled. SP has wide applications in numerous disciplines.[1] Behind its applications and usefulness, several deeper issues have yet to be properly distinguished. Moreover, resolving one does not necessarily lead to the resolution of the rest. We will further argue that a conflation of those issues is in fact a factor in misreading the entire story about the paradox. Lately, however, it is almost conventional wisdom among scholars in this field to take the core of SP to be exclusively causal. In the words of Peter Spirtes, Clark Glymour, and Richard Scheines, "[t]he question is what *causal dependencies* can produce such a [case], and that question is properly known as 'Simpson's paradox'" (Spirtes, Glymour, and Scheines 2001, p. 40; emphasis added). Judea Pearl, for example, writes that "the spice of Simpson paradox has turned out to be non-statistical (i.e., causal) after all" (Pearl 2009, p. 177). In a very recent paper, epidemiologists Miguel Hernán, David Clayton, and Niels Keiding echo the same refrain:

> [Simpson's] paradox and error arise only when the problem is stripped of its causal context and analyzed merely in statistical terms, or when

> non-causal concepts like . . . collapsibility [are] allowed to guide the analysis. Once the casual goal is made explicit and causal considerations are incorporated into the analysis, the course of action becomes crystal clear. (2011, p. 784)

One purpose of this paper is to contest the conventional wisdom that traces SP to causality insofar as its central themes are concerned. Three questions need to be distinguished with regard to the paradox: (i) Why or in what sense, is SP a paradox?, (ii) What are the conditions for the emergence of this paradox?, and (iii) What should one do when confronted with a typical case of the paradox (to be called hereafter the "what-to- do" question).[2] We will argue that SP has to do with causality only if we ask the what-to-do question. For the sake of brevity, we will confine ourselves to the views of Spirtes, Glymour, and Scheines's causal account, often called the Carnegie Mellon University (CMU) theorists' account, as our rejoinder, if it is correct, is adequately general to be applicable to any other causal accounts of the paradox.[3]

The first section of the paper contains examples of Simpson's paradox. In the second section, we will provide a logic-based account that addresses the first two questions about the paradox. In the next section, we discuss Spirtes, Glymour, and Scheines's causal account of the paradox. In section 4, we provide a counter-example to the causal account. Section 5 is devoted to a comparison between our account and the causal account. There, we describe an experiment regarding Simpson's paradox and discuss its bearing on choosing between these two accounts. Section 6 evaluates an objection to our account. We point out that the causal account, while addressing the what-to-do question of the paradox, fails to appreciate the significance of all three questions. Finally, we will plead for a peaceful co-existence for both causal and logic-based accounts for a better understanding of different features of the paradox.

1. SIMPSON'S PARADOX

Consider the two examples of the paradox shown in Table 1 and Table 2. Table 1 represents an example of a formulation of the paradox in which the association in the sub-populations (departments) with higher acceptance rate for females is *reversed* in the combined population, with overall higher acceptance rates for males. Table 2 is an example that shows the paradoxical effect when the association between "gender" and "acceptance rates" in the sub-populations *ceases* to exist in the combined population. Though the acceptance rates for females are higher in each department, in the combined population, those rates cease to be different.

Table 1. Simpson's paradox (type I)

Two Groups	Dept. 1		Dept. 2		Acceptance Rates		Overall Acceptance
	Accept	Reject	Accept	Reject	Dept. 1	Dept. 2	Rates
Females	180	20	100	200	90%	33%	56%
Males	480	120	10	90	80%	10%	70%

Table 2. Simpson's paradox (type II)

Two Groups	Dept. 1		Dept. 2		Acceptance Rates		Overall Acceptance
	Accept	Reject	Accept	Reject	Dept. 1	Dept. 2	Rates
Females	90	1,410	110	390	6%	22%	10%
Males	20	980	380	2,620	2%	12%	10%

