Resolving Goodman's Paradox

How to Defuse Inductive Skepticism

Taner Edis
Department of Physics, Truman State University, Kirksville MO 63501, U.S.A. *

Abstract. Subjective Bayesian inference is unsuitable as an ideal for learning strategies to approximate, as the arbitrariness in prior probabilities makes claims to Bayesian learning too easily vulnerable to inductive skepticism. An objective Bayesian approach, which determines priors by maximizing information entropy, runs into insurmountable difficulties in conditions where no definite background theory is available. However, this lack of background knowledge makes the maximum entropy argument directly applicable to the process of drawing samples from a population. As a result, evidence can be seen not just as eliminating a number of incompatible hypotheses out of an infinity of possibilities, but as being *representative* of the true state of affairs. Hence inductive skepticism can be avoided, as demonstrated by a resolution of Goodman's 'grue' paradox. This leads to a clearer understanding of the vital role abductive processes and tools like simple generalization play in learning.

Keywords: Goodman's paradox, induction, inductive skepticism, statistical inference, Bayesian learning, maximum entropy, machine learning, abduction, generalization

1. How can we learn?

If we have observed twenty black crows, are we then entitled to expect other crows are also black? Such questions signal trouble; after all, showing that the ways we generalize from experience are not arbitrary is a notoriously intractable problem. Scientists and engineers are usually happy to assume we can learn, staying away from the eternal philosophical handwringing over the possibility of knowledge.

However, methods of statistical inference have become basic equipment in our sciences, and workers in fields like statistical physics or machine learning often need to ask how reliable inference is best accomplished. If, for example, Bayesian reasoning is taken as an ideal for methods of machine learning to approximate, the issue of choosing priors will arise, and inductive skepticism lurks below the surface. From a more philosophical point of view, if inductive skepticism means the success of learning is a total mystery, this is a motivation to think a materialist view of mind breaks down.[1] So a scientific approach to minds requires some defusing of skepticism, even if a quest for a complete justification of induction has an air of futility.

* (e-mail: *edis@truman.edu*. Supported in part by NASA grant NAG5-3660.)

Debates over cognitive relativism further highlight the problem. Statistical inference and conventional AI rely on an empiricist approach: there is data, which is then fed into the machinery of method, in order to support or undermine contending hypotheses. But the method or 'inference engine' is not uniquely determined; it appears others could be chosen without logical contradiction. And this method cannot be supported by its past success, as that would be arguing in a circle. It stands as a groundless, foundational principle—which is a polite way of saying it is completely arbitrary.

Of course, there are ways to avoid some of these pitfalls. For example, we can take a coherentist view where nothing in our web of knowledge serves as a foundation all else one-sidedly depends upon. Our methods, in other words, are best seen as results rather than principles immune to criticism. They can be supported by a detailed account of how we learn, richly connected to theories concerning physics, cognition, and the history of science. To avoid getting trapped in a highly coherent fantasy, we also need reality checks. So we can try to achieve 'correspondence through coherence' (Davidson, 1983; Prado, 1987); alternatively, we might finally anchor our web upon 'self-presenting properties' at the bottom of our observations (Chisholm, 1989).

These are rough themes; details depend on particular philosophers. A crucial ingredient, however, appears to be an openness to novelty. History need not rubber-stamp science, it may surprise us by finding conceptual revolutions are best explained by social causes external to scientific practice. An experiment might cast doubt on prevailing theory. In this way, self-support of our style of learning avoids being a vicious circle, looking more like a positive feedback loop (Giere, 1988). Though we start with folk-theories which seem little more than prejudices, we can conduct genuinely independent investigations allowing us to converge to stable results resistant to critical revision.

Such a picture suggests we can develop reliable methods which, while not enjoying absolute certainty, are not completely arbitrary either. To learn, we can depend on statistical methods, heuristic devices, and abduction strategies including simple generalization (Leake, 1995). This toolbox is subject to revision and improvement along with the knowledge it underwrites. Most gratifyingly, we can replace the inference engine of old-fashioned AI with a structure more reminiscent of the interdependent networks of neuroscience.

For all its attractions, though, this picture of learning is still vulnerable to inductive skepticism. Its advantage is that without the extra complications brought on by keeping the three tiers of data, method, and theory sharply separated, the source of our troubles is much clearer. Learning depends critically on our being able to gain a foothold in un-

derstanding *new* situations—where our background theories constrain us the least; when we start to explore new territory by simple classification and generalization. And the claim of an inductive skeptic is just that when we observe twenty black crows, that is the extent of our knowledge, full stop. We can conclude nothing about the unobserved. Hence knowledge can never gain a foothold, and so for all our sophistication, we still build castles in the sky. We might avoid vicious circles, but without any purchase on new situations, we will still get nowhere.

2. Bayesian inference and Goodman's paradox

2.1. A spatial version of Goodman's paradox

To better appreciate the challenge and relevance of inductive skepticism, we need to state the problem in more rigorous terms. A Bayesian formulation is a good candidate to accomplish this. On the positive side, arguments from betting behavior, requirements of consistent reasoning, and other considerations converge on a Bayesian probability calculus with impressive consistency (Howson and Urbach, 1993; Cox, 1961; Snow, 1995). And it is easier to explore basic issues in probabilistic reasoning with a Bayesian approach, rather than relying on the cumbersome devices of classical statistics. On the negative side, it seems as many philosophers and statisticians find fatal flaws in a Bayesian approach as those who think it is obviously correct (Mayo, 1996). Prominent stumbling blocks Bayesians face are their association with 'subjective probability' notions and the difficulties in choosing prior probabilities at the start of the process. However, these problems are in fact closely related to the apparent arbitrariness which fuels inductive skepticism. Hence both the strengths and weaknesses of Bayesian inference make it ideal for discussing inductive skepticism.

