Probability Kinematics and Causality

Richard Jeffrey

Princeton University

Brian Skyrms's *Dynamics of Rational Deliberation* (1990) answers questions I have puzzled over for thirty years, and suggests fresh approaches to other questions. Here I want to explore an answer, in Skyrms's dynamic terms, to an old question on my agenda concerning Humean (or Kantian, or pragmatic) probabilistic accounts of causal talk:

> What features of your judgmental probabilities show that you view truth of one proposition as promoting truth of another, rather than as being promoted by truth of that other, or as being promoted by truth of some third proposition which also promotes truth of the other?

Innocence of causal ideas was meant to be a virtue of *The Logic of Decision* (1965), opening the possibility of a non-circular analysis of talk about causality; but in the wake of Robert Nozick's "Newcomb's problem and two principles of choice" (1969) that became problematical, for it seemed to limit the theory to "normal" cases, where any influence between acts and "states" (consequences, relevant conditions) runs from acts to states. Brian Skyrms (1980) was one of the earliest architects of a decision theory giving causal factors their own explicit place. The second edition of my book (1983, 1990) offered a way ("ratificationism") of dealing with some Newcomb problems, but not all. The present proposal grows out of my later attempt (1988) to make good the deficiency.

Skyrms's (1980, pp. 128-139) causal decision theory replaces the usual act-state matrices by 3-dimensional act-state-cause matrices whose components are the partitions you envisage of available acts, states, and causal hypotheses. The result is like a deck of cards, one for each causal hypothesis K, each card printed with an act-state matrix of conditional expectations of utility (sc., "desirabilities") E(u|S&K&A), and conditional probabilities P(S|K&A). To the card corresponding to any one K, Skyrms applies *The Logic of Decision* to get figures of merit for all acts A, i.e., your conditional expectation of utility given A on the assumption that K is the true causal hypothesis. These are the familiar weighted averages $E(u|K\&A) = \sum_S E(u|S\&K\&A)P(S|K\&A)$ of desirabilities, where the weights are conditional probabilities. Your overall figure of merit for act A is a weighted average of those K-relative figures, the weights being your unconditional

PSA 1992, Volume 2, pp. 365-373

probabilities for the K's: $U(A) = \Sigma_K P(K)E(u|K\&A)$, i.e., writing $f(K) = E(u|K\&A)$, $U(A) = E(f)$.

Skyrms's theory delivers what I take to be correct advice about how to behave in ("Newcomb") problems where you see your acts as mere signs of conditions that you would promote or prevent if you could. It is a static theory, in which one reasons in terms of a single probability function P throughout deliberation. Here I present a dynamic adaptation of my (1965) theory that delivers the goods in Newcomb problems (as well as normal decision problems). It depends on a certain interpretation of, or substitute for, causal talk. The underlying thought is that your directions of causal imputation correspond to no momentary features of your developing judgmental probability function P, but to features of your overall program for updating P. It is programmed constancy of conditional probability as unconditional probability of the condition changes that shows the program to be treating the condition as a source and the other propositions as sinks in the updating process. In normal decision problems acts A are sources. In Newcomb problems the sources are causal hypotheses K. In normal decision problems P(S|A) remains constant as your probability distribution over the A's changes, thereby driving the updating process. In Newcomb problems it is changes in your probability distribution over the K's that drives the process, in which the A's are sinks; changes in your views about what act you are likely to perform are driven by changes in your judgments about the causal hypotheses. (But these latter changes will depend epistemically on changes in your probabilistic judgment about what act you will perform.) Thus, in a prisoners's dilemma for "clones" (i.e., prisoners who think their values, beliefs, and thought processes are much alike), the prisoners may see themselves as having a common tendency toward cooperation or defection. Then Ms. Row's probability distribution over values of K may be expected to change with her probability for cooperating, seen as evidence bearing on the K's; but conditionally on any particular value of K, her probabilities P(state|act&K) = P(state|K) for her states (= Column's acts) are fixed as P develops, driven by changes in P(K).

