SIMPSON'S PARADOX: A LOGICALLY BENIGN, EMPIRICALLY TREACHEROUS HYDRA

1. Introduction

If the term 'paradox' is understood to refer to arguments which have premises that are taken to be true that entail conclusions which are false, then Simpson's Paradox is mislabled. On a broader understanding of the term 'paradox', a set of sentences which appear to be collectively incompatible can count as paradoxical if the incompatibility is only apparent. Simpson's Paradox belongs to this second category. The statistician G. U. Yule is credited with first pointing it out in 1903; it was introduced into the philosophical literature by M. R. Cohen and E. Nagel in 1934, and it was the topic of a brief, witty, article by the statistician E. H. Simpson in 1951.[1] Cohen and Nagel used it to set a problem as an exercise; Nancy Cartwright [1979] and Brian Skyrms [1980] resurrected it from philosophical dormancy.[2] The paradox has been alleged to provide counter-examples to argument forms which are valid in the propositional calculus and counter-examples to the Sure Thing Principle of decision theory.[3] The alleged counter-examples are spurious, however, and the paradox is benign for valid inference and rational choice. Nevertheless, the basis of the paradox poses genuine problems for inferences from data to probability assignments to hypotheses, for models of causal inference, and for probabilistic analyses of causation. These problems persist when actual and possible empirical set-ups that manifest paradoxical structures are analysed. Coming to grips with them can help to explain the otherwise perplexing features of such set-ups.

2. An example of the paradox[4]

Suppose that a new drug is under test to determine whether it provides an effective treatment for an illness. In order to find out whether it is effective,

"Simpson's Paradox: A Logically Benign, Empirically Treacherous Hydra" by Gary Malinas,
The Monist, vol. 84, no. 2, pp. 265–283. Copyright © 2001, THE MONIST, Peru, Illinois 61354.

a percentage of patients are treated with the drug and a control group is given a placebo. When the results of the trial are tabulated, the drug appears to be an effective treatment. Fifty-four percent of the treated patients recover and only 44% of the patients who were given placebos recover. Now suppose that testosterone is one of the components of the new drug, and the question arises whether the drug is more effective for males or more effective for females or whether its effects are independent of gender. When the populations of treated and untreated patients are partitioned by gender, however, it turns out that the recovery rates for both males and females who are given placebos are higher than the recovery rates of those who were given the new drug. E.g., it is consistent with the recovery rates for the total population that 33% of the untreated males recover and only 27% of the treated males recover, and that 66% of the untreated females recover and only 64% of the treated females recover. So the drug appears to be effective when the total population is taken into account, but it does not appear to be effective for the male members of the population and it does not appear to be effective for the female members of the population. The following tables verify these relationships.

Total Population

	Recover	Do not Recover
Received Treatment	105	90
Did not Receive Treatment	40	50

Males

	Recover	Do not Recover
Received Treatment	15	40
Did not Receive Treatment	20	40

Females

	Recover	Do not Recover
Received Treatment	90	50
Did not Receive Treatment	20	10

3. Percentages and probabilities

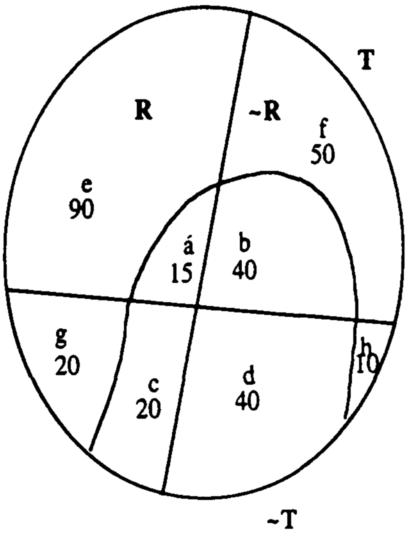
The example used to illustrate Simpson's Paradox was given in terms of patients, treatments, genders, and recovery rates. An urn model can be provided which has the same arithmetical properties as the example, but instead of patients there are balls in the urn, and each ball is inscribed with three symbols, one from each of the sets {T, ~T}, {R, ~R}, {M, ~M}. E.g., a given ball might carry the inscription [R, ~T, M]. The distribution of inscriptions on the balls is stipulated to conform to the tables above. Now the percentages of distributions of inscriptions on the balls can be represented as probabilities; the expression 'Prob(R/T)' is read as 'The probability that a ball is inscribed with an R given that it is inscribed with a T'. The model exhibits the following probability relationships:

