Bayesian Ptolemaic Psychology

Clark Glymour[1]

Carnegie Mellon University

and

Florida Institute for Human and Machine Cognition

1. Ptolemy versus Copernicus in Psychology.

Between Claudius Ptolemy and Johannes Kepler there lay both 1500 years and a methodological chasm. That gap reopens from time to time, and I maintain it has in our own day in cognitive psychology, and in particular in and about Bayesian accounts of the results of various psychological experiments. That is my thesis.

Ptolemy's device, the epicycle on deferent, allowed him to account very accurately for motions with respect to the fixed stars of the sun, the moon, and the five observable planets. Indeed, thanks to Harold Bohr, we now know, as Ptolemy did not, that any periodic motions can be approximated arbitrarily well by iterations of epicycles. So far as the data concern only semi-periodic motions with respect to the fixed stars, Ptolemy's framework can fit anything. Copernicus' framework is not so generous; it requires strict relations among the observable motions of the sun and the planets, and Kepler saw in those relations the very explanatory virtues of Copernican theory, and the comparative explanatory defects of Ptolemaic theory. In Kepler's hand, the Copernican identification of observable features of solar/planetary motion—angles of separation, oppositions,

etc.—with orbital features reduces various empirical regularities to mathematical necessities.

Philosophers and statisticians have often since tried without much success to formalize the general explanatory difference between theories of Ptolemy's ilk and those of Copernican explanatory style, but anyone who has spent some effort with the history of science will usually know each kind when she sees it. The difference can be seen, for example, in the 19th century atomic theory (theories, really) and the theory of equivalent weights. In contemporary psychology there are a lot of Ptolemaic cousins, and increasingly they are Bayesian.

We have seen Ptolemaic theories in various computational theories of cognition, notably, for example, those of Allen Newell and John Anderson. Each of them postulates a universal programming system, with changing psychological glosses. The programming system, SOAR or ACT-R, has an unlimited supply of parameters, adequate to account for any finite behavior, that is, for an behavior subjects can exhibit. The theories provide are no tight constraints of the kind Kepler celebrated, or unlikely but true predictions, such as the phases of Venus. Any physiological or anatomical interpretations of data structures are packed on, ad hoc. Nonetheless, they are celebrated theories, and that says something about the scientific sensibilities of the psychological community. Allen Newell's SOAR briefly had its own society, with limited admission, like the Viennese Psychoanalytic Society once upon a time, and the ACT programs win awards. I once asked the editor of a book describing psychological experiments and their simulation in production system programs that claims in its introduction that the experiments given evidence that Production Systems are the programming system of the mind, which experiments could not be simulated in BASIC. He was candid enough to say "none." In the same annoying spirit, I once asked Newell for any unlikely prediction of his theory. He produced an unconvincing paper in response. Neural net models of psychological processes in which nodes are ambiguously cells or abstract representations of concepts are in the same Ptolemaic mode. We know that neural networks form a scheme for universal approximation of any function; what we see in neural network accounts of psychological

experiments is the ingenuity of the programmer, not a physiological theory except by metaphor.

In the psychology of learning and judgement, theories formulated as programming systems with a gloss are bit by bit coming to be replaced by theories formulated as Bayesian learners, with a hypothesis space, a set of parameters for each hypothesis, a likelihood function relating possible data and each hypothesis and set of parameter values, a prior probability distribution over the hypotheses and their parameters, and the standard Bayesian updating by conditionalization. They are Ptolemaic in an obvious sense: an infinity of parameters is available, and by appropriate parameter choices any consistent preferences can be represented as in accord with maximum expected utility. I will suggest later that neither are they plausibly procedural hypotheses or "rational" hypotheses. My complaint is not lodged against Bayesian statistics, or against attempts to demonstrate, both empirically and theoretically, that aspects of cellular physiology—spiking frequencies, for example--implement probabilistic calculations.

A growing number of psychologists have lately taken up representation of simple causal dependencies by graphical causal models and have investigated and debated a range of considerations about how causal relations are learned: from passive observation of associations, from associations produced by interventions, by Bayesian methods, by matching models to constraints shown by patterns in data, etc. (Glymour, 2003). As in philosophy, the psychological literature shows a predilection for accounts of learning as Bayesian updating, more I think from slavery to fashion than from consideration of what happens when the idea is applied. I will describe three recent examples, each about adult judgement.