To express inductive skepticism in a Bayesian idiom, we can use Nelson Goodman's well-known 'grue' paradox or 'new problem of induction' (Goodman, 1983). Observing all emeralds have been green leads us to expect they will remain so. But the same evidence also supports the hypothesis that emeralds are grue, where 'grue' means something is green up to the year, say, 2010, and blue afterwards. As always with inductive skepticism, the problem comes with inferring the unknown from the known: since our evidence does not discriminate between green and grue, our expectation that emeralds will remain green cannot be based on evidence. It is not enough to invoke uniformity; then the question becomes, uniformity of what, green or grue? The sudden discontinuity at 2010 looks suspicious, but from the point of

view of someone to whom 'grue' and 'bleen' comes naturally, 'green' is a strange color which changes from grue to bleen at 2010.

Goodman's paradox is still liable to seem rather contrived. A clearer statement closer to realistic problems of inference can be obtained by recasting the paradox in spatial terms rather than relying on future discontinuities. After all, the heart of the paradox concerns inferring the unknown, whether separated from us in space or in time.

Imagine an expedition to a planet in a distant galaxy. From their spaceship, they send down an AI-endowed robot to make a visual survey under the thick cloud cover. The robot soon notices some green boulder-like objects which emit radio noise. The survey team dub these 'norks', and observe a thousand large, green norks before calling it a day.

The team then maps radio signals from the whole planet and finds about a million norks. They remark what a strange planet it is, covered with green radio noise emitters. But Mission Control has included a philosopher in the crew; she corrects the survey team, saying they have not found that norks are green, only that norks in the area they explored were green. Radio signals show norks on the other side of a mountain range the robot never crossed. Those are wholly distinct and separate; the survey learned nothing about *them.* One of the AI programmers objects. After the first twenty or so norks, all which were green, the robot generalized, adopting the hypothesis that all norks were green. So did the crew in the spacecraft. And as other norks kept turning up green, their confidence increased. After all, these norks could have been blue instead. They were not.

The philosopher agrees. Call the hypothesis 'all norks are green' g, and the evidence of a thousand green norks e. The survey starts out with a prior probability for g, which is $P(g)$. As green norks accumulate, the probability that norks are green increases. In Bayesian terms, the calculation is straightforward:

$$P(g|e) = \frac{P(g)P(e|g)}{\sum_j P(e|h_j)P(h_j)} \tag{1}$$

where the h_j are all the (mutually exclusive) hypotheses about nork color distribution under consideration. The posterior probability of g is $P(g|e) = kP(g)$; and since $P(e|g) = 1$,

$$k = \frac{1}{\sum_j P(e|h_j)P(h_j)} > 1 \tag{2}$$

So $P(g|e) > P(g)$. Since the probability of g grows with confirming evidence, it appears that simple generalization does, in fact, set us on the path to learning.

Unfortunately, the philosopher now points out, this is an illusion. $P(g|e) > P(g)$ only because e eliminates certain rival hypotheses such as b, 'all norks are blue'. Others survive. Consider the spatial version of the grue hypothesis: $\tilde{g}$, standing for 'norks on the explored side of the mountains are all green, and those on the other side are all blue'. The evidence e does not discriminate between g and $\tilde{g}$; in fact, it confirms both to the same degree: $P(\tilde{g}|e) = kP(\tilde{g})$. The posterior probability of g may have increased, but what knowledge the survey appears to have gained about the color of unobserved norks derives entirely from the initial probabilities prior to any evidence. Assume, for the sake of illustration, that the only hypotheses under serious consideration were g, $\tilde{g}$, b, and $\tilde{b}$. Observing e eliminates b and $\tilde{b}$, which may, in fact, result in the survey judging it more likely that norks on the other side are green rather than blue. But this will happen only if $P(g) > P(\tilde{g})$. These probabilities were set *before* any experience of norks.

Normally, the philosopher continues, our background knowledge sets the prior probabilities. We might, for example, know that an animal species usually has common traits like color, so someone observing only black crows would conclude most crows are black. If norks were just a newly discovered animal, we could say $P(g) > P(\tilde{g})$. Unfortunately, norks are located on a strange planet in a far galaxy, and are dissimilar to anything previously encountered. They are as close to being 'entirely new and strange' as one could want, and so our background knowledge is of little use. The normal pattern for norks may well be to have different colors on either side of a mountain range; no one knows. Of course, hypotheses like g come to us more naturally than $\tilde{g}$, but that says more about us than about norks.

The AI programmer asks whether even though norks are almost wholly new and strange, we still know enough to expect nature is usually uniform. In that case, g would have a prior edge over $\tilde{g}$. The philosopher, however, points out that this would merely postpone the question of selecting priors under completely novel circumstances. If our bias towards uniformity is itself a product of learning, this implies a previous state of even less knowledge. If not, it is an arbitrary foundational principle.

Now thoroughly confused, the survey team sends the robot down to the other side of the mountains, and find nothing but green norks. But this does not quiet their worries. There are many more areas they have not explored, and their philosopher can make a similar argument about the unseen norks in any of these areas.

2.2. Implications for subjective Bayesianism

This story is full of morals for Bayesian inference. First of all, subjective Bayesianism, which is perhaps the most popular Bayesian philosophy, fails to sustain learning in novel conditions. Leaving the choice of priors arbitrary is not good enough.

There is, of course, much that is plausible about subjective Bayesianism (Howson and Urbach, 1993). Individual statisticians bring differing backgrounds and hence different expectations to a problem, which cannot be ignored. And equation (1) ensures prior probability distributions which are not wildly different will quickly converge on near-identical posterior probabilities with accumulating evidence. In most practical applications, where statisticians start with roughly similar prior expectations, subjective Bayesianism will do fine. Even starting from very different priors, provided they have no zeros, Bayesians will converge in the long run.