	S	¬S		S	¬S	
A	3rd best	best	A	p	1-p	E(u|A) = p·(3rd best)+(1-p)·(best)
¬A	worst	2nd best	¬A	1-q	q	E(u|¬A) = (1-q)·(worst)+q·(2nd best)

Table 1 (a) Utilities (b) Conditional probabilities (c) Conditional expectations of utility

Table 1 is a partial representation in my 1965 framework of the general 2-by-2 Newcomb problem, abbreviating p = P(S|A) and q = P(¬S|¬A). (A full representation would identify the causal hypotheses K.) If preference goes by conditional expectation of utility ("desirability"), then by Table 1(c) the condition under which you prefer ¬A to A is that

$$p\cdot(\text{best}-\text{third}) + q\cdot(\text{second}-\text{worst}) > \text{best}-\text{worst}$$

Where the difference between best and third best equals the difference between second best and worst (as in the original Newcomb problem), the condition under which you prefer act 0 is

$$p + q > (\text{best}-\text{worst})/(\text{second}-\text{worst})$$

In the original Newcomb problem best, second, third, worst are 1001, 1000, 1, 0, respectively (i.e., utilities of receiving those numbers of thousands of dollars), so that you prefer ¬A iff p+q > 1.001. Then if you see the acts as having even a very slight predictive significance for the corresponding states, e.g., p = q = .501, you will choose act ¬A. Yet, the problem-statement strongly suggests that A is preferable to ¬A:

Someone with a fair record of accuracy in such matters has fallibly predicted your choice, and irreversibly deposited a million dollars in your bank account if and only if the prediction was that you would reject an additional thousand. There is no time-travel going on; just shrewd reading of character. As the question of the million in your bank account is already settled when you decide to accept (A) or reject (¬A) the extra thousand, it seems merely supersitious to refuse the $1000 sweetener.

This version of the problem, due to Howard Sobel, is equivalent to Nozick's, in which the thousand and the million are in two boxes, and accepting or rejecting the thousand is a matter of taking both boxes or only one.

There are some — e.g., Bar-Hillel and Maragalit (1972) — who think that taking just one box, declining the thousand, ¬A, is the reasonable move; and statistics would seem to bear them out. "If it is irrational to decline the extra thousand, I prefer the million to rational poverty." And it is sometimes held that *The Logic of Decision* resolves the prisoners's dilemma by recognizing the cooperative solution (¬A) as rational, since there is some reason to think each prisoner's decision a good enough predictor of the other's, e.g., when the utilities *best* = 0, *second* = -1, *third* = -9, *worst* = -10 are negatives of times served, so that conditional expectation of utility relative to act ¬A is maximum as long as $p + q > 10/9 = 1.111\ldots$, as when $p = q = .51$.

In normal decision problems you see your choice of an act as the source of changes in your probability distribution P, with alternative acts promoting or inhibiting or independent of various states. The case where you treat some third factor K as representing tendencies or *tropisms* toward particular acts and states is the defining characteristic of Newcomb problems. These are problems in which some tropism is seen as the source, a causal factor K driving both acts and states. This means:

In Newcomb problems, changes in your probability distribution over values of the tropism, X, drive the updating process, and probabilistic dependencies such as $P(S|A) < P(S|\neg A)$ have no causal significance, but are rooted in differences in the evidentiary significance of the acts for values X=x of the tropism, as when ¬A is better evidence than A for high x.

Thus, in the original Newcomb problem, when you and the predictor are responding to a common inclining cause of the act and the state, and ±S means that ±A is predicted, it will be the conditional probabilities $P(A\&S|X=x)$ that are programmed as constant while $P(X=x)$ varies. It is the mark of ordinary decision problems, where acts are seen as driving states, that $P(S|A)$ is meant to remain constant as $P(A)$ varies.

Now in Newcomb problems you see the the tropism as driving both act and state. In particular, your probability for the state conditionally on any particular value of the tropism, and conditionally on any particular pair of values for the tropism and the act, might simply be the value of the tropism:

$$P(S|A\&X=x) = P(S|X=x) = x$$

(In such cases you are treating x as the objective probability of S — the *objective chance*, or *propensity* that S is true.) Then your unconditional probability for S will be a weighted average of the various values x this conditional probability can assume, between 0 and 1, weighted with your unconditional probabilities for those values. As we are dealing with a continuous distribution of judgmental probability over the unit interval [0,1], it is natural to think of the matter in terms of a probability density function f assigning non-negative values f(x) to points t in the unit interval in such a way

https://doi.org/10.1017/S0270864700009449 Published online by Cambridge University Press

that there is unit mass under the f curve over [0,1]. The value x = P(S) defines the position of a fulcrum on which a uniform mass distribution under the f curve would just balance, i.e., $P(S) = \int_0^1 xf(x)dx$.