2. Waving Hands at Bayes

A recent paper by Kushnir and Sobel (in press) will serve as my first stalking horse, in part because it turns on an experiment that poses inference tasks that are reasonably close in structure to real scientific inference problems, for example in genetics (Ideker, et al., 2000). The paper is a revision of an earlier conference proceeding in which the authors

claimed that the experiment shows that their experimental subjects used Bayesian rather than "constraint" based inference methods, a claim downplayed in the later paper, but one of the authors (personal communication), Sobel, still maintains that the experiment provides support for the claim that subjects are doing "qualitative" Bayesian reasoning. He does not say what that is.

 Kushnir and Sobel gave their subjects ("participants" is the politically correct term) the following instruction:
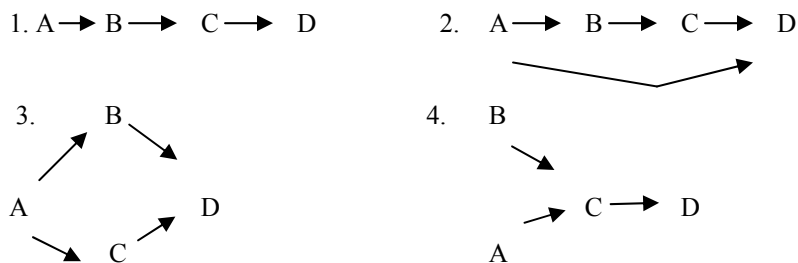
> "In Dr. Science's laboratory, he has created a number of games. Each game has four lights, colored red, white, blue, and yellow. Each light also has zero, one, or many sensors. Some sensors are sensitive to red light, others to blue light, others to white light, and others to yellow light and sensors will activate the light that it is connected to. For example, if the red light is connected to a yellow sensor, then whenever the yellow light activates, the red light will also activate. But, because this happens at the speed of light, all you will see is the red and yellow lights activating together. It is also possible that a light has no sensors attached to it, and therefore is not activated by any other light. Sometimes Dr. Science is very careful about how he wires the lights together. At other times, he is not as careful, and the sensors do not always work perfectly."

The subjects' task was to determine which lights had sensors for which other lights, which amounts to determining a possibly cyclic directed graph on four vertices. Subjects could push a button to illuminate any single light on any trial, and could also disable the sensors on any single light on any single trial; they were required to make at least 25 trials and, when they were satisfied that enough data had been acquired, then to report a yes/no decision (and level of confidence) for each of the 12 possible edges. Each light remained off unless its button was pushed or one of its sensors "detected" another light. The probability of a light being on given that its button was pushed was 1. We are told that "Each sensor caused activation 80% of the time" which is an insufficient specification of the joint probability distribution, given that a light button is pushed, of the activation (on/off) states of the remaining lights: it does not tell us the probability of activation—i.e., illumination-- when two or more of a light's sensors receive a signal. Moreover, the instruction given the subjects contradicts the actual set-up: subjects are

told that sometimes sensors are wired carefully and sometimes not, so they should expect that some sensors work 100% of the time and some sensors that do not. I leave that complication aside.

Subjects were trained with a deterministic system to recognize that use of the control—putting a "bucket" over one of the lights--which effectively disconnected a chosen single light from the system, could disambiguate some situations: "They were shown through a series of generative interventions that purple caused green and that grey caused purple, but were unable to tell whether grey was also a direct cause of green. They were then shown that by putting the bucket on purple and activating the grey light, they could distinguish whether grey was a direct cause of green or only an indirect cause of green (via the purple light)." Giving such instruction implies that the psychologists' thought they were investigating implicit features of explicit strategies.

Subjects were then tested, each with data generated from their interventions on 4 alternative structures:

1. A → B → C → D   2. A → B → C → D

3.   B   4.   B
   A   D   A   C → D
      C

Lights of four colors were randomly assigned the roles of A, B, C and D.
Subjects did very well.

The psychologists' interest was in demonstrating that subjects more accurately determined structural relations if the participants themselves chose the interventions (*autonomous* condition) than if they were given instructions, i.e., ordered, as to which button to push and which control to impose (*following* condition). The sequence of interventions in the following condition were in each case those of some other

autonomous subject. The psychologists suggest—indeed in a previous paper, claimed outright--that there is a Bayesian account of their results, but they give no details. I will return at the end of this essay to explanations of this effect, which almost certainly cannot be as the psychologists suggest. In any case, to give a Bayesian account of the difference between autonomous subjects and followers, there must first be a Bayesian account of learning in the autonomous condition. What could that be?

In order to learn which graphical structure obtains, a Bayesian agent must have prior probabilities over the 4096 possible graphs on four vertices, and for each graph must specify the probability for any data sequence conditional on that graph. With these probabilities, the agent can in principle compute her probability for the data she has seen, and she can, again in principle, compute the posterior probability of each graph by Bayes Rule.