Prob(R/T) > Prob(R/~T)
Prob(R/TM) < Prob(R/~TM)
Prob(R/T~M) < Prob(R/~T~M)

The feeling that these inequalities are paradoxical is assuaged when it is recalled that probabilities can be represented as weighted averages. E.g.,

Prob(R/T)=Prob(R/TM)Prob(M/T)+Prob(R/T~M)Prob(~M/T)

If the weights are sufficiently skewed, as they are in the above tables, a reversal of probability relations holds in the subsets of R under the further partitions of T and ~T by M and ~M. The following diagram illustrates the set-up described by the tables above. The set of M's is represented by {a,b,c,d} and the ~M's by {e,f,g,h}.



The diagram illustrates the consistency of the triad

$a/a+b < c/c+d$, and
$e/e+f < g/g+h$, but
$a+e/a+b+e+f > c+g/c+d+g+h$

where the last inequality represents the consolidated data for the total population in the example.

4. Boundary conditions for Simpson set-ups

The inequalities in the above example are preserved when the data are uniformly multiplied by any positive number. That suffices to show that there are infinitely many such set-ups. More generally, it is possible to have $\mathrm{Prob}(A/B) \approx$ zero and $\mathrm{Prob}(A/{\sim}B) \approx 1/n$, with $n = 1$, and

$\mathrm{Prob}(A/BC) \geq n\mathrm{Prob}(A/{\sim}BC)$, and
$\mathrm{Prob}(A/B{\sim}C) \geq n\mathrm{Prob}(A/{\sim}B{\sim}C)$.

Hence, it is not only inequalities that can be reversed in repartitions of a sample space, but also equalities can be perturbed when data are repartitioned or consolidated.[5] This latter fact has bearing upon the reliability of causal inferences which aim to "screen off" spurious correlations by locating common causes for them. This is discussed below in Section 6.2.

The diagnosis of Simpson's Paradox in its probabilistic form showed that the reversal effects were due to skewed weights. Such skewing can be arithmetically countered by normalising the data which represent proportions before they are used to represent percentages or probabilities. The effect of normalising data is to provide constant denominators for the ratios which are used to represent percentages and to compute weights in the representation of probabilities as weighted averages. In the example of the trial for the new drug, there is a difference in the percentages of patients who received the placebos and those who received the drug amongst both the males and females. The aim of normalising the data on treatments is to set up one-to-one correlations between the representation of treated males and untreated males, and similarly for females. Normalising data to provide constant denominators is a sufficient condition for the alignment of inequalities which are exhibited by consolidated data and are represented in 2×2 tables to agree with the alignments of inequalities which are exhibited by the 2×2×2 tables from which the 2×2 tables are derived.[6] For example, when the data from the drug trial are normalised on treatments, the drug does not appear to be effective for males, for females, nor for the total population. However, while normalising data may be sufficient to block the reversals which characterise the paradoxical cases, it does not always lead to the right conclusion as to what the actual probability relationships are. This is illustrated by the fact that there is more than one partition which "cross grains" the partition of recoveries by treatments in the example, and different patterns of probability relations will be exhibited by subsets of the data under different repartitions. Again, for example, keeping the figures for treatments and recoveries constant, a repartition by age might show positive correlations between recoveries for patients who are treated and are under fifty years of age, and similarly for those who are fifty years of age or older. Normalising the data on treatments can then show a positive correlation between treatments and recoveries for those under fifty, those fifty or over, and for the combined populations. The following tables, based on the same figures for recoveries in the total population, illustrates this possibility. Let 'U' represent the property of being under 50 years of age, and '~U' the property of being 50 or older.

	RU	~RU	R~U	~R~U
T	20	15	85	75
~T	10	20	30	30

Here, treatment appears more favourable for recovery in both members of the partition {U, ~U}. If we "normalise" the tables to put the T's in one-one correspondence with the ~T's, we obtain the following tables:

	RU	~RU	R~U	~R~U
T	20	15	85	75
~T	12	23	80	80

Again, treatment appears favourable for recovery in both tables and the combined table taken from them. While normalising data is sufficient to prevent reversals of relations between percentages and probabilities when data are consolidated, it is insufficient for deciding what the correct relations are. Normalised data from different partitions of the same raw data can imply incompatible conclusions.