Before proceeding, let us examine a particular model of the dynamics of Newcomb problems, i.e., a pair of families f_t, g_t, of probability density functions over the unit interval, representing alternative courses of evolution of your judgment concerning the values of the tropism. As t increases from 0, the f_t family moves from a uniform distribution $f_0(x) = 1$ for all x in [0,1] through distributions concentrating probability more and more toward the right-hand end of the interval, i.e., except for normalization, $f_1(x) = x$, $f_2(x) = x^2$, $f_3(x) = x^3$, ... After normalization (so that the area under the f_1 curve from x=0 to x=1 becomes 1), these become $f_1(x) = (t+1)x^t$. The f_1 family represents one possible route from a judgmental state of total ignorance about where K lies in the unit interval through states of increasing conviction that the value is far to the right in that interval. Similarly, the g_t family represents a route from the same initial state of ignorance, $g_0(x) = f_0(x) = 1$, through distributions concentrating probability further and further to the left of the unit interval, i.e., except for normalization, densities $g_0(x) = 1\text{-}x$, $g_2(x) = (1\text{-}x)^2$, $g_3(x) = (1\text{-}x)^3$, ... In general, normalized, these are $g_t(x) = (t+1)(1\text{-}x)^t$.

Now in this problem your judgments about what act you will perform, and what the state of nature will be, are driven by your judgments about the tropism, as follows, where P_t and Q_t are your probability functions as determined by f_t and g_t, respectively, and E_t and F_t are the corresponding expectation functions. The computations $E_t(X) = \int_0^1 xf_t(x)dx$, $F_t(1-X)^2 = \int_0^1 (1\text{-}x)^2 g_t(x)dx$, etc., yield the following results.

$$P_t(A) = P_t(S) = E_t(X) = Q_t(\neg A) = Q_t(\neg S) = F_t(1\text{-}X) = (t+1)/(t+2)$$
$$P_t(A\&S) = E_t(X2) = Q_t(\neg S\&\neg A) = F_t(1\text{-}X) = (t+1)/(t+3)$$
$$P_t(S|A) = Q_t(\neg S|\neg A)) = (t+2)/(t+3)$$
$$P_t(\neg S|\neg A) = Q_t(S|A) = 2/(t+3)$$

Then as $P_t(A)$ approaches 1, driven by movement of the densities f_t, the conditional probability matrix in Table 1(b) goes through the forms shown in Table 2(a), while as $Q_t(A)$ approaches 0, driven by densities g_t, the matrix goes through the forms shown in Table 2(b). Note that in the two cases the odds on S conditionally on A and on ¬A are as in Tabes 3(a) and (b).

	S	¬S
A	(t+2)/(t+3)	1/(t+3)
¬A	(t+1)/(t+3)	2/(t+3)

	S	¬S
A	2/(t+3)	(t+1)/(t+3)
¬A	1/(t+3)	(t+2)/(t+3)

Table 2 (a) P(±S|±A) (b) Q(±S|±A)

	S:¬S
A	t+2:1
¬A	t+1:2

	S:¬S
A	2:t+1
¬A	1:t+2

	S	¬S
A	3rd best	best
¬A	worst	2nd best

Table 3 (a) P(S|±A):P(¬S|±A) (b) Q(S|±A):Q(¬S|±A) (c) Utilities

Evidently, as t increases without bound $P_t(S|\pm A)$ approaches 1 and $Q_t(S|\pm A)$ approaches 0. Then as you become nearly certain about your act, i.e., as $P_t(A)$ approaches 1 (left) or $Q_t(\neg A)$ approaches 1 (right), you become nearly certain about the corresponding state, i.e., in the original Newcomb problem, that that very act will have been predicted, or in the similar prisoners' dilemma that the other prisoner is performing that very act. The utility matrix (c) is unchanging and identical throughout the approaches (a, b) to the two decisions. On the way to a decision to make A true (a), you

