Causal, Casual and Curious

Judea Pearl*

The Sure-Thing Principle

Abstract: In 1954, Jim Savage introduced the Sure Thing Principle to demonstrate that preferences among actions could constitute an axiomatic basis for a Bayesian foundation of statistical inference. Here, we trace the history of the principle, discuss some of its nuances, and evaluate its significance in the light of modern understanding of causal reasoning.

Keywords: JCI, Judea Pearl, UCLA

1 Introduction

The sure-thing principle (STP) was introduced by L.T. Savage [1] using the following story:

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

Illuminated by this story, the principle appears innocent and compelling, if not tautological. Savage duly recognized its universality and said: "I know of no other extralogical principle governing decisions that finds such ready acceptance." He then expressed it in more general terms as follows (slightly paraphrased):

Definition 1 (*The Sure-Thing Principle*)

"[*Let f and g be any two acts*], *if a person prefers f to g, either knowing that the event B obtained, or knowing that the event not-B obtained, then he should prefer f to g even if he knows nothing about B.*"

2 A principle and its philosophical roots

To appreciate the role that the STP plays in causal reasoning, the reader should note that the principle cannot be derived from classical logic, probability calculus or any other foundational formalism that does not deal explicitly with causation. This becomes clear when we recall that "acts" tend to combine differently from "events" or "propositions," hence they are not covered by logic or probability calculus. For example, if we were to replace preference for actions (e. g., "A person prefers acting f to acting g") with preference for events (e. g., "A person prefers finding f to finding g") than the STP is plainly false as we know only too well from Simpson's paradox [2, 3]. One can easily construct examples (and datasets) in which chances of success are higher for a person who smokes, either knowing that the person is a male or knowing that the person is a female, yet when the gender is unknown, chances of success are higher for those who do not smoke. It is much harder, and seems impossible, to imagine such reversal when f and g

*Corresponding author: Judea Pearl, Computer Science Department, University of California, Los Angeles, Los Angeles, CA, 90095–1596, USA, E-mail: judea@cs.ucla.edu

are deliberate actions, chosen by a rational agent, rather than passively observed events such as "finding that the person smokes." The former change the outcome, the latter merely predict it.

But Savage had much greater ambitions than to formulate a theory of actions; he was after an axiomatic theory for rational decisions based on subjective probabilities. This was seven years after the publication of the second edition of von Neumann and Morgenstern's (VNM) "Theory of Games and Economic Behavior" [4, 5], the book that established the axiomatic foundation for utility theory. Savage was not entirely satisfied with this foundation because VNM took probability theory for granted, even in cases where the uncertainties involved are subjective, and do not lend themselves to repeated observations. Savage, in contrast, sought to show that the rules of probability calculus are inevitable even in decisions involving subjective judgment of uncertainty. To accomplish this harder task, he needed a set of axioms that avoid probabilistic equalities or inequalities such as those invoked in VNM's axioms of utility theory.

To appreciate the enormity of the task, note that the sure-thing principle (Definition 1) is loaded with epistemic relationships such as "knowing that the event B obtained," or, "if he knows nothing about B." For students of probability theory these present no difficulties, because they are translated immediately into probabilistic expressions in which we either condition on B or do not condition on B. The rest is pure algebra. Savage, however, refused to take the appropriateness of the conditioning operator as God-given, and set out to prove that it is dictated by more compelling postulates.

Acknowledging this difficulty, Savage says: "The sure-thing principle cannot appropriately be accepted as a postulate ... because it would introduce new undefined technical terms referring to knowledge and possibility that would render it mathematically useless without still more postulates governing these terms. It will be preferable to regard the principle as a loose one that suggests certain formal postulates well articulated with P1 [the transitivity of preferences]" (ibid p. 22).