We can specify prior probabilities in many ways. For example, all graphs can be given an equal prior; or all graphs with the same number of edges can be given the same prior with the total probability for each edge number equivalence class equal, or perhaps weighted in favor of graphs with fewer edges; or we can assign a prior, k, for each edge independently, and take the prior probability of any graph to be equal to the product of the prior probabilities of its edge presences and edge absences. Since nothing is known beforehand, the edge priors should presumably be equal. So a graph with 1 edge will have a prior probability proportional to $k (1-k)^{11}$ and a graph with two edges, any two edges, will have a prior probability of $k^2(1-k)^{10}$, and so on. I will illustrate a Bayesian procedure (for the first trial only—for reasons to be explained) with the third way of specifying the prior probabilities.

To specify the probability of a data sequence for any particular graph requires a "parameterization" of the directed graph, with parameter values that determine the probabilities for any trial outcome, and a prior probability distribution over the parameters. I make the simplest specification I can think of consistent with the psychologists' set up: the probability of any light being on given that its button is pushed

is 1. The probability of any light being on given that it is covered is 0. For any other condition, the probability that light X is on is:

(1) $Pr(X = on) = Pr(\sum_{Y \in \mathbf{OnPar}(X)} q_Y$

Where Y has value 1 if light Y is on and has value 0 otherwise, $\sum$ is Boolean addition, **OnPar**(X) is the set of parents of X (lights for which X has a sensor) that are on, and q_Y is a parameter taking values in $\{0,1\}$. The parameters q_Y are independent in probability of each other and of all of the variables representing lights.

Formula (1) is the familiar noisy-or gate (Pearl, 1988). If X has a single parent, Y, that is on, the probability that X is on is $Pr(q_y = 1)$. If X has two parents, Y and Z, that are on, then the probability that X is on is $Pr(qy = 1) + Pr(qz = 1) - Pr(qy = 1)Pr(qz = 1)$, and if three such parents, Y, Z, W, $Pr(qy = 1) + Pr(qz = 1) + Pr(q_w = 1) - Pr(qy = 1)Pr(qz = 1) - Pr(qz = 1) Pr(q_w = 1) - Pr(qy = 1) Pr(q_w = 1) + Pr(qy = 1)Pr(qz = 1)Pr(q_w = 1)$. I will assume, again for the simplest and most easily computed case, that $Pr(q_x = 1) = p$, the same for all X and for all graphs. For each graph, each prior probability for the graph and each value of p, a unique joint probability is specified for the values of all uncovered variables given that a particular button is pushed.

A Bayesian agent must combine the prior probability of each graph with a prior probability distribution *over* all possible values of p. Let P_G (d; p) be a probability distribution for the outcomes on trial d, that results from a value p for the probabilities in (1), let P_G be the subjective probability distribution over values of p, and let D be the set of d—each of which is a list of which lights are on and which light is pushed—for a set of trials. Again to make matters as simple as possible, I assume P_G is a uniform distribution over p. Then:

(2) $Pr(D \mid G) = \Pi_{d \in D} \int_{PG(d;p) \in PG} P_{G,}(d; p) \, d\mathrm{p}$

(3) $\Pr(D) = \Sigma_G \Pr(D \mid G) \Pr(G)$

And, by Bayes rule

(4) $\Pr(G \mid D) = \Pr(G) \Pi_{d \in D} \int_{PG \in PG(d)} P_{G,}(d; p) \, dp \, / \, \Sigma_G \Pr(D \mid G) \Pr(G)$

And finally, for any edge E, let $G(E)$ be the set of graphs that have E as an edge. Then

(5) $\Pr(E \mid D) = \Sigma_{G \in G(E)} \Pr(G \mid D)$

Formula (2) is only valid if the trials are independent. Arguably, they are not. d is in effect conditioned on which buttons are pushed and which buckets are used. So the probability of D|G is in effect conditioned on the entire sequence of button pushings and bucket placements. (I don't see any way to get around this, because since people are engaged in discovery, the button pushings from trial to trial are not independent. The button pushings may depend upon the outcomes of the previous trial. Of course for each d, the only thing that matters is the buttons pushed on that individual trial.

$P(G|d1,d2) = P(G)P(d1,d2|G)/P(d1,d2)$

$P(d1,d2|G) = P(d1|G,p) \, P(d2|G,p,d1)$

$P(d1|G,p) = P(d1|G,p,act1)$ where act1 has probability 1

$P(d2|d1,G,p) = \Sigma \, P(d2|d1,G,p,act2) \, P(act2|d1,G,p) = \Sigma \, P(d2|G,p,act2)P(act2|d1)$

$P(d1,d2|G) = P(d1|G,p,act1) \, \Sigma P(d2|G,p,act2)P(act2|d1)$

where act2 may be a function of d1.