2. OUR LOGIC-BASED ACCOUNT OF THE PARADOX[4]

We begin with an analysis of the paradox in response to question (ii) above. Consider two groups [A, B] taken to be mutually exclusive and jointly exhaustive. The overall rates for each group are [α, β] respectively. Each group is partitioned into categories [1, 2], and the rates within each partition are [A_1, A_2, B_1, B_2]. Let's assume that f_1 = the number of females accepted in D_1, F_1 = the total number of females applied to $D_{1,}$ m_1 = the number of males accepted in D_1, M_1 = the total number of males applied to D_1. Then $A_1 = f_1/F_1$, and $B_1 = m_1/M_1$. Defining f_2, F_2, m_2, and M_2 in a similar way, we get $A_2 = f_2/F_2$ and $B_2 = m_2/M_2$. Likewise, we could understand α and β as representing the overall rates for females and males, respectively. So the terms $\alpha = (f_1 + f_2)/(F_1 + F_2)$ and $\beta = (m_1 + m_2)/(M_1 + M_2)$. To help conceptualize these notations in terms of Table 1, we provide their corresponding numerical values. $A_1 = \frac{180}{200} = 90\%$, $A_2 = \frac{100}{300} = 33\%$, $B_1 = \frac{480}{600} = 80\%$, $B_2 = \frac{10}{100} = 10\%$, $\alpha, = \frac{280}{500} = 56\%$, and finally $\beta = \frac{490}{700} = 70\%$. Since α, β, A_1, A_2, $B_{1,}$ and B_2 are rates of some form, they will range between 0 and 1 inclusive. We further stipulate the following definitions:

$C_1 \equiv A_1 \geq B_1$
$C_2 \equiv A_2 \geq B_2$
$C_3 \equiv \beta \geq \alpha$
$C \equiv (C_1 \;\&\; C_2 \;\&\; C_3)$

In terms of Table 1, these definitions become C1: 90% ≥ 80%, C2: 33% ≥ 10%, C3: 70% ≥ 56% and thus C is satisfied. But C alone is not a sufficient condition for SP. We could have a case where $A_1 = B_1$, $A_2 = B_2$, and $\beta = \alpha$ resulting in no paradox, yet C being satisfied. Hence, we stipulate another definition:

$C_4 \equiv \theta > 0$,
where, $\theta = (A_1 - B_1) + (A_2 - B_2) + (\beta - \alpha)$

For the data in Table 1, θ equals 10% + 23% + 14%. Again, C_4 alone is not sufficient for SP since we could have a case where $A_1 > B_1$, $B_2 > A_2$, and $\beta > \alpha$ resulting in no paradox (C is violated) and yet C_4 being satisfied.[5] Hence, a situation is a case of SP if and only if:

(a) $C \equiv (C_1 \;\&\; C_2 \;\&\; C_3)$ and
(b) $C_4 \equiv \theta = \{(A_1\text{-}B_1) + (A_2 - B_2) + (\beta - \alpha)\} > 0$.

Each condition (a or b) is necessary, but they jointly constitute sufficient conditions for generating SP (Bandyopadhyay et al. 2011).

There are four points worth mentioning. First, Clark Glymour (2004) would call our account an application of the "Socratic method" in which we provide necessary and sufficient conditions for the analysis of a concept.[6] Second, the characterization of the puzzle in terms of our two conditions captures the paradoxical nature of the data in the examples given, namely the reversal or the cessation of an association in the overall population; they are in no way ad hoc. Third, the paradox is "structural" in character, in the sense that the reasoning that leads to it is deductive. (Consider our examples, which involve simple arithmetic. The overall rates of acceptance for both females and males follow from their rates of acceptance in two departments taken separately.) Note that both conditions of the paradox can be defined in terms of the probability theory, which is purely deductive (Blyth 1972). Fourth, unless someone uses the notion of causation trivially, for example, believes that 2 + 2 "causes" 4, there is no reason to assume that there are causal intuitions lurking in the background. We will return to the last point in greater detail in the following sections.

To answer the first question (i), we now provide an explanation of how the paradox arises in people's minds and why it is found perplexing. The "how the paradox arises in people's minds" part is different from the conditions for emergence of SP given above. It is the reasoning process that subjects undergo that leads them to a paradoxical conclusion. For our purposes, we have reconstructed our type I version of SP in terms of its premises and conclusion to show how the paradox arises. However, the point of the reconstruction will be adequately general to be applicable to all types of SP. Before the reconstruction, we introduce a numerical principle called the collapsibility principle (CP), which plays a crucial role in the reconstruction. The CP says that relationships between variables that hold in the sub-populations (e.g., the rate of acceptance of females is higher than the rate of acceptance of males in both sub-populations) must hold in the overall population as well (i.e., the rate of acceptance of females must be higher than the rate of acceptance of males in the population). There are two versions of CP corresponding to the two types of SP. The first version of CP (CP1) says that a dataset is collapsible if and only if [(A1 > B1) & (A2 > B2) → (α > β)]. The second version of CP (CP2) states that a dataset is collapsible if and only if [(A1 = B1) & (A1 = B2) → (α = β)]. We will, however, find that both versions of the principle are, in fact, false with regard to the paradox. That is, CP → ~SP, whether it is CP1 or CP2, where "→" is to be construed as the implication sign.