https://doi.org/10.1017/S0270864700009449 Published online by Cambridge University Press

see yourself approaching the next-to-worst outcome — but you see the alternative as even worse. On the way to a decision to make A false (b) you see yourself as near the next-to-best outcome — but you see the alternative as even better! Then a decision to make A true is ratifiable in the sense that as you approach it, ¬A continues to look worse than A; but the more promising decision to make A false is unratifiable, for as you become surer that you will opt for ¬A, the payoff from that option looks worse than the payoff from the other.

Then if preference goes by conditional expectation of utility, A is available to you as a preferential choice: as you opt for it, that option looks preferable to the alternative. But ¬A would not be a preferential choice; you may opt for it, but in doing so you will be choosing the then less-preferred option. On a dynamic reading, *The Logic of Decision* delivers the 2-box solution to Newcomb's problem.

That's the case, anyway, on the particular model we have been considering, of your approaches to the two possible decisions, i.e., a plausible model for the problem as stated, far from being the only plausible model of it, but perhaps representative of all such models in delivering the 2-box solution. But what is sorely needed, here, is a cogent general characterization of the class of plausible models.

In variants of the problem, 1-box solutions are plausible (so I think) and are delivered by the present machinery. A prime example is the case where the predictor is thought to be 100% reliable. Then all your probability will be concentrated upon two tropisms, H_0 and H_1, where H_i = you will take i+1 boxes and the predictor expects you do so. Here $P(A\&S|H_i) = i$, and in Table 1(b), p=q=1 and the other two entries are 0, so that you do not see this as a decision under uncertainty; rather, you see ¬A (take 1) as surely better than A (take 2), and (according to the present machinery) are well advised to act accordingly, for as $E(u|\neg A) > E(u|A)$ all the way through your approach to either choice, that machinery identifies ¬A is your only preferential choice.

In Newcomb problems theoretical considerations about preëxisting tropisms drive the agent's practical judgment about what to do, and it is precisely this that puts them out of focus as decision problems — the defining characteristic being felt as a denial of normal free agency. The quest for believable Newcomb problems is the quest for believable hypotheses of the right sort about tropisms. Philosopers are often willing to accept schematized hypotheses, granting that talk about objective chance can be cashed out so as to satisfy the principle that one's judgmental probability for any proposition A is one's expectation of a random variable $X(w) = ch_w(A)$ — read, "the objective chance in world w that A is true" — having the characteristic that for each w, the function $ch_w(A)$ is a probability measure in the technical sense. Thus the condition is $P(A) = E(ch_w(A))$, so that in case $ch_w(A)$ has only a countable set of distinct values x, each having a judgmental probability, i.e., $P\{w:ch_w(A)=x\}$, the condition is

$$P(A) = \sum_x x \cdot P\{w:ch_w(A)=x\}$$

or, what comes to the same thing,

$$P(A|\{w:ch_w(A)=x\}) = x$$

This (especially in the second form) is what David Lewis (1986, p. 87) calls the (shudder) "principal principle" connecting the two concepts of judgmental probability (P) and objective chance (ch). Others, for the same reason, call it the principle of "direct inference" from objective chance to judgmental probability.

https://doi.org/10.1017/S0270864700009449 Published online by Cambridge University Press

But I prefer to replace this principle by a schema (or treat this principle as a schema) indicating the general conditions under which we treat natural random variables as sources of causal influence, conditions in which it is changes in judgmental probability distributions over the possible values of such variables that drive judgmental probability updating programs. I illustrate it here with a particular natural random variable of a sort often used in that way in textbook examples, i.e., the proportion p(w) of green balls in an urn, where A is the proposition that the next ball drawn will be green. The schema is asserted for all proportions, x.

$$P(A|\{w:p(w) = x\}) = x$$

Now this is no principle, if a principle is a requirement reasonably imposed in every case. Thus it would not be reasonable to apply this schema in case the drawing is made with eyes open, in good light, with view to drawing a green ball. Rather, this is a condition defining circumstances under which the updating process is driven by one's judgmental probability distribution over the possible proportions of green balls in the urn. The statement of these circumstances, which hold in some cases but not in all, explicitly refers to your judgmental probabilities P as well as to the objective conditions, $p(w) = x$. The schema does not prescribe updating schemes, but classifies them; given an updating scheme, the schema identifies the random variable that drives it.