Consider now how to compute the posterior probabilities of edges for a *first* trial in which W is covered and X is pushed with the result that Z goes on but Y does not. That enables us to compute by considering only graphs on the three variables X, Y, Z, with six possible edges.

Since Z but not Y is on when X is pushed, and W is covered, only graphs with an X -> Z edge are consistent with the data. Since W is covered, we need attend only to subgraphs

on the variables, X, Y, Z. There are 32 graphs on X, Y, Z containing an X -> Z edge. I divide them into 3 classes as follows: Class 1 contains neither X -> Y nor Z ->Y; Class 2 contains X ->Y but not Z -> Y, or Z -> Y but not X -> Y; Class 3 contains both X -> Y and Z -> Y.

For each value of p, each Class 1 graph implies that the probability that Z, X and only Z, X are illuminated when X is pushed and W is covered is p. Class 1 contains

1 graph with 1 edge and has prior probability $\approx k(1-k)^5$

3 graphs with 2 edges, each graph with prior probability $\approx k^2(1-k)^4$

3 graphs with 3 edges, each graph with prior probability $\approx k^3(1-k)^3$

1 graph with 4 edges, with prior probability proportional to $k^4(1-k)^2$

For each value of p, each class 2 graph implies that the probability that Z, X and only Z, X is illuminated when X is pushed and W is covered is p(1- p). Class 2 contains

2 graphs with 2 edges, each graph with prior probability proportional to $k^2(1-k)^4$

6 graphs with 3 edges, each graph with prior probability proportional to $k^3(1-k)^3$

6 graphs with 4 edges, each graph with prior probability proportional to $k^4(1-k)^2$

2 graphs with 5 edges, each graph with prior probability proportional to $k^5(1-k)^1$

For each value of p, each Class 3 graph implies that the probability that Z, X and only Z, X are illuminated when X is pushed and W is covered is $p(1 - p)^2$. Class 1 contains

1 graph with 3 edges, with prior probability proportional to $k^3(1 - k)^3$

3 graphs with 4 edges, each graph with prior probability proportional to $k^4(1-k)^2$

3 graphs with 5, edges, each graph with prior probability proportional to $k^5(1-k)^1$

1 graphs with 6, edges, with prior probability proportional to $k^6(1 - k)^0$

Now we can compute the probabilities of various edges. For an edge such as Z -> X, or Y -> Z or Y -> X that make no new pathways from source to Y save through X, the posterior probability is the prior probability. For an edge such as X -> Y, since half of the probability of Class 2 graphs is for graphs with an X -> Y edge, it follows that the posterior probability of the X -> Y edge is equal to:

$\int_0^1 \{ 1/2(p(1\text{-}p) [2k^2(1\text{-}k)^4 + 6\ k^3(1\text{-}k)^3 + 6k^4(1\text{-}k)^2 + 2k^5(1\text{-}k)]) + (p(1-p)^2 [k^3(1-k)^3 + k^4(1\text{-}k)^2 + k^5(1\text{-}k) + k^6) \}dp$

_____

$\int_0^1 \{ (p[\ k(1\text{-}k)^5 + 3k^2(1\text{-}k)^4 + 3k^3(1\text{-}k)^3 + k^4(1\text{-}k)^2]) + (p(1\text{-}p) [\ 2k^2(1\text{-}k)^4 + 6\ k^3(1\text{-}k)^3 + 6k^4(1\text{-}k)^2 + 2k^5(1\text{-}k)]) + (p(1-p)^2 [k^3(1-k)^3 + k^4(1\text{-}k)^2 + k^5(1\text{-}k) + k^6]) \}dp$

The integrals are straightforward but tedious. The posterior does not simplify to any fraction independent of k, since it equals a ratio of the form

$$\frac{B + C}{A + 2B + C}$$

where A, B, C are positive polynomials in k. The posterior of Z -> Y is similar. Note that if each graph, rather than each edge, were equally likely, the computation would be slightly simpler. For the numerator and denominator in the posterior, one would need to count the number of graphs in each of the relevant classes above and multiply the cardinality of the graph class by the appropriate function of p, integrate over p, and sum.

If both Y and Z are on when X is pushed and W is covered, then the data require at least one of the following: an X ->Y and X -> Z edge; an X->Y and Y -> Z edge; or an X -> Z and Z – Y edge. I will pass on the calculation.

If neither Y nor Z are illuminated, every graph on X, Y, Z is consistent with the data, and the graphs fall into several distinct classes with the likelihood as a function of p the same within each class and different between classes. Again, I skip the calculation.