Recall that A_1 and A_2 stand for the rates of acceptance for population A in departments 1 and 2, respectively. Similarly, B_1 and B_2 stand for the rates of acceptance for population B in departments 1 and 2, respectively. In contrast, α and β are rates of acceptance for A and B populations in the overall school. More explicitly, if we use our earlier notations of f_1, F_2, m_1, M_2, then CP1 implies

$$[(f_1/F_1) > (m_1/M_2) \ \&\ (f_2/F_2) > (m_2/M_2) \rightarrow \left({}^{f_1+f_2}\!\big/_{F_1+F_2} \right) > \left({}^{m_1+m_2}\!\big/_{M_1+M_2} \right)].$$

Likewise, CP2 says that

$$[(f_1/F_1) = (m_1/M_2) \ \&\ (f_2/F_2) = (m_2/M_2) \rightarrow \left({}^{f_1+f_2}\!\big/_{F_1+F_2} \right) > \left({}^{m_1+m_2}\!\big/_{M_1+M_2} \right)].$$

In the type I version of SP as in Table 1, CP1 becomes false. The same result can be obtained for the type II version of SP in Table 2 where CP2 will turn out to be false. As we can see, CP is a numerical inference principle devoid of any causal intuition. Here is the reconstruction of type I version of SP:

(1) Female and male populations are mutually exclusive and jointly exhaustive; one can't be a student of both departments and satisfy two conditions (a) and (b) in our characterization of what is called SP.

(2) The acceptance rate of females is higher than that of males in Department 1 (observed from data).

(3) The acceptance rate of females is higher than that of males in Department 2 (observed from data).

(4) If 2 and 3 are true, then the acceptance rate for females is higher than that of males overall (from CP1).

(5) Hence, the acceptance rate for females is higher than that of males overall (from 2, 3 and 4).

(6) However, fewer females are admitted overall (observed from data).

(7) Overall acceptance rate for females is both higher and lower than that of males (from 5 and 6).

In our derivation of the paradox, premise (4) plays a crucial role. In our type I version, the rates of acceptance for females are greater than those of males in each department. That is, $A_1 > B_1$ and $A_2 > B_2$, but α < β. Thus, CP1 becomes false. In fact, that CP1 is not generally true is shown by our derivation of a contradiction. The same kind of argument can be advanced to show that CP2 is also false. Our answer to the first question, (i), then, is simply that humans tend to invoke CP uncritically, as a rule of thumb, and thereby make mistakes in certain cases about proportions and ratios; they find it paradoxical when their usual expectation that CP is applicable across the board, turns out to be incorrect.

3. THE CAUSAL ACCOUNT OF SP

The CMU theorists have developed a subject matter-neutral automated causal inference engine that provides causal relationships among variables from observational data using information about their probabilistic correlations and assumptions about their causal structure. These assumptions are (1) the Causal Markov Condition (CMC), (2) the Faithfulness Condition (FC), and (3) the Causal Sufficiency Condition (CSC). According to CMC, a variable X is independent of every other variable (except X's effects) conditional on all of its direct causes. A is a direct cause of X if A exerts a causal influence on X that is not mediated by any other variables in a given graph. The FC says that all the conditional independencies in the graph are only implied by CMC, while CSC states that all common causes of measured variables are explicitly included in the model. Since these theorists are interested in teasing out reliable causal relationships from data, they would like to make sure that those probability distributions are faithful; otherwise they will not be able to derive causal relationships.

In Table 2, the dependency we observe between "gender" and "acceptance rate" (in the sub-population) gets cancelled out by their independence (from the overall population). In this case, the CMC alone imposes no constraints on the distributions that this structure could produce, since there is no independence whatsoever from using CMC. If there is an independence relation in the population that is not a consequence of the CMC, then the population, according to these causal theorists, is unfaithful. By assuming FC, they are able to eliminate all such cases of SP from consideration.

One reason for SP being causal, according to this account, is that (for the example given in Table 1) applying to the school is a causal problem involving causal dependencies between "gender" and "acceptance rates." More female students chose to apply to the departments where rates of acceptance are significantly lower, *causing* their overall rates of acceptance to be lower in the overall school. Similarly, with regard to Simpson's own example in the literature, Spirtes, Glymour, and Scheines write that "[t]he question is what *causal dependencies* can produce such a table, and that question is properly known as 'Simpson's paradox.'" (2001, p. 40). Therefore, Simpson's worry has a causal story, that is, the source of the paradox lies in its causal root.

Consider the following two tables to see what the CMU theorists mean. Table 3 is based on data for 80 patients. 40 patients were given treatment T and 40 assigned to a control, ~T. Patients either recovered, R, or didn't recover, ~R. There were two types of patients, males (M) and females (~M).