There are also well-known difficulties with subjective Bayesianism (Kyburg, 1983; Mayo, 1996). The 'long run' can be indefinitely long, so there can always be Bayesians with different enough starting points to disagree after any amount of evidence. Let c stand for a hypothesis that is absurd in light of the evidence from modern science, s, so that $P(s|c) = \delta \ll 1$. A proponent of c might, however, come from a cultural background where c is considered almost certain, so he starts with $P(c) = 1 - \epsilon$, where $\epsilon \ll \delta \ll 1$. In that case, $P(c|s) \approx 1 - \epsilon/\delta \approx 1$. Such scenarios are not entirely unrealistic, since the scientific community does clash with communities with radically different conceptions of the world. In the United States, for example, c could represent the religiously-inspired creationist claim that biological evolution did not occur and that Earth is only a few thousand years old. Creationists, who are a major nuisance for US science education, are also not averse to accusing scientists of illegitimately assuming nature is uniform (Edis, 1997). Subjective Bayesianism, with its arbitrary priors, appears incapable of ruling out even ideas as absurd as creationism.

Goodman's paradox amplifies the subjective Bayesian problems with priors. In the original paradox, no evidence gathered before 2010 distinguishes between green and grue, so there is no possibility of progress and eventual convergence. In the spatial version, no amount of observation on one side of the mountains tells the survey team about norks on the other side. Everything depends on the priors; and subjective Bayesians make no rational distinction between an AI programmer setting $P(g) > P(\tilde{g})$ and a skeptical philosopher preferring symmetry and hence $P(g) = P(\tilde{g})$. Subjective Bayesianism falls into such impasses

too easily; it cannot be the centerpiece of either an account of genuine learning or of a philosophy of science.

2.3. Does objective Bayesianism do better?

Since arbitrary priors invite inductive skepticism, perhaps the solution is to supplement equation (1) with a principle for determining priors in novel situations. Attempting to remove the taint of subjectivity from Bayesian inference, objective Bayesian approaches take this path. There are a number of equivalent ways to formulate a principle for determining uninformative priors, focusing on marginal distributions, right-invariant Haar measures and so forth (Berger, 1985). The most appropriate in the context of Goodman's paradox is the maximum entropy formalism championed by E.T. Jaynes. We want to determine the starting probabilities prior to all experience. This suggests looking for a prior which reflects the fact that we are maximally ignorant, possessing no empirical information. *We should assume the least about the unknown.* But in that case, we can use information entropy as a measure of missing information. We maximize the entropy

$$H = -\sum_j P(h_j) \log P(h_j) \tag{3}$$

subject to constraints if some knowledge about the problem is available. When hypotheses are parametrized by continuous variables, the sum becomes an appropriate integral (Jaynes, 1968).

This is an attractive picture. In the spirit of pure Bayesianism, empirical data come into play only through equation (1), while the initial prior is objectively determined by maximizing ignorance. Of course, this prior is often practically impossible to compute for real-world statistical problems. This limits the range of application of maximum entropy methods, though it is invaluable in certain statistical mechanics problems (e.g. Edis, 1993) and as a technique for signal analysis and image reconstruction (Wu, 1997; Erickson et.al., 1998). Even so, it is clear that maximizing H favors broad distributions. When one of the $P(h_j)$ is much larger than the others, H is lowered; after all, narrowing the possibilities in this way implies increased knowledge about the state of a system. Therefore, even when an uninformative prior cannot be exactly calculated, we can rule out absurdly narrow priors such as $P(c) = 1 - \epsilon$ in the creationism example. Objective Bayesianism, in other words, promises to avoid the problems generated by arbitrary priors, even if in practice it would serve mainly as a way to prevent wildly different starting points for a more subjective style of Bayesian inference.

Unfortunately, objective Bayesianism fails to get around Goodman's paradox. Consider, once again, the problem faced by the survey team who have just encountered norks. The natural set of hypotheses to consider are those which state the color of every nork on the planet, such as g, b, $\tilde{g}$ and so forth. Since the norks are entirely unfamiliar, H has to be maximized without any constraints. This results in a uniform distribution: $\forall i, j\, P(h_i) = P(h_j)$. In particular, $P(g) = P(\tilde{g})$. Objective Bayesianism would seem to suggest the survey team can learn nothing about unobserved norks from the ones they have already seen.

In fact, the problem runs deeper. The above treatment of the h_j gives each color distribution equal weight. An alternative might be to assume the norks were drawn from a larger population in such a manner that each nork has a probability θ of being green and $1-\theta$ of being blue. In that case, our hypotheses would be parametrized by θ, and we would need to maximize the entropy of $P(\theta)$. Jaynes (1968) finds that the prior is the irregular distribution $P(\theta) \propto [\theta(1-\theta)]^{-1}$. This gives more weight to more uniform color distributions, resulting in $P(g) > P(\tilde{g})$.

In other words, contrary to Jaynes, maximizing entropy does *not* produce a distribution reflecting complete ignorance. The procedure relies on the availability of a background theory which specifies the appropriate statistical model. It is only in this context that we can speak of uninformative distributions for the model parameters and thereby unambiguously assign initial probabilities to hypotheses. This is why maximizing entropy can work in statistical physics or image reconstruction; in statistical physics, for example, the density of states is determined by quantum mechanics, leaving no room for uncertainty about the weight for each configuration. The problem with the norks involves not only a lack of knowledge of their precise color distribution, but also a lack of a background theory which could fix the prior. All objective Bayesianism does here is state more rigorously how background knowledge is to be reflected in the prior; the problems the philosopher posed for the survey team still remain.