Accordingly, Savage abandoned the epistemic formulation of the sure-thing principle (Definition 1) and replaced it with one that is based on unconditional acts in multiple decisions problems, but avoids such notions as "knowing" or "not knowing."[1]

For most of us, who are willing to accept probability calculus as a normative standards for handling uncertainty, Definition 1 should pose no difficulty at all. All that needs to be done is to formalize the sure-thing principle in probability terms, take Bayes conditioning as the appropriate operator for updating knowledge, and investigate what else needs to be assumed to make the principle valid.

3 Easier said than done

The first demonstration that Savage's sure-thing principle may be invalid in some situations was presented by Colin R. Blyth [6], the Canadian-born mathematician who coined the name "Simpson's paradox." Blyth was able to contrive a sequential guessing game in which a strategy that violates the sure-thing principle yields a higher payoff than the one dictated by the principle. Blyth's construction (provided in Appendix) is rather intricate, and tends to hide the key element responsible for the failure of the sure-thing principle. Qualitative and more transparent counterexamples were constructed by the philosophers Gibbard and Harper [7] and Richard Jeffrey [8]. Jeffrey's example is particularly transparent, because it extends Savage's businessman story without resorting to numbers or to Simpson's reversal. Here it is:

> Change Savage's example to make the election be merely for the office of mayor, and suppose that the businessman thinks – perhaps correctly, and perhaps with excellent reason – that his buying the property would improve the Democratic contender's chances of winning.

Imagine that the businessman believes that the Democratic candidate, if elected mayor, would be a disaster to the city, regardless of whether he buys the property of not. Under such circumstances, it is quite

---

1 Samet [16] formulated the sure-thing principle in epistemic terms using a formal definition of "knowing."

reasonable that buying the property would be a good post-election deal, regardless of which candidate wins, yet a terrible pre-election deal, prior to knowing the winner, in blatant violation of the sure-thing principle.

It is not known whether Savage was aware of this obvious flaw in his formulation of the sure-thing principle because, in his presidential election example, actions were tacitly assumed incapable of affecting B, the election outcome.[2]

To repair this flaw, Jeffrey suggested a qualification of the sure-thing principle which restricts its application to cases where "states are independent of acts." In the businessman example, the sure-thing principle would be declared inapplicable if the state (i. e., the winner's identity) would be dependent on the act (buying the property).

Jeffrey's qualification is obviously an overkill. Adopting it would amount to excluding all cases in which states and acts are dependent, for example, studies in which individuals with certain characteristics (say sex or educational level) are more likely to seek treatment than other individuals, a common case in observational studies. As it turns out, the sure-thing principle can be reinstated by excluding only *causal dependence* of states on acts, allowing for *statistical dependence* between them, as when treatments are self-selected. This distinction, which cannot be expressed in the language of probability theory, was first noted by Gibbard and Harper [7] who applied Stalnaker [9] and Lewis [10] possible-world counterfactuals to decision theory and unveiled many of the features of *do*-calculus. Jeffrey could not allow for this distinction because he wanted to remain loyal to probability theory, and refused at all cost to acknowledge its limitations. This orthodoxy led him to develop "Evidential Decision Theory," a highly controversial school of philosophy in which acts are treated as events and Bayes' is the only conditioning operator allowed (see [11, 12], pp. 108–109)).

4 The principle in a causal setting

Adopting the causal-independence qualification, we can express the sure-thing principle succinctly, in *do*-calculus notation, and obtain its more refined formulation:

Definition 2 (*The Causal Sure-Thing Principle (CSTP)*)

Let f and g be two acts, and B any event that is equally probable under f and under g, that is,

$$P(B|do(f)) = P(B|do(g)). \tag{1}$$

If a person prefers f to g, either knowing that the event B obtained, or knowing that the event not-B obtained, then he ought to prefer f to g even if he knows nothing about B.

In Pearl ([12], p. 181) this version of the sure-thing principle was phrased in terms of increasing probabilities (of success) in each subpopulation, rather than preferences that hold knowing or not knowing a given event. It states:

Definition 3 (*Causal Sure-thing principle – Population Version*)

An action A that increases the probability of an event E in each subpopulation must also increase the probability of E in the population as a whole, provided only that the action does not change the distribution of the subpopulations.