One would think that treatment is preferable to control in the combined statistics, whereas, given the statistics of the sub-groups, one gathers the impression that control is better for both types of patients. The what-to-do question is "Would one recommend the treatment to a patient, irrespective of his/her gender?" Spirtes, Glymour, and Scheines recommend control. Call this first example the medical example. In a second example (see Table 4), however, we are asked to consider the same data, but now regarding varieties of plants (white [W] or black variety [~W]), R and ~R as yields

Table 3. Simpson's paradox (medical example)

Two Groups	M		~M		Recovery Rates		Overall Recovery
	R	~R	R	~R	M	~M	Rates
T	18	12	2	8	60%	20%	50%
~T	7	3	9	21	70%	30%	40%

Table 4. Simpson's paradox (agricultural example)

Two Groups	T		~T		Yield Rates		Overall Yield
	Y	~Y	Y	~Y	T	~T	Rates
W	18	12	2	8	60%	20%	50%
~W	7	3	9	21	70%	30%	40%

(high[Y] or low yield [~Y]), and M and ~M as tall and short plants ([T] or [~T]).

Given Table 4, the overall yield rate suggests that planting the white variety is preferable since it is 10% better overall, although the white variety is 10% worse among both tall and short plants (sub-population statistics). Which statistics should one follow in choosing between which varieties to plant in the future? The CMU theorists' recommendation is that in this case, one should take the combined statistics (and thus plant the white variety), which is in stark contrast to the recommendation given in the medical case. In short, the causal theorists provide varying responses to the what-to-do question in the medical and agricultural examples. There is no *unique* response regarding which statistics, sub-population or whole, to follow in every case of SP.

Consider the "causal feature" in their causal account concerning the medical example. The novelty of their approach exploits the idea of intervention with regard to these cases. They construe "intervention" as something that directly controls manipulated variables in such a manner that makes them probabilistically independent of all their other causes when the rest of the casual structure remains intact. Thus, "gender" turns out not to be an effect of "treatment." When we "intervene" in the technical sense to impose a treatment on a new subject, gender and treatment not only are, but must be, probabilistically independent. The reason for treating "gender" and "treatment" to be independent is that when the subject's gender is not considered, the value of the variable "gender" has no effect on our choice, given the casual structure. Thus, in the medical example, Spirtes, Glymour, and Scheines recommend control. In the agricultural example, by contrast, the decision to plant a particular variety does not influence the genetic features in terms of both their association between height and color, and other possible effects on the plant. So whatever causal dependency there is between "height" and "color" in the sample will continue to exist in the population. Thus, recommending the combined population statistics, according to them, is the correct choice.

4. A COUNTER-EXAMPLE TO THE CAUSAL ACCOUNT

It is not easy to come up with an example that precludes invoking some sort of appeal to "causal intuitions" with regard to SP. But what follows is, we think, such a case. It tests in a crucial way the persuasiveness of the CMU theorists' account.[7] Suppose we have two bags of marbles, all of which are either big or small, and red or blue. Suppose in each bag, the proportion of big marbles that are red is greater than the portion of small marbles that are red. Now suppose we pour all the marbles from both bags into a box. Would we expect the portion of big marbles in the box that are red to be greater than the portion of small marbles in the box that are red? Most of us would be surprised to find that our usual expectation is incorrect. The big marbles in the first bag have a higher ratio of red to blue marbles than do the small marbles; the same is true about the ratio in the second bag. But considering all the marbles together, the small marbles have a higher ratio of reds to blues than the big marbles do. To help understand this scenario, consider Table 5.

Table 5. Simpson's paradox (marble example)

Marbles of Two Sizes	Bag 1		Bag. 2		Red Marbles rates		Overall Rates for Red Marbles
	Red	Blue	Red	Blue	Bag 1	Bag 2	
Big Marbles	180	20	100	200	90%	33%	56%
Small Marbles	480	120	10	90	80%	10%	70%

In Table 5, we find that in both bags, the ratio of red to blue marbles is greater for the big marbles than it is for the small ones (Bag 1: 90% > 80% and Bag 2: 33% > 10%). However, when all marbles are pooled together in one bag, the small marbles have a higher ratio of red to blue marbles than do the big marbles (in the combined bag: 70% > 56%). We argue that this is a case of SP since it has the same mathematical structure as the type I version of SP. There are no causal assumptions made in this example, no possible causal "confounding," and yet it seems surprising. That is the point of the counter-example. We believe this counter-example shows that at least sometimes, there is a purely mathematical mistake about ratios that people customarily make.