It might seem that an objective Bayesian might just need to take one step backward to solve this problem. After all, background theories themselves can be assigned probabilities, so we might try to maximize $-\sum P(T) \log P(T)$ over all possible theories T in order to account for an uncertainty of background. But now, even if we could conceive of a way to approximate this quantity, the question becomes what could serve as a theory of all theories which will assign weights to the T's. It is tempting to think simplicity might be the key, following Ockham's Razor and the common intuition that simpler theories are more likely to be true. However, it is severely doubtful that the requisite experience-independent measure of simplicity could be found, especially given

the idea's vulnerability to skeptical arguments at this point (Sober, 1994). This would also be a curiously Platonic turn for a philosophy of inductive inference, since most of what is interesting about the world becomes inherent in the logically predetermined priors, reducing the function of empirical observation to eliminating some possibilities.[2]

So it appears objective Bayesianism cannot get around inductive skepticism either. In fact, no pure form of Bayesianism can, since the root of the problem is not with any particular method of assigning priors but with the whole notion of setting priors independent of all experience. We must go beyond pure Bayesian approaches to account for learning in truly novel circumstances.

3. Representative sampling and generalization

3.1. Back to basics

Assigning probabilities to an open-ended set of hypotheses is not safe. So perhaps other interpretations of probability, particularly the frequentist, will offer clues about how to proceed. While classical statistical inference is just as vulnerable to inductive skepticism, frequencies often provide a more intuitively plausible conception of probability.

In that case, let us go back to basics. Consider a bag containing one white and one black ball, from which we blindly select one. Few would quarrel with a statement that there is a $\frac{1}{2}$ probability of picking either ball. A frequentist can observe that in the long run, the relative frequency of white balls in a series of trials will approach $\frac{1}{2}$. Someone who thinks of probabilities as propensities could agree that the physical situation is such that someone will pick either ball one half the time. And an entropy-maximizing Bayesian will assign equal probability to white and black since we possess no information prompting us to favor one over the other.

Among the various options, however, the epistemic probability which Bayesians use is the most resilient in the face of skeptical objections. To begin with, propensities depend on the background theory. It is easy to imagine that, for example, the white ball might have a more pleasing texture, so that it will be selected more often. We could make sure conditions are completely symmetric except for color, controlling all the relevant factors we know of. But even then, perhaps the person picking the ball has an occult affinity towards white, so that unbeknownst to all, it is a virtual certainty that they will select the white ball. Though an absurd possibility, this emphasizes how judgments of propensity make no sense outside of a fairly solid background theory. Propensity

notions, however useful in context, will not help us understand learning in entirely novel conditions.

A frequentist view has a similar problem. The white ball is one out of a population of two; this notion of a relative frequency is straightforward. But frequentism, to the extent that it treats single-event probabilities as meaningful at all, ties them to frequencies in an idealized infinite series of trials. The long run frequency will approach $\frac{1}{2}$ if the balls are drawn at random; in other words, if the trials produce a random series. Normally, we could infer that a series is most likely random if we have a reliable background theory to that effect, or if we can generalize from observing an algorithmically random finite series. This, of course, invites skeptical objections. Frequentism implicitly assumes either background knowledge or that we can extrapolate; it does not get around inductive skepticism.

If, however, probabilities express degrees of belief, they need not assume knowledge about the ball selection process. We draw one ball out of a population of two, where we are completely ignorant about any biases in the physical selection. Unlike the example of the norks, there is no ambiguity concerning the population we sample, hence maximizing the entropy directly gives the single-event probability. Another way to see this is to notice that maximizing the entropy when setting probabilities for the balls concerns *our* ignorance. We cannot rule out an occult affinity for white, but we do not know about an affinity for black either. The symmetry between white and black is entirely due to our ignorance, not any known or assumed properties of the selection process.

The balls-in-bag example, though trivial, highlights how probability judgments are on solid ground when they refer to sampling definite populations. Both Bayesian and frequentist interpretations of probability run into difficulties when extending this basic, so-called 'classical' notion of probability—Bayesians to ill-defined populations corresponding to hypothesis sets, and frequentists to idealized infinite series. This suggests that to understand learning about norks, we should focus on real populations and the process of drawing a sample.

In the nork problem, the expedition has surveyed a thousand out of a million norks. As any pollster knows, almost all samples of this size are *representative* of the population; a randomly selected sample is very likely to reflect the characteristics of the whole. This is, of course, a much-used principle of statistical reasoning, and it also has previously been argued to help demonstrate the rationality of induction (Stove, 1986).[3] So observing that the sample the expedition took is likely to be representative is a vital step towards resolving Goodman's paradox.

That almost all samples of a thousand out of a million are representative is easy to illustrate. Using the standard hypergeometric probability distribution for sampling without replacement, we can calculate that if exactly half the total population were green, approximately 99.86% of samples of a thousand will show between 450 and 550 green norks. The likelihood that the sampled frequency is within ± 0.05 of the frequency in the general population is even larger if the proportion of green norks is less than or larger than one half.

Such frequencies of adequately representative samples can straightforwardly be related to our basic balls-in-bag probability. Since the expedition was completely ignorant of the colors of norks, and thus also of any bias in their sampling process, the simple maximum-entropy argument applies: they must assign equal probability to all the possible thousand-nork subsets. It is as if they had all thousand-nork combinations out of a million in a bag, and picked one arbitrarily. Therefore the sample they obtained is likely to be representative of the color distribution of the whole population. So the robot should, in fact, generalize so as to expect that all norks are green, rather than that they change to blue on the other side of the mountains.

To clarify this reasoning, it will help to note the differences between learning about the norks and the more familiar type of statistical sampling problem. The symmetry between samples of norks is entirely due to initial ignorance. Nothing about the total nork population need be assumed; not that their colors are independent, nor that their colors are randomly distributed so that nothing but the overall frequencies of colors usefully describes the population. Though formally similar, learning about norks is not the same as, for example, estimating the proportion of defective widgets given a random sample from a production run, where a definite statistical model is assumed and a single parameter is estimated.