The calculation of posterior probabilities from succeeding trials can become more intricate, as the probabilities of edges cease to be uniform and independent. Further, if the choice of which button to push and light to cover depends on the outcomes of previous

trials, a Bayesian agent is essentially engaged in a sequential decision problem in which she is obliged to calculate the most informative next experiment. That means, for each trial, taking account of the previous trial, she must calculate the expected change in her degrees of belief from each of the 16 possible experiments she can make. Even for this simple problem, those calculations are not easy and there are a lot of them (Tong and Koller, 2001),

There are only three resources available to provide a Bayesian explanation of the difference between followers and autonomous subjects: they differ in their prior probabilities for the edges; they differ in the likelihoods—the probabilities of data given each graph and parameter probabilities—or they differ in the relevant data. The likelihood of the data given a graph, a value of p, and a choice of button to push and light to cover are a logical matter, the same in both cases; the relevant data the subjects see are the same—which lights illuminate when which buttons are pushed when which lights are covered. Two differences remain possible. One is the possibility of differences in prior probabilities in the two conditions. One source—the only one I can think of—for different prior probabilities in the two conditions is the suspicion, or degree of belief, in the following condition that the selection of button to be pushed and illumination of other lights is confounded by unknown factors. That is, something unrevealed causes the selection of the button for X and the illumination of light Z when the button for X is pushed. (A related explanation has been offered for our belief in freedom of the will (Glymour, 2004)). The effect of allowing this hypothesis in the case in which W is covered and Y does not go on, calculated above, is to decrease the prior probability of the X -> Z edge compared with the value in the autonomous condition, other things equal. While such an hypothesis predicts a difference in posterior probabilities for followers and autonomous subjects, it does not suffice to explain why the followers are less accurate unless, by sheer chance (whatever that is for a Bayesian) the followers priors were sufficiently farther than the autonomous from the true frequencies of edges. Finally, followers and autonomous subjects could differ in their probability distribution for choice of actions. For example, if someone concocts a known strategy for choosing button pushings, it could be incorporated into Bayesian calculations. A set of button pushings

not under their control could have a different probability distribution, perhaps leading to greater uncertainty. Each subject has her own preferred strategy in which there is a unique set of best decisions for the next trial given the outcomes of previous trials. Autonomous subjects, but usually not followers, get to use their preferred strategy. Followers instead must use a probability distribution over the strategies of whichever unknown autonomous subject they shadow. It is not clear, however, that, since the follower sees (indeed carries out) the actual sequence of decisions made by the autonomous subject she shadows, that the probabilities she has over the autonomous strategies make any difference (compared to the autonomous subject she is shadowing) in the end when she must report her judgements of the probabilities of edges.

The complexity of calculations required for a simple, coherent Bayesian account that is reliable over the general problem the psychologists pose seems to me make it implausible that their subjects instantiate a Bayesian procedure, and of itself a Bayesian account offers no evident means to explain the superior performance of subjects in the autonomous condition. To merit the title of explanation, the suggestion that subjects use "qualitative" Bayesian procedures is obliged to show a procedure that is not "quantitative" and its connection with some real (i.e., quantitative) Bayesian procedure. The possibilities are myriad: Use only inequalities; assign incoherent priors or compute incoherent posteriors; use some alternative to conditioning for updating

What is a "constraint based" learning procedure? Some data mining algorithms for causal relations build a causal structure, or class of such structures from estimates of probability relations among the observed variables obtained from the sample data, for example from results of tests of hypotheses of conditional independence. But the psychologists' problem is special in that aspects of the structure are revealed trivially, or by simple counts, without recourse to probability estimates. For example, assuming the probability of sensor error is the same for all sensors, all of the following rules are valid or hold almost certainly

in the limit as the number of trials N increases without bound. They make good decision procedures even for small N when p is > ½ .

Rule 1. If X is pushed several times and Y is illuminated more often than is Z, and W is covered, then there is a directed edge from X to Y.

Rule 2: If X is pushed, and Y is illuminated, then there is a directed path from X to Y not through any non-illuminated light.

Rule 3: If when X is pushed repeatedly, Z is never on unless Y (respectively, Y or W) is on, but not vice versa, then every directed path from X to Z is through Y (respectively Y or W).

Rule 4: If in trials which X is pushed, neither Y nor other lights are illuminated, then there is not a directed edge from X to Y.

The rules are obvious or should become obvious with a few experimental button pushes. They can be applied in a variety of strategies to solve the problems the psychologists pose, with small memory and computational requirements and with or without the use of the covering control. Merely for example, the following procedure:

    Note every possible edge as "can't tell."