Some causal theorists might be tempted to contend that even in this marble example, there is confounding between the effects of the marble size on the color with the effects of the bag on the color. However, this confounding is not a causal confounding on which the causal account rests since one *cannot* say that Bag 1 *has caused* big marbles to become more likely to be red, or that Bag 2 has caused big marbles to become more likely to be blue. In short, one must admit that the above counter-example does not involve causal intuitions, yet it is still a case of SP.

It must also be admitted that there are all sorts of complexities about going from correlation to causation (see Russo 2009, for more on applications and methods of causality in social sciences). Correlations are not causes, though correlations are part of the evidence for causes. But what is paradoxical about SP has little to do with these complexities; there is simply a mistaken inference about correlations, which are really just ratios. Of course, when there are different correlations available that may seem to support conflicting causal inferences, the inference from correlations to cause becomes much more difficult; no one could reasonably deny that. We certainly admit that surprising facts about proportions come up frequently when we infer causes from proportions. This is when our mistakes about proportions seem most troubling to us. But the paradoxical nature of the examples really lies in the mistaken assumptions about the correlations (ratios) themselves.

5. COMPARISON

We contend that whether SP has anything to do with causality depends on which question (i, ii, or iii) we are asking. Although the first two questions are no doubt distinct, our formal reconstruction of the paradox provides a unified account of them, which empirical studies we have carried out both illustrate and amplify, We now discuss the results of such an experiment.[8] It involves presenting SP in a mathematical form as well as describing an example of SP in non-mathematical language, and getting student responses in both cases.[9] The non-mathematically explained case of SP is:

> There are only two high schools in a certain school district. Given that the graduation rate for girls in School #1 is higher than the graduation rate for boys in School #1, and that the graduation rate for girls in School #2 is higher than the graduation rate for boys in School #2. Does it follow that the graduation rate for girls *in the district* is higher than the graduation rate for boys *in the district*?

Which one of the following is true?

(a) Yes, the graduation rate for girls is greater than it is for boys *in the district*.
(b) No, the graduation rate for girls is less than it is for boys *in the district*.
(c) No; the graduation rates for girls and boys are equal *in the district*
(d) No inference could be made about the truth or falsity of the above because there is not enough information.

The mathematical case of the paradox is:

1. (f1/F1) > (m1/M1).
2. (f2/F2) > (m2/M2).
3. Does it follow that

$$\left(f_1 + f_2 \big/ F_1 + F_2\right) = \left(m_1 + m_2 \big/ M_1 + M_2\right)?$$

Which one of the following is true?

(a) Yes, the first expression is greater than the second.
(b) No, the first expression is less than the second.
(c) No, the first and second expressions are equal.
(d) No, inference could be made about the truth or falsity of the above because there is not enough information.

The correct answer to both questions is (d). Data were collected from 106 students. We found that for the non-mathematical question, students chose response (a) 83% of the time, which involves the mistaken use of the collapsibility principle, which is a non-causal numerical inference principle. They correctly responded choosing (d) only 12% of the time. For the mathematical question, 29% of the students picked the right answer (d), whereas they committed the error of choosing (a) 57% of the time. A test of the null hypothesis of "no difference in the rate of error" between the two versions of the questions produced evidence of a statistically significant difference in the error rates (P-value < 0.0001).[10] This just means that the two types of questions produce different error rates.[11] Similar surveys over many years of students in philosophy classes have manifested the same patterns of responses. The varying error rate in the two types of questions is clear evidence for the statistical difference between them without implying that this statistical difference has any deep philosophical bearing on our discussion, since a large number of students committed the same type of error by misapplying the collapsibility principle in both cases.

The math version of the paradox exactly mirrors our test case, which does not involve any causal intuition whatsoever. In turn, the math version also has a similar structure to the non-math version of our experiment involving the paradox. Consequently, it would be a mistake to think that the subjects' responses have exploited a causal intuition underlying different versions of the paradox; there is no difference between these two experiments, as they exhibit a similar mathematical structure. Most subjects mistakenly applied the non-causal principle CP.

Consider the CMU theorists' comment that Simpson's worry (whether to recommend "treatment" or "leave the patients untreated") has a causal story. They show clearly that the source of their recommendation about SP lies in their causal analysis, especially when, using intervention and other causal machinery, they recommend "control" in the medical example and recommend planting the white variety in the agricultural example, finding the same causal dependency between "height" and "color" both in sample and population. Two points need to be mentioned clearly here.