This ignorance means the expedition cannot rely on the procedures statisticians use to ensure a sample is representative. In particular, randomization makes no difference. Discussions of statistical inference often conflate two different senses of randomness. The first is that of patternlessness, the way a series obtained through flipping a fair coin has no infinite subset predictable by any algorithm (Chaitin, 1987). Another common use of "random" refers to the arbitrariness of the sampling process, so that each subset of the whole population was equally likely to be chosen. Indeed, using a table of random numbers is an excellent way to achieve this arbitrariness. However, in the case of the norks, these two senses need to be distinguished. Ignorance, or the fact that norks are entirely new and strange, is sufficient to ensure the sample is arbitrary. It is, of course, possible that the sample is biased—

that the robot happened to land in an unrepresentative pocket of green norks. And one reason why Goodman's paradox has such force is that the sample is clearly not random in the sense of being patternless, because it is restricted to one side of the mountains. It seems plausible that nork colors might in fact have a pattern of dependence on mountain ranges; it would have been much safer to randomly choose the norks and get a sample from much more widely scattered locations. But if the nork example is to stand for a completely novel situation, we cannot rely on on this bit of background knowledge either. *Any* sample is equally likely to hit an unrepresentative pocket; the robot's clustered observation and a randomly generated sample are no different in this regard. The same applies to the original, temporal version of Goodman's paradox. If we know *nothing* about the time dependence of emeralds, the fact that our sample is confined to the past is irrelevant: it is as likely to be representative as a case where we could gather a sample which was better scattered throughout spacetime.

Another aspect of the difference between the nork problem and sampling widgets is also worth noting. Obtaining information from a sample immediately brings up the question of quantifying the strength of support for the robot's generalization, and leads straight into the thicket of debate pitting classical confidence intervals for parameter estimates against their Bayesian analogues. Our minimalist version of probability will not help resolve such matters. And since the expedition has no prior knowledge that the overall frequency of nork colors is the only useful descriptive variable, it is more appropriate to conceive of the robot's task in broader terms. As it examines the norks, the robot must be ready to generalize from other patterns and explore causal links which will bind knowledge about norks more securely into an overall web of knowledge. Obtaining an increasingly accurate estimate of any single parameter describing the nork population is too limited a task; instead, its learning must remain part of task of building a more complex theoretical understanding. Generalization from a likely representative sample, however well-justified, is still but a first step in this process. Hence attaching a precise number to its confidence in the generalization would be more misleading than helpful.

3.2. Breaking the symmetry

At this point, we might wonder if the argument from representative sampling cheats somewhere. It seems not to do justice to the compelling symmetry in Goodman's paradox, where it seems like whatever justification for a generalization to g we come up with can be stated equally well in terms of $\tilde{g}$.

Consider the following objection the expedition's philosopher might advance. During its exploration, the robot made a list with entries like "the n'th nork seen was G," with G for green, B for blue and so forth. But, the philosopher says, the robot could just as well have been programmed to follow a different labeling scheme which replaced G with $\tilde{G}$ for "green on this side of the mountains, blue on the other side," and B with $\tilde{B}$. Though eccentric, this would be a perfectly equivalent way of recording the same information. But then, if such a differently programmed robot had observed the same group of norks as the original, it should have generalized to $\tilde{g}$. After all, prior ignorance of nork colors goes for the $\tilde{G}$ and $\tilde{B}$ labels as well, so the sample taken is most likely representative. Since the frequency of $\tilde{G}$'s in the sample is almost certainly close to that in the overall population, the robot can confidently pick $\tilde{g}$ over g.

This objection, however, is flawed. Imagine that the philosopher had instead declared she had some peculiar but certain knowledge about norks: that green norks on the explored side of the mountain have a small hole on top, but that blue norks, if any, lack this hole. On the other side, however, only blue norks have the hole and not the green. The expedition now sends the robot down to the *same* side of the mountains it previously explored, confirming that the norks are green, and that, looking closer, they have small holes on top. Yet if the philosopher's knowledge was accepted as certain, it is clear that this new trip to the surface did not acquire any new information. In particular, she cannot claim that since all norks observed have had holes on top, the robot should generalize to expect norks all have holes, which means they are blue on the other side of the mountains. As far as the holes are concerned, the second sample from the same side is completely biased and wholly uninformative. The correct inference is still the generalization from the first, arbitrary sample. Introducing the strange information about holes after the fact changes nothing.

The switch between G and $\tilde{G}$ is exactly parallel, since the 'new' character $\tilde{G}$ is linked to G and the mountains by definition, and thus with certainty. And again, it makes no difference. *After* any given sample S is taken, the philosopher can always identify a G' such that G' and G coincide for an $S' \supset S$ but diverge for the rest of the population, and claim that a different generalization based on G' is justified. But then she can no longer claim S was arbitrary. Another way of expressing this asymmetry is to notice the entropy-maximizing probability distribution expressing ignorance about the G-properties of the population is *not* identical to that for complete ignorance about properties labeled by G'. The change from G to $\tilde{G}$ was not a neutral relabeling; it assumed information which was not available.

3.3. Evil Demons

This goes a long way to resolving Goodman's paradox, but not yet the full distance. The symmetry between g and $\tilde{g}$ is broken, after all, for what amounts to historical reasons. Though after-the-fact changes do not matter, it appears justified generalization depends critically on knowing beforehand that G and B were appropriate for describing objects rather than $\tilde{G}$ and $\tilde{B}$. At the least, there is a sense in which our prior ignorance about new and strange objects like norks is not quite complete. For *us,* G corresponds directly to how we perceive objects, and we legitimately express our ignorance in these terms. But this dependence on how *our* minds work is disturbing; if an alien mind to whom $\tilde{G}$ came naturally had observed the same sample of norks, would it have been justified to generalize to $\tilde{g}$?