    Choose a button at random and push it 5 times;

    Apply each of the 5 rules to the result and remember the required and prohibited directed edges and directed paths;

    Do the same with another button

    Until all buttons are tested

    For each "can't tell" edges that remains, say X -> Z, cover each of W, Y in succession and push X 5 times. Apply the rules; if neither covering eliminates the X -> Z edge by the rules, then mark the X -> Z edge required.

E.g., for structure 3, starting with pushing D, then C then B then A, we should find:

> Push D: By rule 4, eliminate all edges out of D into A, B, C
>
> Push C: By rule 4, eliminate all edges out of C into A, B; and by rule 1 C -> D is required
>
> Push B: parallel to C
>
> Push A: By rule 2 and results already obtained, A -> B and A-> C are required.
>
> There remains a "can't tell" A -> D edge.
>
> Cover B and push A: By rule 3, eliminate A -> D

Of course, I don't claim that the psychologists' subjects, or any one of them, used just this procedure. The point is that their problem is reliably solved by simple, non-Bayesian, low memory and low computational counting procedures. One way or another, I bet that is what their subjects are doing.[2]

3. Common Cause versus Common Effect

The first of a series of interesting experiments by Steyvers, et al. investigated how well subjects could use co-occurrences to distinguish between a system in which two of three variables were independent causes of a third and a system in which a single variable was a common cause of two others. Subjects were told that there

---

[2] What then explains the difference in accuracy between the following and autonomous subjects? Lots of things, quite possibly. Commitment, perhaps. Some of the following subjects may not have embraced the goal given that they did not choose the actions. Confusion and memory limitations perhaps. Some of the following subjects may have had a fixed strategy in mind and not realized how to adapt it to the choice of interventions they were given. The set-up of Kushnir and Sobel's study does invite studies of human behavior in comparison with other norms. For example, suppose with low frequency the lights sometimes come on spontaneously, and we confine ourselves to acyclic graphs. For this setting, recent work by Frederick Eberhart suggests that the graphical causal structure can be most efficiently and accurately identified by strategies that push multiple buttons simultaneously. I should not be surprised if a slight modification of results in Eberhardt et al. (2005) applies as well to the semi-deterministic set up the psychologists use. It should be of interest whether people spontaneously use that recourse, and whether they do better if it is allowed than if they are restricted to pushing at most one button at a time.

are three aliens who communicate telepathically, that telepathic communications usually succeed but sometimes fail, and were asked to determine whether a particular alien (C) receives words from both of the others or sends words to both of the others. They were told the two structures are equally likely.

Subjects were given a block of eight trials of word occurrences for a given structure, and were given multiple blocks of trials with structure randomized across blocks. After each trial, subjects were required to provide a conjecture as to the structure for the block in which the trial occurred. No feedback was given to the subjects. Unknown to the subjects, the transmission success probability was set at .8; if two transmissions in the same trial were received, one was chosen at random; if no successful transmission occurred to an alien, the alien displayed a word at random from among the vocabulary of 10 words.

Steyvers et. al. found that about half of the subjects scored at chance; of the remaining, a fraction scored above chance with little variation between trials, and a highest scoring fraction scored above chance and improved across trials within a block, but not across blocks.

What explanations are possible for these results? The subjects were college students, and the half that scored at chance may simply not have cared about the problem. The half that scored above chance and did not improve on trials within a block presumably used some simple heuristic that gave them a better than ½ chance at the right answer on any single case; the group that improved with trials presumably adapted their hypothesis to increasing evidence. A simple, reliable

heuristic is available and obvious: guess that C is a common cause if A = B, and guess that C is a common effect otherwise. Simply guessing that C is a common cause if the number of cases in which A = B is greater than the number of cases in which A ≠ B, and guessing that C is a common effect otherwise, gives a learning rule. These simple procedures are in close agreement with the average performance of the two groups of subjects Steyvers et. al. identify whose judgements are correct more often than chance.