First, there is no point in denying that there are causal considerations involved in the medical and agricultural examples. They have no doubt contributed to our understanding regarding how to address the what-to-do question. Doing is almost always causing something to happen or to be the case. So to know what to do, we generally need to know how to cause something. In the agricultural example, if the decision is about which variety to plant to get the best yield, then it seems that causal issues settle the case. If we have no way of controlling how many tall plants are produced except by choosing whether to plant the white or black variety, and what we are interested in is the strategy that will produce the highest yield, then the causal story settles which yield statistics to pay attention to. However, if the decision question is instead about how to develop varieties that produce higher yields, then perhaps one would want to focus on the fact that the most important factor to work on is the size of the plants. So for that question, the sub-group statistics would be relevant.

The second point is about our assumption that the causal decision theory is correct. It is hard to see why one would recommend doing something that is merely correlated with a good result if there is no relevant cause underlying the correlation. Given a set of options constituting a decision situation, decision theory recommends an option that maximizes utility. It makes an appraisal of an option's utility by computing that option's expected utility. This account exploits probabilities and utilities of an option's possible outcome to calculate its expected utility. Here, the probabilities in question are dependent on the option. What is distinctive about causal decision theory is that it adopts the dependence of probabilities on the option to be causal rather than merely evidential (Joyce 1999; Levi 2000; Weirich 2012). Since the what-to-do question is a decision theoretic question and we agree with causal theorists that causal considerations settle the issue, we will assume the recommendations provided by causal decision theory to be correct with regard to both medical and agricultural examples.

Given what has been discussed so far about the causal theorists' stance toward the what-to-do question, one realizes that they have in fact addressed that question. We don't deny that causal inference plays a crucial role in addressing the what-to-do question. We do, however, argue that they fail to provide an adequate response to the first two questions. We have already provided a counter-example showing that SP has nothing to do with causality insofar as the first two questions are concerned. But SP still seems surprising because violation of the collapsibility principle leads to a "paradoxical" result. In addition to the fact that the CMU theorists have not provided an explanation for its surprising nature, they have not actually provided conditions for the paradox to arise—question (ii) about the paradox. In short, Spirtes, Glymour, and Scheines address the third question, but not the first two questions, thus failing to distinguish the three types of questions with regard to the paradox. In this way, our work and that done by the CMU theorists can be seen as complementing each other. They address the very important practical question of what-to-do in case of SP, and we provide an account of the paradox, illuminating its non-causal nature. Together, we are able to resolve some of the most pressing issues about SP. Table 6 summarizes how the two approaches have addressed three types of questions.

Table 6. Simpson's paradox and three types of questions

Approaches	Why Paradoxical?	What Conditions Needed for SP?	What to Do?
Causal (CMU)	No explanation provided	No specific conditions provided	Exploits the idea of intervention
Logic-Based	The failure of collapsibility principle	Two conditions provided	Agrees with the causal approach supplemented with causal decision theory

6. A POSSIBLE OBJECTION TO OUR ACCOUNT

One objection that has been raised recently against our account is that the real crux of the paradox lies in knowing *why* it has happened rather than *how* to recognize the paradox when it did happen.[12] According to this objection, causal theorists are interested in the deeper "why" question. The objector even contends that it is not hard to provide a causal story behind our counter-example.

If one were asked why the ratios of small red to small blue and large red to large blue marbles in the bag are what they are, one could provide plausible causal explanations. The manufacturing or packaging process might have favored this ratio, for example. The objector continues that perhaps blue marbles are made of more brittle materials and so they break and are defective more often (or the blue material is more expensive and so the manufacturer, desiring a good ratio of blue to red marbles, loads the bag with small blues and large reds, or so on). Moreover, we will always look for a causal account rather than rest content with a statistical anomaly.

People may make this causal assumption, but we must remember our original question: What makes the SP paradoxical? The above objection of the causal theorists changes the question. Instead of asking what makes SP paradoxical, it asks why we got the paradoxical data. We are not committed to denying that there is typically a causal story about how we came up with the data, although it could perfectly well be a coincidence. And indeed, we have investigated how evidence can be assessed for SP to be able to rule out its occurrence just by chance.[13] Hence, (1) our claim about the paradoxical nature of SP is independent of our ability to come up with plausible explanations, (2) we have gathered empirical evidence to support our claim that people extend the collapsibility principle across the board, and (3) we have demonstrated how the collapsibility principle in SP cases leads directly to contradiction.

7. CONCLUSION

In the first epigraph for this paper, as the CMU theorists write, "[t]he skeptic about causality pushes the brake pedal to make his car slow, flips a switch to make a lamp glow, puts his money in the bank to collect interest." In the same vein, Pearl notes that physicists have "continued to write equations in the office and talk cause-effect in the cafeteria. . . . Physicists talk, write, and think one way and formulate physics in another." In the light of the data presented above, how do we make sense of what they claim? Their underlying theme is that many skeptics' theoretical attitudes toward causality do not agree with their psychological attitudes toward it. Yet in a straightforward sense, our experimental data have shown that psychological attitudes concerning causality do not always enter into SP examples. In fact, generation of the paradox has nothing to do with causality, since the principle of collapsibility is non-causal. The role of causation in explaining what SP is and why it occurs could be left aside, even though causality remains important in resolving the most well-known instances of SP as well as the what-to-do question.