---

2 We know however that he was aware of Simpson's paradox; Blyth [6] thanks Savage for telling him of Cohen and Nagel [17] where Simpson's reversal is first demonstrated. In a personal conversation (December 2015), Blyth recalled that Savage at first was concerned, but later denied that the game is a counterexample to his sure-thing principle. It is. As we demonstrate in Appendix.

The derivation is the same in both versions, and relies on the fact that the expected utility is a weighted average of the conditional utilities with *identical* weights under f and under g.[3]

5 The Will Rogers phenomenon

The restriction expressed in eq. (1) is probabilistic, not individualistic. This means that the STP permits the action to change individuals in each subpopulation, B and *not-B*, as long as it does not change the distributions of the subpopulations.

To demonstrate its ramification, consider an example known as "Will Rogers phenomenon" in which moving an element from one set to another set raises the average values of both sets. It is based on the following quote, attributed (perhaps incorrectly) to comedian Will Rogers:

> When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states.

The teasing implication behind the quote is that Oklahomans who moved were at the lowest intelligence scale in their state, and still above the intelligence level of an average Californian.

To see the relation to the sure-thing principle, let the immigration movement be the action considered, and let the subpopulations B and *not-B* stand for Oklahoma and California, respectfully. Clearly the action raised the average intelligence in both subpopulations. Yet, contrary to the causal sure-thing principle (Definition 3), it did not change the average intelligence in the combined states. The reason lies in violating the causal independence provision of eq. (1); the action changed the relative sizes of the two sub-populations, B and *not-B*.

A real-world example of the Will Rogers phenomenon is often detected in a medical process called *stage migration* [13]. In medical stage migration, improved detection of illness leads to the movement of people from the set of healthy people to the set of unhealthy people. This results in an apparent increase of life span for both groups, though there is no improvement in treatment for any.

6 Discussion

What can we learn from the sure-thing principle? Clearly, the utility of the principle does not lie in its inferential power. Savage acknowledged it and noted: "It is all too seldom that a decision can be arrived at on the basis of the principle used by this businessman, but ..." and here he identifies to the uniqueness and importance of the principle, "I know of no other extralogical principle governing decisions that finds such ready acceptance." Thus, the principle is unique in that it garners strong and universal acceptance despite the fact that it is not grounded in either logic or probability. Not many extralogical principles enjoy such combination of conviction and consensus.

By asking what kind of logic is needed to sanction the sure-thing principle we are asking in essence what kind of logic governs human thoughts, especially thoughts pertaining to decision making. Any logical system purporting to represent human thought must entail the sure-thing principle as a theorem. Relatedly, cognitive scientists who build robots to understand causal talk should ensure that the logic driving those robots sanctions the sure-thing principle.

What makes the sure-thing principle unique is its qualitative nature and the fact that it does not require any assumption beside the action-independence provision. By examining the vocabulary of this principle, and the conditions under which it is valid, we get a glimpse at how causal knowledge is stored in the mind.

---

3 Aumann et al. [18] have shown that the sure-thing principle is only valid when the subpopulations form a partition.

I made similar observations when asked to explain why people regard Simpson's reversal as a paradox, namely why they feel violated upon hearing about a drug that is bad for men, bad for women, and good for people [2]. Of course such a drug is a physical impossibility, but where do we get this deeply held conviction from, if it is not a theorem in standard logic?

The universality of this intuition suggests that a structure like causal Bayesian networks is used by people to represent and process causal knowledge. Aside from sanctioning the sure-thing principle, Bayesian networks also enjoy transparency and parsimony, two features that are not shared by other theories of actions, (e. g., possible-world counterfactuals [10] or potential outcome theory [14, 15]). We further know that causal Bayesian networks support the inferential machinery of the *do*-calculus. It is not surprising, therefore, that, with the exception of a few held back disciplines, graphical models became the standard carriers of causal knowledge.