5. *Two spurious problems*

It has been mooted that Simpson's Paradox provides counter-examples to classically valid arguments and to the Sure Thing Principle in decision theory. While the alleged counter-examples are fallacious, they do illustrate how easy it is to fall into the traps which paradoxical data facilitate.

5.1. Arguments of the following form are valid in the propositional calculus. Premises: If p then r. If q then r. Conclusion: If p or q then r. Now consider the following dictionary.

p = A male patient takes the drug.
q = A female patient takes the drug.
r = Taking the drug is less favourable for recovery than not taking the drug.

It can appear on the basis of the tables and probabilities given in the examples of the drug trial and the urn model that this dictionary provides a counter-model to the PC-valid argument form.[7] A closer look at the structure of the alleged counter-model and the data dispels this appearance.

A plausible reading of the suggested counter-argument assigns it the following form with (1) and (2) true, but (3) false.

(1) If x is male, then Prob(Rx/Tx) < Prob(Rx/~Tx)
(2) If x is female, then Prob(Rx/Tx) < Prob(Rx/~Tx)
(3) If x is male or x is female, then Prob(Rx/Tx) < Prob(Rx/~Tx)

Is this, taken in conjunction with the tables provided, an instance of and a counter-model to the form that is classically valid? No. That it is not a counter-model is apparent when the probability values from the model are explicitly provided. The ratios to which the conditional probabilities are equivalent can be represented as percentages. Then we have the following:

If x is male, then .27 < .33.
If x is female, then .64 < .66.
But, if x is male or x is female, then .54 > .44.

The contents of the consequents are determined by the sets selected by the restrictive clauses of the antecedents and the rules governing the function Prob(../—). The propositional variable *r* does not have a univocal interpretation under the readings proposed for 'Taking the drug is less favourable for recovery than not taking the drug'. As *r* does not have a univocal interpretation on the proposed reading, the informal formulation of the argument is not an instance of the valid-argument form.

Nonetheless, the argument does *appear* to be an instance of a valid form, and for many, it is intuitively *surprising* that the premises can be true and the conclusion can be false. A possible explanation for this is that within a broad, though limited, range of cases, the probability relations posited in the premises do entail a like alignment of the probability relations in the conclusion where data are consolidated. E.g., as noted in Section 4, when data are normalised, reversals of probability relations cannot occur. The case where data are "normal" is a special instance of a more inclusive class of arithmetical constraints on data which, if satisfied, are sufficient for the preservation of probability relations when data are consolidated.[8] If data that feature in set-ups where probabilistic inferences are intuitively drawn often or typically do fall within those arithmetical constraints, the inferences from probability relations that are supported by data from elements of partitions to like-probability relations when data are consolidated in conclusions will be truth-preserving (for those cases). In these cases, the weights that feature in the representations of the relevant conditional probabilities as weighted averages are not sufficiently skewed to reverse the probability alignments when data are consolidated. The general case, however, includes the cases that fall within the boundary conditions for reversals of probability relations when data are consolidated or repartitioned. This explains the intuition that the data from the example support a counter-model to a valid argument form by taking the intuition to be based upon a disposition to over-generalise from cases where the weights do not perturb probability relations to the many cases where they do perturb them. Of course, it is an empirical matter whether the correct explanation of the apparent counter-example is due to an over-generalisation or some other quirk of intuitive reasoning in which useful but rough heuristic rules can lead to untoward conclusions.[9]

5.2. The Sure Thing Principle (hereafter, STP) asserts that

> If you would definitely prefer *g* to *f*, either knowing that the event C obtained, or knowing that the event C did not obtain, then you definitely prefer *g* to *f*.[10]

Now consider the urn model described above and the following two-player zero-sum game. Players select one of two options. Player 1 goes first and makes choices on the basis of STP, if it is applicable. Player 2 is required to take whichever option remains open. Balls are returned to the urn after they are drawn. The options are as follows:

> **Option 1:** Draw balls at random from the urn until you get one that contains a ~T. Bet one unit that it contains an R.
>
> **Option 2:** Draw balls at random from the urn until you get one that contains a T. Bet one unit that it contains an R.
>
> Before you exercise either option, you are told whether the selected ball contains an M or a ~M.