In general, p is any random variable satisfying the schema for all of its values x. For the particular random variable in our example, p = proportion of green balls in the urn, the value of p was used as a conditional probability, $P(A|\{w:p(w) = x\}) = x$, but that is not required in general. Thus, suppose your judgmental probability that a penny spun on edge will land head up is a function h(o/r) of the ratio of the diameter of its obverse to that of its reverse — the penny being a section of a cone. In such a case, p(w) might be o(w)/r(w), i.e., not a probability, and the schema would be

$$P(A|\{w:p(w)=x\}) = h(x)$$

Knowing h, one could rewrite this in a form in which the value of the random variable is your judgmental probability itself:

$$P(A|\{w:h(p(w))=y\}) = y$$

But identification of such a function h is more likely to be the desired outcome of an incomplete project, so that the schema is a hopeful characterization of some future judgmental probability function. The project might involve extensive experimentation with coins having various values of p, with actual relative frequencies of heads for particular values x of p serving to identify the values y = h(x).

I think it important to mark the distinction between p as an actual accomplishment (Shafer 1981) and as a project. That loose distinction sits on top of a continuum — a multi-dimensional one at that — with extremes exemplified at the "accomplishment" end by textbook examples like the one in which p is the proportion of green balls in an urn, and at the project end by the EPR attitude toward quantum mechanics (Einstein&al. 1935) as demanding completion by a still unspecified naturalistic grounding for p. There is a useful comparison with the semantics of counterfactual conditionals — say, Stalnaker's (1968), which uses a selection function, f, where f(A, w) is w itself if A is true in w, and otherwise is the world most like w in which A is true. My view of f — like Stalnaker's, I think — is like my view of ch, i.e., a tribe of context-dependent notions. Similarity of worlds is what we make it. Thus, with w as the real world, if I do not now drop this glass, w′ = f("I drop this glass now", w) is undefined in the absence of a specification of the analog of an updating program for P — a program allowing determination in principle of what truths are to carry over from w to w′, and of the truth values in w′ of falsehoods in w. The truth value in w of "This glass would break now if I dropped it now" is the truth value in w′ of "This glass breaks now." Certainly; but that does not yet determine the conditional's truth value in w — even "in principle." To identify a particular selection function is to identify the features of each w that are to remain fixed when we go to f(w, A). The cases where counterfactual talk is unproblematical are those in which w is the actual world and tacit understandings identify the constant features of f relative to a tacitly understood class of A's. What we must guard against is the natural but bogus thought that since such understandings are familiar and plentiful, "f" already makes sense for arbitrary w's and A's, before we do the work of specifying f(w, A) for all w and A, i.e., work far beyond our power to accomplish — except in special, hothouse cases.

But of course the dynamic account of deliberation presented here is no less schematic than Stalnaker's account of counterfactual conditionals. Instead of prescribing appropriate dynamics for problems as presented, it requires deliberating agents to map the approaches to their various possible acts via updating schemes consonant with their understanding of those problems. What's provided is only a framework, to be filled in by the deliberating agent or whoever undertakes to analyze the agent's predicament. But the filling material is subjectivistic Bayesian stuff — not counterfactual conditionals, or objective chances, but programs for judgmental probabilities evolving over time, albeit programs for which the user's manuals might helpfully use the counterfactual idiom and speak of causal influence and objective chance, for it is in the context of such definite programs that those idioms are of use.

I conclude with assorted references to the literature.