We have shown that Simpson's paradox can be generated in a straightforward deductive way. Among its premises is concealed a distinctly human dimension. In recent years, there has been a great deal of discussion of human frailty in connection with an individual's assessment of probabilistic statements (Kahneman, Slovic, and Tversky 1982; Kahneman 2011). Our resolution of the paradox has illuminated another aspect of human frailty. We explained its apparent paradoxical nature by invoking the failure of our widespread intuitions about numerical inference. The failure of collapsibility, which is non-causal, in Simpson's paradox-type cases is what makes them puzzling.

This explanation paints a human face onto the rather abstract structure of Simpson's paradox.

As George Berkeley once observed, philosophers "have first raised a dust and then complain that [they] cannot see" (Berkeley 1710/1970). Failing to see the relevance of the three types of questions is what we consider to be the dust that revolves around the paradox. Once the dust has settled, we notice that our experiment regarding the paradox has also brought a new flavor to doing "experimental philosophy," since it allows us to decide between two competing accounts of the paradox—causal and logic-based. However, strictly speaking, which account is the correct one depends entirely on which of the three questions we are asking. This oversight, we contend, is the root cause of the debate over the true nature of Simpson's paradox that the causal theorists have sorely missed.

From a different perspective, however, there is a reason to be grateful for a peaceful co-existence for both causal and logic-based accounts. We discussed the significance of the three questions in unlocking the riddle of Simpson's paradox. While the logic-based account illuminates the first two questions, the causal account addresses the most difficult, the what-to-do question. This paper pleads for combining different approaches together for a better understanding of the paradox. Only then can we achieve an increased understanding of its various aspects. Otherwise, like two groups of blind people touching an elephant and concluding that the part they have touched is the true nature of the elephant, we risk overlooking the overall picture that Simpson's paradox has three dimensions.

Montana State University
University of Hyderabad

NOTES

Several versions of this paper have been accepted for presentation in several places including Germany, Holland, India, and the United States. This version of the paper was presented at the American Philosophical Association, Eastern Divisional Meetings in 2012 and at Jadavpur University in 2013. We would like to thank the audiences in those places, including Prajit K. Basu, John G. Bennett, Sadhan Chakrabarty, Lopamudra Choudhury, Gordon Brittan, Jr., Abhijit Dasgupta, Dan Flory, Daniel Hausman, James Mattingly, Luke Rothschild, Indrani Sanyal, Madhucchandha Sen, Smita Sirker and Billy Smith for their helpful comments. We are also indebted to three anonymous referees of this journal for their helpful comments on earlier versions of the paper. We are very much indebted to John G. Bennett and Gordon Brittan, Jr., for numerous e-mail communications over the years regarding the paradox, to Davin Nelson for showing us how to think differently about the paradox, and to the editor of this journal for the inputs received during re-submission of the paper.

[1] Nancy Cartwright (1979) is the first philosopher to highlight the significance of Simpson's paradox to other philosophers.

[2] Daniel Hausman was perhaps the first philosopher who drew our attention to the significance of these three types of questions (in an e-mail communication). However, we are presumably the first to note their significance in Bandyopadhyay et al. (2011). Later, Pearl picked up this distinction (Pearl 2014).

[3] This does not, however, imply that there are no differences between the different casual accounts regarding Simpson's paradox. Pearl, for example, has developed a calculus of causality to handle causal cases, including Simpson's paradox. In addition, according to both him and some other causal theorists, although the collapsibility principle goes hand in hand with the paradox, it is not the central idea in

unlocking its riddle, as the principle is fundamentally non-causal. To contrast Pearl's account with that of the CMU theorists, the CMU theorists have proposed a constraint on observational data so that the data do not generate Simpson's paradox, whereas Pearl does not offer such a constraint. Consequently, the need for the collapsibility principle does not arise for their account. For more on the CMU theorists' view, see section 3. However, for our present purpose, what matters is the common assumption shared by both Pearl and the CMU theorists about the causal resolution of the paradox.

[4] We base this section on our earlier work (Bandyopadhyay et al. 2011).