The Steyver's, et al., explanation is different. The subjects have prior probabilities P(CC) and P(CE) for the common cause and common effect models. The subjects know the two structures will be given with equal probability, so their priors are equal. They judge by the posterior ratio

(1) $\varphi = P(CC|D)/P(CE|D)$

where the trials are independent

(2) $\log(\varphi) = \sum_{t=1}^{T} \log \frac{P(D_t|CC)}{P(D_t|CE)}$

and the probabilities are determined by the true values of the probability of transmission success (α) and the number of words (n) according to

Table 1. The four data patterns (*D*) and their probabilities under a common cause (*CC*) and common effect model (*CE*) with α=0.8 and *n*=10.[3]

D		P(D|CC)	P(D|CE)
1) A=B=C	All same	.67	.096
2) A=C, B≠C or B=C, A≠C	Two adjacent same	.30	.87
3) A=B, A≠C	Two outer same	.0036	.0036
4) A≠B, B≠C, A≠C	All different	.029	.029

Further, subjects differ in their learning behavior according to two parameters. One is an exponential decay term δ for the contribution of earlier trials to the posterior:

---

[3] From Steyvers, et al.

$$(1) \ \log(\varphi) = \sum_{t=1}^{T} \left[\log \frac{P(D_t \mid CC)}{P(D_t \mid CE)} \right] e^{-\delta(T-t)}$$

The other parameter allows decisions to be made not strictly on the odds ratio. That is, for some non-negative quantity γ:

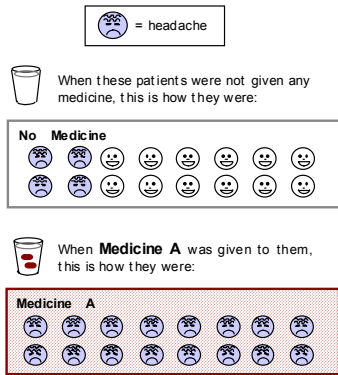$$(2) \ P(CC) = \frac{1}{1 + e^{-\gamma\varphi}}$$

To carry out this computation, subjects must in each block remember which words are the same or different for each alien in the order of the trials so far, must know the true values of the transmission probability, α, the number of words, n, must use them to compute, as in table 1, the ratio of the probabilities of each trial outcome on each hypothesis , must decrement that trial term exponentially by place order, take the log, sum the results over all trials so far, take the anti-log of the sum, and substitute the result for φ in equation (3) above. This, I submit, is not a plausible procedural hypothesis.

The simple heuristic explanation could have been tested by including blocks of trials in which C was masked. Unfortunately, that was not done.

4. Bayesian Cheng Models

Patricia Cheng has produced a series of studies exploring how large sample ("asymptotic") judgements of the strength of a putative causal factor in producing an effect are inferred from time-ordered co-occurrence data, and she has systematized the results in an interesting theory that can be reconstructed as inference to parameter values in noisy-or-gate and noisy-and-gate graphical causal models. Inevitably, her theory and some simple experimental results have been Bayesed.

Hu et al. consider experiments in which subjects are shown data under two conditions, before and after a treatment.

Example of an experimental display, showing patients who had not (top) or had (bottom) received an allergy medicine, and who either had or had not developed headaches.

The point of the experiments is to demonstrate that subjects have a strong preference for attributing a causal role when the association between the putative cause and the effect is invariant, either always, or almost always, generating the effect or always, or almost always, preventing the effect. I do not contest that conclusion at all, and anyone who has spent much time with naïve subjects' judgements of cause and effect will not be surprised. Hu et al., claim a much stronger conclusion: subjects are doing Bayesian updating with non-uniform priors biased towards necessity or sufficiency of the absence or presence of a factor. According to their account, subjects must decide between two graphs, in the first of which B is a cause of E and in the second, B and C are causes of E. As usual the decision depends on the support

(4) $\qquad \log \dfrac{P(Graph1 \mid D)}{P(Graph0 \mid D)},$

To obtain the ratio priors and likelihoods are required; likelihoods require integration over parameters

(5) $\qquad P(D \mid Graph1) = \int_0^1 \int_0^1 P(D \mid w_0, w_1, Graph1) P(w_0, w_1 \mid Graph1) dw_0 dw_1$

$\qquad P(D \mid Graph0) = \int_0^1 P(D \mid w_0, Graph0) P(w_0 \mid Graph0) dw_0$

For Cheng noisy-or-gate models, in which b, the presence or absence of B, is 1, and the presence (1) or absence (0) of C is c, and e^+ means e occurs, and w_0 and w_1 are the causal powers of B and C, respectively, to produce e,

(6) $P(e^+ \mid b, c; w_0, w_1) = 1 - (1 - w_0)^b (1 - w_1)^c$

Now with data in which there are $N(c^+)$ cases in which C occurs, and $N(e^+)$ cases in which e occurs, and similarly for $N(c^-)$

(7)
$$P(D \mid w_0, w_1, Graph1)$$
$$= \binom{N(c-)}{N(e+,c-)} w_0^{N(e+,c-)} (1 - w_0)^{N(e-,c-)}$$
$$\binom{N(c+)}{N(e+,c+)} [1 - (1 - w_0)(1 - w_1)]^{N(e+,c+)} [(1 - w_0)(1 - w_1)]^{N(e-,c+)}$$

$$P(D \mid w_0, Graph0)$$
$$= \binom{N(c-)}{N(e+,c-)} \binom{N(c+)}{N(e+,c+)} w_0^{N(e+,c-)+N(e+,c+)} (1 - w_0)^{N(e-,c-)+N(e-,c+)}$$