This question does not directly challenge the inference that norks are probably green. For example, the expedition could construct something like this alien mind by inserting a small box into the circuit connecting the robot's camera with its CPU. Inspired by thought experiments about inverted qualia, this box would convert green signals from the camera into blue and vice versa, but it would be activated on only one side of the mountains and not the other. Now, the robot would expect to continue to see green norks; translated into the expedition's terms, this means generalizing from $\tilde{G}$ observations to $\tilde{g}$. But this does not restore the symmetry between g and $\tilde{g}$. It is clear the robot's very hardware deceives it about norks; the inverting box was designed with prior knowledge of the normal process of taking samples, just to make it biased. However, what if the expedition lacked this knowledge? The survey team knows a lot about observation. Engineers can easily find the inverting box, and neuroscientists can certify that the human optical nerve has no such impediment. But if not handed down from heaven, this knowledge must itself have been learned somehow. So we must ask where we would stand if we were truly facing *entirely* new circumstances. In that case, we would be in the same boat as the robot; all we have to go on are our perceptions and not direct access to reality, so it becomes a serious question whether we, like the robot, are systematically misled by our perceptual wiring.

This is reminiscent of the species of skepticism which wonders if our senses are deceived by an Evil Demon. Indeed, Goodman's paradox can be viewed as an ingenious combination of Humean and Cartesian skepticism. This brings up a host of traditional philosophical issues concerning Evil Demons and how language represents reality, as well as new questions concerning the ways we classify observations recently raised by philosophers (Hirsch, 1993). Though it is impossible to ad-

dress many of these here, the notion of evidence being representative points a way out of this version of skepticism as well.

We can begin with a standard coherentist response to Evil Demon scenarios. A believer in a Demon would be so comprehensively paranoid, she would have no hope of giving reason even for her paranoia. In contrast, trust in our cognitive abilities is a self-supporting belief (Lehrer, 1990). We can take it as a starting point for constructing theories and see what happens—including, possibly, revising our views so as to converge on paranoia.

This is correct. Trial-and-error is the only way to begin writing on a blank slate; in building a robot which does not know even the ways to classify and label objects, all we can do is start with some arbitrary perceptual equipment and test it. We need, however, to put some backbone into this argument, especially into the notion of support.

Consider, first, learning when we do have something to build upon. Generalizing about norks did not rely on prior knowledge about the nork population to fix the appropriate statistical model. So not only is there a possibility of failure as in any statistical inference, but we must be responsive to systematic failures which inform about the nature of the nork population. Imagine that norks were in fact blue on the other side of the mountain, and that the expedition discovers this by a comprehensive investigation. Then they visit a string of planets studded with norks, and all exhibit a pattern of blue and green on either side of mountain ranges. In that case, the survey team could not expect nork colors to be uniform on the next planet. They can now also generalize at the level of the population of nork-bearing planets, expecting a two-color pattern. In other words, the expedition is no longer ignorant about norks, and so looking only at one side of any mountains can no longer be considered arbitrary sampling.

Now imagine that even on earth, practically every green population turned blue over barriers. Trees switch between blue and green on crossing a river; the neighbor's lawn looks blue beyond the fence; the same spectral lines look blue or green depending on the star the light comes from. In such a world, the higher level generalization would be very broad indeed, and in fact, newly encountered norks would themselves be expected to conform to this pattern. All our color-regarding behavior will be influenced by this deep and persistent pattern; it will become second nature to us so much that a forest which *did not* change color on different sides of a river would strike us as a strange anomaly. Observing a cluster of green objects, we would generalize to expect they will be blue beyond any mountains. In effect, we will have learned $\tilde{G}$ and $\tilde{B}$ are the correct ways to describe objects, since this produces successful generalizations.

Let us now apply this rough sketch of learning to a blank slate of a robot, which starts out with less smarts than an exceptionally stupid bacterium. It would be more appropriate to conceive of its learning in more pre-theoretical, hardware terms. Its first task, after all, is to get its perceptions lined up with reality so that it can generalize in matters important to it. So we should think of a population of self-replicating robots, with very basic but varyingly wired perceptual systems. Their reproduction depends on correct inferences; they might, for example, thrive among plants and die in water, and have only visual information to help them infer which area is which. So it is vital they perceive colors in a way which will successfully generalize. In a world where objects keep switching between our blue and green, it would be most efficient if the robots directly see the world in $\tilde{G}$ and $\tilde{B}$ terms, rather than compensate with a layer of software-style learning.

There is no need to belabor the point; it is abundantly clear a process like Darwinian evolution can produce well-adapted, cognitively sophisticated creatures who successfully generalize and exploit regularities in their local environments. The question is, of course, whether any of this matters; it looks too much like the classic mistake of circularly justifying induction by its sparkling record of past success.

The difference is in a matter of degree. To the inductive skeptic, the unknown is unknown, period; the efforts at generalization of a super-stupid robot and a sophisticated survey team are much the same. However, if evidence is representative rather than only something which eliminates incompatible hypotheses, the generalizations available to a robot close to a blank slate are much weaker that those of one which has adapted to a complex local environment. The blank-slate robot almost certainly will start with a string of failures, adapting very crudely at first. It is trying to extrapolate from a very small sample of its environment, and so this will be unreliable. The survey team and the well-adapted robot do much better. They construct many interlocking layers of generalization, including inferences based on the frequency of success which would be expected from correct perception of regularities; they draw on multiple independent lines of evidence that they have an accurate picture of their local environment. They not only have a much larger sample but multiple convergent ways of extrapolating from it.

This amounts to a search for coherence; not as a vague hanging-together but as a means of reality-testing. A well-adapted robot will achieve a cognitive picture tuned for successful local inference. So for such a robot, or the survey team, norks are not as much an unknown as for a robot close to a blank slate. They can study norks with less of an expectation of basic readjustments of perception and with more confidence that they can classify and generalize properly.