On a given round of the game, if both players turn up balls which have an R, or which have a ~R, the round is cancelled and another round with bets in place is played. A player gains a win when his selected ball contains an R and the other player's ball contains a ~R. The players' aim is to adopt the option that maximises their chances of drawing balls that are inscribed with R's. Player 1 reasons as follows. Suppose he is told a given ball has an M. Then Option 1 gives him a .33 chance of it having an R compared with Option 2's .27 chance of its having an R. Next, suppose he is told that a selected ball has a ~M. Then Option 1 gives him a .66 chance of it having an R compared with Option 2's .64 chance of it having an R. Accordingly, STP appears to apply, and Player 1 selects Option 1. Player 2 is thus required to take Option 2. However, STP appears to give Player 1 bad advice. Fifty-four percent of the balls inscribed with a T are inscribed with an R, and only 44% of those inscribed with a ~T are inscribed with an R. Playing Option 1, Player 1 is more likely to have his R's matched by Player 2, thereby cancelling the round, and is less likely to match Player 2's R's, thereby losing the round. This remains the case despite the correlations of Rs with Ms and with ~Ms. Has STP given Player 1 bad advice, or has he applied STP inadvisably?

For STP to be applicable to his preferences, Player 1 needs to prefer Option 1 given M and given ~M. His reasoning adopts these preferences on the basis of the probability relations

Prob(R/TM) < Prob(R/~TM)
Prob(R/T~M) < Prob(R/~T~M).

Do these rationally support a preference for Option 1 on being told that M, or ~M, in the setting of the game? No. It will help to see why the probability relations do not support these preferences if we consider a different set-up than the one the players actually occupy in which Player 1's reasoning would be sound. Then, this will be contrasted with a variant of the set-up the players actually occupy that is equivalent to it. The fallacy in Player 1's reasoning that leads him to adopt the preferences that are required for STP to be applicable will emerge from the contrast between these two set-ups.

Let balls from our model be placed in two urns. The first, Urn(M), has all and only balls that are inscribed with an M. The second, Urn(~M), has all and only balls that are inscribed with a ~M. The tables that describe this set-up merely relable the tables from the medical example in Section 2 where data are partitioned by gender.

	Urn(M)	
	R	**R**
T	15	40
~T	20	40

	Urn(~M)	
	R	**~R**
T	90	50
~T	20	10

Player's options and the criteria for winning and losing are unchanged. The information that M, or ~M, indicates the urn from which a ball originates. However, that information is not relevant to the players' choices. In this game, Option 1 (the ~T option) does dominate Option 2, and it has a positive expectation of showing a profit regardless of the urn from which a selected ball originates. The ratio of ~T's to R's is greater than the ratio of T's to R's in each urn. It is significant that this is not the set-up that the players actually occupy and it is a different game from the one that they are playing.

Next, consider a set-up where balls from our model are again sorted into two urns. The first, Urn(T), contains all and only balls inscribed with a T, and the second, Urn(~T), contains all and only balls inscribed with a ~T. Criteria for winning and losing are unchanged. Players' options are to play the game with balls drawn from Urn(~T), option 1, or from urn(T), option 2. This game is equivalent to the one the players actually are playing. In it, 54% of the balls in Urn(T) have R's, and 44% of the balls in Urn(~T) have R's. Unlike the game where urns are homogenous with respect to M's and ~M's, the urns in this set-up have a mixture of M's and ~M's. The relevance of that mixture is displayed by the representations of Prob(R/T) and Prob(R/~T) as weighted averages. The information that a ball has an M, or a ~M, does not perturb these ratios, and they are different ratios than those that feature in the set-up with Urn(M) and Urn(~M).

The error in Player 1's reasoning was to suppose that the information concerning M's prior to each play of the game placed him in a situation analogous to the game where drawings are from Urn(M) and urn(~M). The contents of these two urns preserve the skewed distributions that support a dominance argument for Option 1. In that game, STP applies, and it gives good advice. In the game where drawings are from Urn(T) and Urn(~T), however, there is no sound dominance argument for Option 1 given M or given ~M. The mixtures in the redistribution correspond to different ratios of R's to T's and ~T's than the ratios that feature in the game with Urn(M) and Urn(~M). These ratios are not perturbed by the "news" that a selected ball has an M (or ~M). So, this set-up does not support a preference for Option 1 on being told that M or that ~M. STP is not applicable because rational players will not have the preferences that its applicability requires, i.e., in the game where balls are drawn from Urn(T) and Urn(~T), a preference for Option 1 over Option 2 given M, and given ~M.