As far as I know, it was Frank Arntzenius (1990) who first published the suggestion that constancy of conditional probability under varying probability of the condition is the missing ingredient in recent attempts to define probabilistic causality. He has called my attention to Huw Price's (1993) arguments against that idea, in this volume. (Note that Arntzenius's paper in *this* volume does not address the constancy question.) I am sympathetic to Price's (and, as he points out, Frank Ramsey's) view of causal asymmetry as rooted in our perspective as agents. I was delighted to find Arntzenius floating and defending in print the suggestion that invariance of conditional probabilities completes probabilistic analyses of causality; I had thought it might, but didn't feel well enough acquainted with the literature to stick my neck out. But whether or not it yields an analysis of homely notions of causality, I see invariance as the active ingredient in impingements of the notion of causal efficacy on decision theory, and I expect that Arntzenius's (1990) arguments will stand up as demonstrating a similar role for invariance pretty generally.

Like de Finetti (1931, etc.) I am a subjectivist. Thus, following Steven Leeds (1984) and Brian Skyrms (1984) I see the probabilities in quantum mechanics as subjective; I see the theory as telling you to accept certain constraints on your judgmental probability distribution, e.g., fixing certain conditional probabilities. In effect, the theory's sentences about probability are in the imperative mood; using the theory requires following those instructions. Bas van Fraassen (1989, pp. 197-205) pushes this point of view further, using Haim Gaifman's (1988) idiom, in which the theory is said to offer expert probabilistic opinion within its domain.

The present analysis provides a dynamic rationale for the scheme of probability kinematics floated in Jeffrey (1965 chapter 11; see also 1992 chapter 7). In the design of mechanisms, kinematics is the discipline in which rigid rods and distortionless wheels, gears, etc. are thought of as faithful, prompt communicators of motion. The contrasting dynamic analysis takes forces into account, so that, e.g., elasticity may introduce distortion, delay, and vibration; but kinematical analyses often suffice, or, anyway, suggest important dynamical considerations. That is the metaphor behind the term "probability kinematics," and behind use of the term "rigidity" for constancy of conditional probabilities.

References

Arntzenius, F.(1990), "Physics and Common Causes", *Synthese* 82: 77-96.

Bar-Hillel, M. and Margalit. A. (1972), "Newcomb's Problem Revisited", *British Journal for Philosophy of Science* 23: 295-304.

de Finetti, B. (1931), "Probabilismo", *Logos* 14, pp. 163-219, Napoli. Translation: "Probabilism", *Erkenntnis* 31: 169-223, 1989.

Einstein, A., Podolsky, B. and Rosen, N. (1935), "Can Quantum-mechanical Description of Physical Reality be Considered Complete?", *Physical Review* 47: 777-780.

Gaifman, H. (1988), "A Theory of Higher-order Probability", in Brian Skyrms and William Harper (eds.), *Causality, Chance, and Choice*. Dordrecht: Reidel.

Jeffrey, R. (1965, 1983), *The Logic of Decision*. New York: McGraw-Hill.

_ _ _ _ _ . (1988), "How to Probabilize a Newcomb Problem", in *Probability and Causality*, J. H. Fetzer (ed.), pp. 241-251. Dordrecht: Reidel.

_ _ _ _ _ . (1992), *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Jeffrey, R. and Michael Hendrickson (1989), "Probabilizing Pathology", *Proceedings of the Aristotelian Society* 89, part 3: 211-225.

Leeds, S. (1984), "Chance, Realism, Quantum Mechanics", *Journal of Philosophy* 81: 567-578.

Nozick, R. (1969), "Newcomb's Problem and Two Principles of Cchoice", in N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*. Dordrecht: Reidel.

Price, H. (1993), "The Direction of Causation: Ramsey's ultimate contingency", *PSA 1992*, vol. 2.

Shafer, G. (1981), "Constructive Probability." *Synthese* 48: 1-60.

Skyrms, B. (1980), *Causal Necessity*. New Haven: Yale.

_ _ _ _ _ _. (1984), *Pragmatics and Empiricisn*. New Haven: Yale.

_ _ _ _ _ _. (1990), *The Dynamics of Rational Deliberation* Cambridge: Harvard.

Stalnaker, R. (1968), "A Theory of Conditionals", in N. Rescher (ed.), *Studies in Logical Theory*. Oxford: Blackwell, pp. 98-112.

van Fraassen, B. (1989), *Laws and Symmetry*. Oxford: Clarenden Press.