One still needs $P(w_0, w_1 \mid Graph)$, for which Hu et al. specify:

(8) $P(w_0, w_1 \mid Graph1) = \dfrac{1}{Z} \left[e^{-\alpha w_0} e^{-\alpha(1-w_1)} + e^{-\alpha(1-w_0)} e^{-\alpha w_1} \right]$

Here α indicates the bias towards necessity and sufficiency. $\alpha = 0$ is a uniform distribution, and Hu et al compare predictions of judgements on the uniform with $\alpha = 30$. Z is a normalizing constant.

Thus, according to Hu et al., to form a judgement as to whether C is a cause, or how likely it is to be a cause, subjects must separately multiply (8) with appropriate parameter values by each equation in (7) and integrate each result over values of w_0 and w_1. They show that the procedure nicely fits their data, but no evidence that subjects carry out any such complex calculations.

5. Modeling Versus Explaining

One sometimes hears such Bayesian models described as "rational" models of judgement, placing them as contributions to a dispute within psychology that has

its modern genesis in the well-known work of Tversky, Kahneman, and others. "Rational" in this context is a misappropriated term of moral psychology, wrenched from most of its sense. "Rational" as Bayesian modelers use the term means judgements that are probabilistically coherent and updated by conditionalization. The connection with rationality as ordinarily understood is through the Dutch Book arguments, which show that if degrees of belief are betting odds on bettable propositions, then there is a sure win combination of bets against anyone whose betting odds are not probabilistically coherent and who does not update them by conditionalization.

One can be rational in this sense and mad as a hatter, quite unable to learn the truth about much of anything. Simply give probabilities 0 to a lot of true stuff, and probability 1 to a lot of false stuff, and those degrees of belief will never change by conditionalization. Arguably, we cannot avoid that state. If, as I suspect and cognitive psychologists used to assume, we are bounded by Turing computability, then if we are probabilistically coherent, we must give probability 0 to an uncomputable infinity of logically contingent propositions. Turing's theorem plus the premise that our cognitions are Turing computable implies that we must be infinitely dogmatic. The newspapers confirm the hypothesis.

Taking computation and rationality seriously means allowing that rational means to ends must be those that can be carried out effectively, in the computational sense of "effective." The goal of inquiry is presumably interesting truth, and we know there are learning problems in which whichever of an infinite set of mutually exclusive alternative hypotheses is true can eventually be learned by Turing bounded learners, but not by any Bayesian Turing bounded learners. Whether there are interesting problems of this kind is unknown, but the very existence of any such examples shows, I think, that the equation of "rational" and Bayesian is unreflective rhetoric.

Ptolemy's astronomy models the motions of the planets very well, but it doesn't correspond to what goes on. If our interest were merely in a mathematical synopsis of the data, or prediction of the distribution of values of more data of the same kind, a model like Ptolemy's is just fine. If one wants to understand what is going on, it doesn't help much. The same is true of Bayesian models of judgement that have little or no procedural plausibility or evidence. At least SOAR and Act-R and so on attempted to describe procedures that brain might carry out through some unknown and unspecified way of compiling higher order computational relations into relations among nerve cells. The Bayesian models I have considered do not attempt even that. They are style over substance.

References

Kushnir, T. and D. Sobel, The importance of decision making in causal learning from interventions. (in press) *Memory and Cognition*.

Glymour, C. (2003) *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge: MIT Press.

Glymour, C.   We believe in freedom of the will so that we can learn. 2004. Comment on D. Wegner, *The Illusion of Conscious Will*, Behavioral and Brain Sciences,.

Ideker, T. V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, L. Hood, 2000. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network**,** *Science* 4 May: Vol. 292. no. 5518, pp. 929 – 934.

Lu, H. , A. Yuille, M. Lilejeholm, P. Cheng and K. Holyoak, (2006). Modeling Causal Learning Using Bayesian Generic Priors on Generative and Preventive Power. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-eighth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Steyvers, M., Tenenbaum, J., Wagenmakers, E.J., Blum, B. (2003). Inferring Causal Networks from Observations and Interventions. Cognitive Science, 27, 453-489.

Tong, S. and D. Koller, 2001. Active Learning for Structure in Bayesian Networks, *Proceedings of the 2001 International Joint Conference on Artificial Intelligence.*