6. *Statistical inference, causal inference, and probabilistic analyses of causation.*

Data sets which have the structure of Simpson's paradox have turned up in actual studies and experiments in the empirical sciences, in accountancy, in legal cases, and even in negotiations for salaries for baseball players in which their batting averages are relevant to the salaries they get.[11] E.g., one player had a higher batting average than another in each of two years, but the latter had a higher batting average over the combined two-year period. Such data sets, as well merely hypothetical data sets which share their structure, are relevant to testing theories of statistical inference, theories of causal inference, and analyses of causation which crucially rely on probability relationships.

6.1. Statistical inference. Theories of statistical inference aim to formulate rules for drawing inferences from data about the frequencies of kinds of events or attributes to conclusions concerning their probabilities. A core issue for such theories is to determine which reference classes support such inferences and which do not support them. In the example of the drug trial, the negative associations between the rates of recovery for treated males and for treated females would seem to support an inference that the probability of recovery for a patient is higher if he or she is left untreated. It was noted above that a partition of patients by age group could suggest the opposite conclusion. Alternatively, if recovery rates for patients were probabilistically independent of both age and gender, and the association between treatments and recoveries was otherwise resilient, the appropriate frequency data to use as basis for inference would be the positive association between treatments and recoveries. This is brought out by the urn model described above. There, the association between R's and T's is independent of any other letters inscribed on the balls. The probability of a ball which is inscribed with a T also being inscribed with an R is uniquely fixed by the ratio of balls inscribed with [TR] and [~TR]. This suggests the following constraint for the plausibility of statistical inferences from data to probability assignments: Inferences from proportions exhibited by data concerning A's and B's to conditional probability assignments which correspond to those proportions, e.g., Prob(A/B), are plausible with respect to a set of factors F, only if Prob(A/B) = Prob(A/B&Fi), for all Fi in F. This condition is met by the urn example concerning distributions of letters on balls; there is insufficient data in the example from the drug trial to tell whether or not it is met with respect to gender and/or with respect to the patients' ages. The underlying problem of which reference class or classes to use as a basis for inference from data to probability-assignments persists even if the partition by age *and* gender is taken as a relevant alternative to either alone. Taking partitions which are fine-grained or "maximally specific" as a basis for inference from data to probability assignments is no less secure from error than relying on partitions which are too coarse-grained and which mask relevant information. When a reversal of probability relations under one partition of a body of data is not matched by a different partition of the data (the typical case) the question arises as to whether to take the weighted average across the data to determine probabilities or whether to normalise the data for purposes of inference, thereby blocking the reversal. In the urn model described above, taking the weighted average gives the right answer for finding the probability of an R on a ball given that it is inscribed with a T. In other cases, normalising the data gives the right answer. Whether it does will often turn on contingent background information which sometimes is, and often is not, available to inquirers.

6.2. Causal inference. When kinds of events or properties are probabilistically relevant to each other, we have *prima facie* reasons to think that they are causally relevant to each other. Simpson's Paradox serves as a piquant reminder of how difficult it can be to reliably infer and work out causal connections from probabilistic relationships. One example of this difficulty is provided by Reichenbach's attempt to identify common causes of events which are correlated but are causally independent of each other.[12] His proposal is that correlated events are causally independent of each other provided that there is some event which "screens off" the correlation. Assume that B is positively probabilistically relevant to A, i.e., Prob(A/B) > Prob(A/~B). He says that C "screens off" B from A and is a common cause of them both just in case Prob(A/BC) = Prob(A/~BC). Recall that conditional probabilities can be represented as weighted averages. Accordingly, if the weights, Prob(B/C) and Prob(~B/C), are appropriately skewed, the screening-off condition can be fulfilled without C serving as a common cause of A and B. To find an appropriate skewing, it is sufficient to provide models which satisfy the following formulae:

Let terms take only positive values in the interval [0, 1]. Let Prob(A/BC) = x, and Prob(A/~BC) = y. On Reichenbach's proposal,

C screens off A from B (and B from A, as independence is symmetrical) if and only if Prob(A/B) > Prob(A/~B) and $x = y$.