Of course, all this is still based on learning about our local pocket of space and time; the inductive skeptic can still ask why things should behave similarly elsewhere. Perhaps norks really are blue on the other side. Perhaps an Evil Demon will turn the sky green and the grass blue tomorrow, just to play a bad joke on us. Or maybe emeralds will turn blue at 2010, because they were grue all along. But we can legitimately expect otherwise. The ignorance of the survey team meant they should treat their sample of norks as representative. Similarly, in the absence of information to the contrary, our well-explored local pocket of spacetime can be taken as representative. This is no guarantee; the survey team should be surprised if the norks are blue over the mountains, but not amazed. Nevertheless, norks are not a complete unknown for them. And we can legitimately ignore the Evil Demon and grue emeralds as serious probabilities.

4. How we learn

Normative accounts of learning are often closely connected to the philosophy of science. After all, science is our best example of nontrivial learning; not only because it is compelling in its depth and detail, but because it has on occasion radically revised certain of our commonsense generalizations about the world. So questions of learning are often posed in a context of trying to understand science, and inductive skepticism surfaces in problems like the underdetermination of theories by evidence.

For many, the key issue has been to find a method or principle to sort out the best theory in play, given the evidence. But attempts rooted in ideas like positivism, logical probability, or falsificationism, have always come up short. And though such accounts should have been normative for low-level machine learning as well as the cognitive heights of modern science, no AI programmer could send a robot out into the world equipped only with falsificationist principles.

Still, since the alternative seems to be a dismal pragmatism which reduces science to a 'way of coping' when not flirting with outright cognitive relativism, the need for basic principles remains strong. In the past few decades, many philosophers have looked to statistics to underwrite science. After all, statistics is all about rigorous reasoning under conditions of uncertainty and incomplete information, and has a large and intimidating mathematical literature. Its techniques are in widespread use in the sciences. Best of all, its norms can be applied practically. Bayesian statistics, especially in a decision theoretic context, seems a candidate 'gold standard' for machine representations of

uncertainty; psychologists can study everyday human reasoning and even animals to see how well we follow the rules of statistical inference (e.g. Cosmides and Tooby, 1996; Brase et al., 1998).

However, even here it is all too easy to find openings for inductive skepticism. Statistical inference is a deductive procedure to draw conclusions from data and a given statistical model; it is no use outside an overall iterative approach involving inductive steps to decide upon appropriate models (Leonard, 1980; Leonard and Hsu, 1999). The inductive element depends on the expert judgment of the statistician; it is not captured by the mathematical apparatus. Though there is much debate within and between various schools of statistics on this issue, none provide an adequate account of model formation, and all are vulnerable to the problems illustrated by Goodman's paradox.

Therefore, inductive skepticism must be defused at the level of rudimentary machine learning rather than that of statistical methods. At first, this does not look promising. Even in representing uncertainty, AI departs from ideals such as the Bayesian picture, treating it as one among many tools thrown together for practical reasons (Krause and Clark, 1993). And examining human reasoning as well as ideas about artificial analogues, we encounter the literature on abduction and hypothesis-making. On one hand, abduction is often considered a source of creative insight, which must involve random generation of novel approaches (Edis, 1998). It also, however, incorporates basic attempts at inference, making connections and extrapolations through analogy and generalization. This aspect of abduction is subject to criticism and justification (Burton, 1999). In short, 'abduction' is a vague concept, describing a motley collection of tools which might seem to need justification in terms of approximating a more rigorous principle of learning.

Nevertheless, much in basic abductive reasoning naturally combines inference and model-making. We form hypotheses because we think they are likely. In fact, we never consider the vast majority of possibilities compatible with the data, ruling out almost all competitors to our hypotheses from the start. This is done largely through classification and generalization, in which the common theme is pattern recognition. Finding patterns and extrapolating from them comes to play both in drawing conclusions in novel circumstances and in introducing new levels of generalization for more complex models. Understanding why this works, then, should take the sting out of inductive skepticism, turning it into a challenge to spur learning about learning rather than an objection which threatens to derail the whole enterprise.

The key is to realize evidence is representative rather than purely eliminative. Borrowing ideas of sampling and maximizing ignorance

from statistics, we can justify this statement. As a result, we *do* expect uniformity and simplicity in nature, but not because this is a prior assumption or a metaphysical requirement of Reason. This could not work as a fundamental principle anyway, as the symmetry between G and $\tilde{G}$ in Goodman's paradox makes clear: uniformity of what? Our notions of simplicity are not logically predetermined; they take shape in adapting to our environment, and we expect this to continue to succeed on grounds of representativity. And representativity derives from nothing but an accurate description of our ignorance.

None of this diminishes the usefulness of statistics. It does, however, suggest a change of emphasis. Alongside using statistical norms to guide work on machine learning, we should also open statistics to the influence of ideas concerning machine learning and abduction. Statistics cannot generate all-encompassing norms of learning; we should instead conceive of it as occupying a middle ground between broad analyses of science and investigations of more basic elements of learning. Tools like generalization are not just primitive echoes of statistics; they set the stage for statistical inference.

In that case, the task that beckons is to forge closer connections between various ways of thinking about learning. Bayesians, for example, might take their perennial problem with priors as an opportunity to try and set priors to better reflect available instruments of learning, particularly sampling. One possible way departs from 'empirical Bayesian' approaches which deal with hierarchical models, using existing data to revise priors (Berger, 1985; Schervish, 1995). Such methods are also useful in interpreting Bayesian probabilities in a more frequentist fashion (Efron, 1996). Another direction to explore is amending Bayesian inference to accommodate imprecise probabilities (Walley, 1991), since precise distributions on parameter values are inappropriate when exploring novel situations and the statistical model is not fixed. There is much research to be performed here, both from a statistical and an AI perspective.

Another benefit of progress on this front would be a more unified view. Theories of science and learning based on probabilistic reasoning have hitherto been disconnected from accounts which emphasize coherence, whether they happen to be inspired by epistemological reflection or by the properties of neural networks. Integrating statistics with the perspective of abduction and model building would introduce a coherentist element, since this appears to be vital to understanding how generalization works.