Appendix

Blyth's counterexample

Take any dataset that supports Simpson's reversal, i. e.,

$$P(C|A) > P(C|\neg A), \tag{2}$$

$$P(C|A, B) < P(C|\neg A, B), \tag{3}$$

$$P(C|A, \neg B) < P(C|\neg A, \neg B). \tag{4}$$

Now consider the following guessing game. You have two options
f: Draw samples at random from the dataset until you get one for which *A* holds, and bet a dollar that *C* is true.
g: Draw samples at random from the dataset until you get one for which *A* does not hold and bet a dollar that *C* is true.
B: property *B* holds for the unit you bet on.

Given that *B* obtained, you would definitely prefer *g* to *f* because, by (3), *g* gives you higher probability of winning a dollar than what you get under *f*. Given that ¬*B* obtained, you would also prefer *g* to *f* because *g* gives you probability $P(C|A, \neg B)$ of winning a dollar which, by (4), is larger than $P(C|A, \neg B)$, what you get under *f*. But not knowing if *B* holds you should prefer *f* to *g* because *f* gives you probability $P(C|A)$ of winning a dollar, while *g* gives you only $P(C|\neg A)$ which by (2) is smaller.

This is an ingenious scheme of converting any data supporting Simpson's reversal into a decision situation in which the sure-thing principle is violated. Note however that any such game does not constitute a counterexample to the Causal Sure-Thing Principle (Definition 2); because it violates the independence provision:

$$P(B|do(f)) = P(B|do(g)).$$

This can be proven by realizing that in the game the equality $P(B|do(f)) = P(B|do(g))$ translates into $P(B|A) = P(B|\neg A)$ and, under the latter, we know that Simpson's reversal is impossible.

References

1. Savage L. The foundations of statistics. New York: John Wiley and Sons, Inc, 1954.
2. Pearl J. Understanding Simpson's paradox. Am Stat 2014;88:8–13.
3. Simpson E. The interpretation of interaction in contingency tables. J Royal Stat Soc Ser B 1951;13:238–41.
4. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton, NJ: Princeton University Press, 1944.
5. von Neumann J, Morgenstern O. Theory of games and economic behavior, 2nd ed. Princeton, NJ: Princeton University Press, 1947.
6. Blyth C. On Simpson's paradox and the sure-thing principle. J Am Stat Assoc 1972;67:364–6.
7. Gibbard A, Harper L. Counterfactuals and two kinds of expected utility. In: Harper WL, Stalnaker R, Pearce G, editors. Ifs. Dordrecht: D. Reidel, 1976:153–69, 1981.
8. Jeffrey R. The sure thing principle. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 2: Symposia and Invited Papers 719–730, 1982.
9. Stalnaker R. A theory of conditionals. In: Rescher N, editor. Studies in logical theory, vol. No. 2, American Philosophical Quarterly Monograph Series. Oxford: Blackwell, 1968:98–112. Reprinted in W.L. Harper, R. Stalnaker, and G. Pearce (Eds.), Ifs, Dordrecht: D. Reidel, 1981:41–55.
10. Lewis D. Counterfactuals. Cambridge, MA: Harvard University Press, 1973.
11. Jeffrey R. The logic of decisions. New York: McGraw-Hill, 1965.
12. Pearl J. Causality: models, reasoning, and inference, 2nd ed. New York: Cambridge University Press, 2009.
13. Feinstein A, Sosin D, Wells C. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. New Eng J Med 1985;312:1604–8, DOI:10.1056/NEJM198506203122504.PMID 4000199.
14. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci 1923;5:465–80.
15. Rubin D. Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol 1974;66:688–701.
16. Samet D. The sure-thing principle in epistemic terms. Tech. rep., The Leon Recanati Graduate School of Business Administration. Tel Aviv: Tel Aviv University, 2015.
17. Cohen M, Nagel E. An introduction to logic and the scientific method. New York: Harcourt, Brace and Company, 1934.
18. Aumann R, Hart S, Perry M, Conditioning and the sure-thing principle. Tech. rep. Center for the Study of Rationality and Institute of Mathematics, The Hebrew University of Jerusalem, 2006.