Representing Prob(A/B) and Prob(A/~B) as weighted averages, the screening-off condition has the following form:

$$x(a) + b(c) > y(d) + e(f), \text{ and } x = y.$$

The probabilistic constraints on the values for {a,b,c,d,e,f} are insufficient to guarantee uniqueness for the screening-off factor C. E.g., C's obtaining could be necessary and sufficient for D's obtaining so that C screens off A from B if and only if D does so as well. This allows for spurious screening-off conditions or, unlikely, for multiple common causes. It also has the consequence that from an arithmetical perspective, any positive correlation between A and B can be associated with some condition C which "screens off" A from B. This follows form the boundary conditions described above. The moral to draw is not that screening off is trivial or arbitrary, but that one needs to specify which conditions which fulfil the screening-off condition will, if they exist, count as relevant to establishing causal independence. The screening-off condition requires supplementation if it is to be taken as sufficient for identifying common causes of correlations of causally independent events or attributes.

6.3. Probabilistic analyses of causation. Reichenbach's screening-off condition is a special case of more general analyses of causation in terms of probabilistic relations. The intuition which such analyses of causation share is that causes increase the probability of their effects. Simpson's Paradox poses a problem for this basic intuition since it shows that probabilistic relevance can be due to the effects of averaging in which causal information is lost. For example, if smokers attempt to counter the cardio-vascular effects of their habit by taking regular exercise, and non-smokers are blasé about exercising, there can be a negative correlation between smoking and heart disease, even if the incidence of heart disease is greater amongst the smokers who exercise compared with the non-smokers who exercise, and similarly for the smokers who do not exercise compared with the non-smokers who do not exercise. Such examples suggest that probabilistic relevance to background factors can better serve as the basis for extrapolating causal relevance. E.g., B is causally relevant to A if and only if Prob(A/B and Fi) is greater than Prob(A/~B and Fi) for all relevant Fi. However, this proposal falls afoul in cases like that described by the urn model above. Suppose that a ball's being inscribed with an M or a ~M is taken to be the only potentially relevant background factor for the probabilities of its being inscribed with other symbols. Then we have the following:

$$\text{Prob}(R/TM) < \text{Prob}(R/{\sim}TM)$$
$$\text{Prob}(R/T{\sim}M) < \text{Prob}(R/{\sim}T{\sim}M)$$

But it would be an error to infer either that ~T or T is causally relevant to R, or that Prob(R/~T) > Prob(R/T). Even if further constraints concerning the discreteness of causes and effects and their temporal order are provided to supplement the probabilistic relations, examples which have the structure of the urn model can be provided to meet these as well. Sometimes the correlations in the cells of relevant partitions are the correct basis for inferring causal relations, and sometimes they are spurious correlations which should be ignored. It remains an outstanding problem for probabilistic analyses of causation to formulate supplementary conditions on probabilistic relations for the purpose of using the latter to infer causal connections.

7. *Simpson's Paradox in dynamic settings*

In the previous section I noted that paradoxical set-ups have occurred in a wide range of actual settings. Further, I noted that when they do occur, causal relationships can be masked by probability relationships and other correlations. For some cases, when these are disentangled, phenomena that are deeply puzzling or that even seem impossible (but for the fact that they occur) can become transparent. An example of phenomena that have puzzled biologists as well as ethicists is the occurrence of altruistic behaviours in species. It is a matter of definition that altruistic behaviour disadvantages the individuals who engage in it while others reap its benefits. How, then, could it become a stable trait of the behaviour of a group in a set-up that evolves across time where the course of evolution punishes organisms that are less fit than those with whom they compete? Even if altruistic behaviour did emerge, wouldn't it be an unstable characteristic of the group that exhibited it? Wouldn't they, or their behaviour, be exploited by competitors and then driven to extinction or near-extinction? Simpson set-ups provide models that allow for the evolution of altruism. Moreover, if some further conditions are imposed on those set-ups, altruism can become a stable feature of them. Elliott Sober describes a simplified model for the evolution of altruism in a biological setting. Once such models are at hand, it is a small matter to reinterpret them as models of possible social systems, e.g., economic or political systems. Mapping social phenomena onto such models or variants of them might be useful for explaining some social phenomena; also, they may be useful for institutional design.