Meanwhile, we might at least put some of the more extreme forms of skepticism behind us. We expect grass will remain green tomorrow, and there is more to this confidence than an act of animal faith.

Notes

[1] Hume and Kant, though in different ways, suggested induction was rational by virtue of the way our minds work. Many philosophers have been attracted to rooting epistemology in psychology or the norms of a culture ever since. However, this is not satisfactory unless there is something magical about our minds so that they participate in some kind of transcendent validation of induction. Hence the appeal of arguments such as Grayling (1985), which proposes a version of idealism.

[2] Regardless of whether a Bayesian approach is adopted, Goodman's paradox cannot be resolved by conceiving of induction in terms of elimination rather than enumeration and generalization as does Papineau (1993). That eliminating alternatives is not sufficient is precisely what Goodman's paradox makes clear.

[3] While Stove (1986) does emphasize the key role of representative sampling, he also misdescribes the inductive skeptical claim as $P(h|e) = P(h)$ for *all* e and h (p. 40), not excluding cases where $e \vdash \neg h$. He also burdens his overall argument with a logical probability framework, and does not adequately address Goodman's paradox.

References

Berger, J.0. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, New York: Springer Verlag.

Brase, G.L., L. Cosmides, and J. Tooby (1998) 'Individuation, Counting, and Statistical Inference: The Role of Frequency and Whole-Object Representations in Judgment Under Uncertainty', *Journal of Experimental Psychology: General* **127**:1, pp. 3–21.

Burton, R.G. (1999) 'A Neurocomputational Approach to Abduction', *Minds and Machines,* **9**:2, pp. 257–265.

Chaitin, G.J. (1987) *Algorithmic Information Theory,* Cambridge: Cambridge University Press.

Chisholm, R.M. (1989) *Theory of Knowledge,* 3rd edition, Englewood Cliffs: Prentice-Hall.

Cosmides, L., and J. Tooby (1996) 'Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgment Under Uncertainty', *Cognition* **58**:1, pp. 1–73.

Cox, R.T. (1961) *The Algebra of Probable Inference,* Baltimore, Johns Hopkins University Press.

Davidson, D. (1983) 'A Coherence Theory of Truth and Knowledge', in *Kant oder Hegel?,* D. Heinrich, ed., Stuttgart: Klett-Kotta.

Edis, T. (1993) 'Unusual Constraints in the Quantum Statistical Mechanics of Josephson Junction Systems', *Journal of Statistical Physics,* **71**, p. 313.

Edis, T. (1997) 'Relativist Apologetics: The Future of Creationism', *Reports of the National Center for Science Education* **17**:1, pp. 17–24.

Edis, T. (1998) 'How Gödel's Theorem Supports the Possibility of Machine Intelligence', *Minds and Machines,* **8**:2, pp. 251–262.

Efron, B. (1996) 'Empirical Bayes Methods for Combining Likelihoods', *Journal of the American Statistical Association,* **91**:2, pp. 538–565.

Erickson, G.J., J.T. Rychert and C.R. Smith, eds. (1998) *Maximum Entropy and Bayesian Methods: Proceedings of the 17th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis,* Boston: Kluwer.

Giere, R.N. (1988) *Explaining Science: A Cognitive Approach,* Chicago: The University of Chicago Press.

Goodman, N. (1983) *Fact, Fiction, and Forecast,* 4th edition, Cambridge: Harvard University Press, chap. 3.

Grayling, A.C. (1985) *The Refutation of Scepticism,* La Salle: Open Court.

Hirsch, E. (1993) *Dividing Reality,* New York: Oxford University Press.

Howson, C., and P. Urbach (1993) *Scientific Reasoning: The Bayesian Approach,* 2nd edition, Chicago: Open Court.

Jaynes, E.T. (1968) 'Prior Probabilities', *IEEE Transactions on Systems Science and Cybernetics,* SSC-4, pp. 227–241.

Krause, P. and D. Clark (1993) *Representing Uncertain Knowledge: An Artificial Intelligence Approach,* Dordrecht: Kluwer.

Kyburg, Jr., H.E. (1983) *Epistemology and Inference,* Minneapolis: University of Minnesota Press.

Leake, D.B. (1995) 'Abduction, Experience, and Goals: A Model of Everyday Abductive Explanation', *Journal of Experimental and Theoretical Artificial Intelligence,* **7**, pp. 407–428.

Leonard, T. (1980) 'The Roles of Inductive Modeling and Coherence in Bayesian Statistics', in J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith, eds., *Bayesian Statistics I,* Valencia: University Press.

Leonard, T., and J.S.J. Hsu (1999) *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers,* New York: Cambridge University Press.

Lehrer, K. (1990) *Theory of Knowledge,* Boulder: Westview, pp. 121–124.

Mayo, D.G. (1996) *Error and the Growth of Experimental Knowledge,* Chicago: University of Chicago Press.

Papineau, D. (1993) *Philosophical Naturalism,* Oxford: Blackwell, pp. 161–167.

Prado, C.G. (1987) *The Limits of Pragmatism,* Atlantic Highlands: Humanities Press International.

Schervish, M.J (1995) *Theory of Statistics,* New York: Springer.

Snow, P. (1995) 'An Intuitive Motivation of Bayesian Belief Models', *Computational Intelligence,* **11**:3, pp. 449–459.

Sober, E. (1994) 'Let's Razor Ockham's Razor', in *From a Biological Point of View: Essays in Evolutionary Philosophy,* Cambridge: Cambridge University Press.

Stove, D. (1986) *The Rationality of Induction,* Oxford: Clarendon, chap. 6.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities,* London: Chapman and Hall.

Wu, N. (1997) *The Maximum Entropy Method,* New York: Springer.