[5] As a heuristic rule, we take A1 to be that sub-group ratio, which is the greater of the two ratios, and B1 as that which is lesser of the two. In Table 1, the ratio of women admitted to department 1 is greater than that of men. Hence, the former will be taken as A1, and the latter will be taken as B1. Similarly, since the ratio of women admitted to department 2 is greater than that of men, the former is taken as A2 and the latter as B2. This avoids the complexity of taking the absolute value of their difference in calculation of θ.

[6] Glymour (2004) contrasts this method with what he calls the "Euclidean" method-based theories where one could derive interesting consequences from them, although Euclidean method-based theories, according to him, are invariably incomplete. It is interesting to note two very different points. First, although Glymour is not fond of the Socratic method on which, however, a large part of Western philosophical tradition rests, our Socrates method-based logical account at the same time is also able to generate some interesting logical consequences (see Bandyopadhyay et al. 2011). Second, it is not only the Greeks who applied this kind of method. In classical Indian philosophical tradition, the Socratic method is also very much prevalent, where a definition of a term is evaluated in terms of whether it is able to escape from being both "too narrow" and "too wide."

[7] This counter-example is due to John G. Bennett.

[8] The suggestion to run an experiment is due to Caleb Galloway.

[9] Here, we are overlooking various subtleties involved in setting up those experiments. We offered versions with two different orders of questions. We did not want students to know what we were planning to test, nor did we want each student to know exactly what the student sitting next to him or her was doing. Often, many of the survey questions are irrelevant to the target questions. For example, we asked: "Is there life in Mars?" For fuller versions of those two experiments administered to the students during their class hours, please contact the authors.

[10] There are, however, objections to the use of significance tests. See Higgs (2013), who recommends curtailing the use of "significance" in statistical tests.

[11] Why the students provided a different type of response with regard to the math formulation of the paradox could be an interesting topic to speculate. Students may be frightened by mathematics, and when faced with (d) "no inference is possible," might consider it to be an easy alternative. However, this type of speculation goes beyond the scope of the paper.

[12] James Mattingly, our APA commentator, has suggested this way-out for causal theorists.

[13] Unpublished.

REFERENCES

Bandyopadhyay, Prasanta S., Davin Nelson, Mark Greenwood, Gordon Brittan, and Jesse Berwald. 2011. "The Logic of Simpson's Paradox," *Synthese*, vol. 181, no. 2, pp. 185–208.

Berkeley, George. 1710. *A Treatise concerning the Principles of Human Knowledge*, ed. Colin Murray Turbayne (Repr., Indianapolis: Bobbs-Merrill, 1970).

Blyth, Colin R. 1972. "On Simpson's Paradox and the Sure-Thing Principle," *Journal of the American Statistical Association*, vol. 67, no. 338, pp. 364–366.

Cartwright, Nancy. 1979. "Causal Laws and Effective Strategies," *Noûs*, vol. 13, no. 4, pp. 419–437.

Glymour, Clark. 2004. "Critical Notice," Review of *Making Things Happen: A Theory of Causal Explanation*, by James Woodward, *British Journal for the Philosophy of Science*, vol. 55, no. 4, pp. 779–790.

Hernán, Miguel A., David Clayton, and Niels Keiding. 2011. "The Simpson's Paradox Unraveled," *International Journal of Epidemiology*, vol. 40, no. 3, pp. 780–785.

Higgs, Megan D. 2013. "Do We Really Need the S-Word?," *American Scientist*, vol. 101, no. 1, pp. 6–9.

Joyce, James M. 1999. *Foundations of Causal Decision Theory* (Cambridge, UK: Cambridge University Press).

Kahneman, Daniel. 2011. *Thinking Fast and Slow* (New York: Farrar, Straus and Giroux).

Kahneman, Daniel, Paul Slovic, and Amos Tversky, eds. 1982. *Judgment under Uncertainty: Heuristics and Basics* (Cambridge, UK: Cambridge University Press).

Levi, Isaac. 2000. "Review Essay on the *Foundations of Causal Decision Theory* by James Joyce," *Journal of Philosophy*, vol. 97, no. 7, pp. 387–402.

Pearl, Judea. 2009. *Causality* (Cambridge, UK: Cambridge University Press).

———. 2014. "Comment: Understanding Simpson's Paradox," *American Statistician*, vol. 68, no. 1, pp. 8–13.

Russo, Federica. 2009. *Causality and Causal Modelling in the Social Sciences* (Dordrecht: Springer).

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search* (2nd edition) (Cambridge, MA: MIT Press).

Weirich, Paul. 2012. "Causal Decision Theory," in *The Stanford Encyclopedia of Philosophy* (Winter 2012 edition), ed. Edward N. Zalta. http://plato.stanford.edu/entries/decision-causal/.