Sober describes a highly simplified and extreme case for illustrative purposes.[13] Assume a total system which consists of equally numerous selfish elements (S's) and altruistic elements (A's) that form two distinct groups with initially skewed distributions of the two kinds of elements. Suppose that the average fitness of a population of elements declines as the percentage of its selfish members increases. The decline is experienced by both the S's and the A's. The S's gain a benefit from their interactions with A's, and the A's absorb the costs of their interactions with S's. The S's gains and losses in their interactions with each other tend to cancel out, and as fewer A's are available to be exploited as the percentage of S's increases in a group, their fitness level declines as well. Let a single selfish element in a population of 99 A's have a fitness level of 4, and the 99 A's a fitness level of 3, where fitness is narrowly interpreted to represent the expected number of offspring in that arrangement. (Say the elements propagate uniparentally.) In an arrangement in which there are 99 S's and only 1 A, the S level of fitness is 2, and the A level of fitness is

1. As the percentage of S's increases, the level of fitness for A's declines continuously from 3 (with 1% S's) to 1 (with 99% S's), and there is a corresponding decline in the level of fitness for S's from 4 to 2. The average level of fitness for the total population declines from just over 3 to just under two as the ratio of S's to A's changes from 1:99 to 99:1. Now consider two groups, the first of which has just one S and 99 A's, and the second of which has one A and 99 S's. Let w represent the average fitness for elements in the two groups. The following table summarises this arrangement.

Group 1	**Group 2**	**Global average**
1S; $w = 4$	99S; $w = 2$	100S; $w = 2.02$
99A; $w = 3$	1A; $w = 1$	100A; $w = 2.98$

In this arrangement, the fitness level of S's is greater than that of A's in both groups, but it is lower than A's in the global average. Now suppose that parents die after reproducing, and that their reproduction exactly correlates with their levels of fitness. The census for offspring and their frequencies compared with frequencies for their parents are provided in the following table:

	Group 1	**Group 2**	**Global ensemble**
Parent Freq.	1%S; 99%A	99%S; 1%A	50%S; 50%A
Offspring Census	4S; 297A	198S; 1A	202S; 298A
Offspring Freq.	1.3%S; 98.7%A	99.5%S; .5%A	40%S; 60%A

After one reproductive cycle, the frequency of A's has declined in each group, but it has increased in the global ensemble. What will happen after a succession of reproductive cycles? The reproduction rule limits the lonely A in Group 2 to simply replacing itself while S's double with each generation. A's' hold on frequency declines, approximately halving, with each successive generation. The frequency of S's in Group 1 also rises with each successive generation, and this is bad news for the A's in Group 1. They are ineluctably driven to a fixation point where their population will stabilise at replacement while S's population continues to double in size with each generation. In this set-up, the early global bloom of altruism is nipped by the local and global winter of selfishness, never to bloom again.

In order for reversal effects due to an initial skewing of elements to be sustained in a system, a skewing comparable to the skewing of the initial set-up has to be sustained as the system evolves. In a system in which S's have an advantage over A's and they are equally distributed, S's will quickly become dominant and drive the A's to extinction or near extinction. However, if the system is structured so that there is an imbalance in the distribution and S's are clustered together with only a few A's amongst them, and A's are clustered together with few S's amongst them, it will be in the interest of the S's clustered with A's to keep other S's at bay, and this is an interest which they will share with the A's. Of course, even if they are successful at keeping other S's at bay, if their local numbers increase at a rate in excess of the increase for A's, as in the above tables for A's and S's, their comparative advantage over other S's will decline unless they redress the balance by expelling some of their numbers to form more clusters of S's. Alternately, S's might kill the offspring of other S's who are competitors in a bi-parental set-up, thereby preserving a skewing effect. A variety of possible mechanisms are able to preserve the skewing. Sober observes that "The groups must form new colonies rapidly enough to offset the within-group process that serves to displace the altruistic trait. Given favourable parameter values, altruism may come to exist at some stable intermediate frequency." He continues, "Also crucial is the question of *how* new colonies are established. If groups are founded by individuals who are alike, this will enhance and preserve intergroup variation and allow group selection to exert a more powerful influence on the advancement of altruism. If, on the other hand, migrants from different groups mix together and then found new groups, between-group variance will be diminished and the evolution of altruism will be more difficult."[14]

Biological examples provide natural settings for exploring the effects of reproduction and replacement. Economic set-ups in which there is competition also have counter-parts of extinction, reproduction, and replacement. Simpson set-ups might occur in them; independently of whether they are found to occur in them, it may be possible (and desirable) to design systems to exploit such possibilities and to secure stable pockets of altruism in a largely open-market system. However, if such set-ups can promote altruism despite evolutionary pressure against it, they have a structure that can also promote other (less desirable) traits against which there is evolutionary pressure, e.g., stupidity and ignorance. They also suggest how a minority political party might gain maximum effectiveness in competition with other parties.

It is an empirical question whether Simpson set-ups occur in nature or society, and if they do, whether they are dynamically stable. Inquiry into whether they occur in large biological, social, or economic systems is in its infancy. One difficulty such inquiries face is intrinsic to Simpson set-ups: any body of data which is rich enough to support inferences to probability assignments can be repartitioned so that relations between proportions in the cells of the partition are reversed when the data are consolidated. This trivial arithmetical fact poses some deep difficulties for empirical inquiry and it exacerbates some already familiar difficulties. When a reversal is observed under partitions of data, it is apt to ask whether they are an artefact of the arithmetic or a stable feature of the kinds which are represented by the partitions? At this juncture applied problems of selecting the right reference classes for statistical inferences, extrapolating causal relations from statistical data, and the classical problem of induction join hands to make the inquirer's job a high-risk occupation.

Gary Malinas

University of Queensland

NOTES

[1] G. U. Yule, "Notes on the Theory of Association of Attributes in Statistics," *Biometrika* 2 (1903) pp. 121–34. M. R. Cohen and E. Nagel, *An Introduction to Logic and Scientific Method* (New York: Harcourt, Brace, 1934). E. H. Simpson, "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society* B 13 (1951) pp. 238–41.

[2] Nancy Cartwright, "Causal Laws and Effective Strategies," *Noûs* 13 (1979) pp. 419–37. Brian Skyrms, *Causal Necessity* (New Haven, CT; Yale University Press, 1980).

[3] Judea Pearl raises the issue of whether Simpson's Paradox provides counter-examples to argument forms which are valid in the propositional calculus in *Probabilistic Reasoning in Intelligent Systems* (San Mateo, CA: Morgan Kaufman, 1988). C. R. Blyth takes paradoxical data sets to provide counter-examples to the Sure Thing Principle in "On Simpson's Paradox and the Sure-Thing Principle," *Journal of the American Statistical Association* 67 (1972) [Theory and Methods Section] pp. 364–66.

[4] This and the following section recycles the tables and diagram used in Gary Malinas, "Simpson's Paradox and the Wayward Researcher," *Australasian Journal of Philosophy* 75 (1997) pp. 343–59.

5. The boundary conditions for Simpson set-ups are given in C. R. Blyth, *op. cit.*, and Y. Mittal, "Homogeneity of Subpopulations and Simpson's Paradox," *Journal of the American Statistical Association* 86 (1991) [Theory and Methods Section] pp. 167–72.

6. Mittal, *op. cit.*, provides a proof that data which are normalised do not give rise to the kinds of reversals which characterise Simpson's Paradox.

7. C.F. Judea Pearl, cited in n.3, above.

8. C.f. Mittal, cited in n. 5, above.

9. There is the further possibility that the proposed dictionary does provide a counter-model to the PC-valid form, and the proposed reading and interpretation of the argument as a case of equivocation does not correctly represent its form. While my analysis of the proposed counter-model and the explanation of why it intuitively appears to be a counter-model does not preclude this possibility, it does shift the onus to the provision reasons that finesse the analysis.

10. L. J. Savage, *The Foundations of Statistics* (New York: John Wiley, 1954), pp. 21–22.

11. For examples, see S. Sunder, "Simpson's Reversal Paradox and Cost Allocations," *Journal of Accounting Research*, 21 (1983), 222–33. R. J. Thornton and J. T. Innes, "On Simpson's Paradox in Economic Statistics," *Oxford Bulletin of Economics and Statistics*, 47 (1985), 387–94. J. E. Cohen, "An Uncertainty Principal in Demography and the Unisex Issue," *The American Statistician*, 40 (1986), 32–39. References from Mittal, cited in n. 5, above.

12. Hans Reichenbach, *The Direction of Time* (Berkeley, CA: University of California Press, 1971), ch. 4.

13. Elliott Sober, *Philosophy of Biology* (Oxford: Oxford University Press, 1993), pp. 98–102. The tables below are recycled from Sober's discussion.

14. Elliott Sober, *The Nature of Selection* (Chicago: University of Chicago Press, 1993), p. 329.