Induction, Samples, and Kinds

Peter Godfrey-Smith

Harvard University

1. Introduction

2. Goodman's Problem and Naturalness Constraints

3. A Second Form of Inference

4. A Nominalist Challenge

5. A Discussion of Cases

6. Conclusion

1. Introduction

This paper will criticize a familiar package of ideas about "inductive" inference, and use the criticism to motivate a different package.

"Induction" is understood here as a pattern of argument or method used to answer questions of the form: "how many *F*'s are *G*?" This question is understood as one about a proportion or frequency. So it could also be expressed by asking "what is the rate of *G* in the *F*'s?" The question "Are all *F*'s *G*?" is a special case. Examples of such questions include:

1. How many teenagers smoke?

2. How many ravens are black?

3. How many emeralds are green?

4. How many people in this room are third sons?

5. How many organisms have the amount of bases cytosine and guanine equal, and the amount of adenine and thymine equal, in their DNA?

6. How many electrons have charge of approximately - 1.6 x 10^{-19} coulombs?

There are lots of ways of answering such questions. In "induction," the questions are answered by noting the relation between *F* and *G* in observed cases and making some sort of extrapolation or generalization. This is presumably done with the aid of background knowledge. But the approach taken is one in which the number of *F*'s seen is supposed to be epistemically important. The classic case of inductive inference is the one where *all* observed *F*'s are found to be *G*, and this is used to conclude that all the unobserved ones are as well.

Here is a view that many people hold: induction is often rational, lest factual knowledge collapse. But as we learned from Nelson Goodman, the *F* and the *G* in a good induction *can't be just anything*. We need a constraint, probably some sort of "naturalness" constraint, on the kinds or predicates involved. Many philosophers would agree with this even though they do not agree what the constraint is or where it comes from.

I will argue that for the kind of inference that counts as induction in the above sense, and that can be generally justified, the predicates used *can* be just anything – or near enough to anything. In particular, a "naturalness" constraint has no basis. Naturalness does have a role in *another* kind of inference that can answer "how many *F*'s are *G*?" questions. When we get to that point you might say that other kind of inference is induction, too. So we run into an issue that is a bit terminological. But the overall view I will defend is that the familiar philosophical concept of "induction" has conflated two kinds of inference, each of which is successfully exploited by science. For each of the two inference patterns, an account can be given of its *in-principle reliability*. That account is a kind of philosophical justification. The package usually known as "induction" does not have that kind of justification, however. It combines elements from each method without combining parts that give rise to an in-principle reliable combination in its own right.

The paper also argues for some meta-epistemological ideas. Older work in normative epistemology often included trying to show why a method will or won't *work* – showing whether the method is responsive to the world in a way that leads to us getting the right answers. There was a retreat from this procedural approach in the 20th century,

when much work focused on logical relationships. In the case of induction, there was also a further role for Goodman's *Fact, Fiction, and Forecast* (1955). Goodman seemed to show the complete foolishness of procedural approaches that had been taken up to that point, and also seemed to show the coherence of a quite different orientation to the problem. This re-orientation holds that the normative analysis of induction works by means of a mutual adjustment of judgments about cases and judgments about general rules, with the aid of intuitions we have on both matters. This is the method Rawls (1971), crediting Goodman, christened as "reflective equilibrium."

I will criticize conclusions reached by that approach and argue for the value of another. In some ways the alternative is a revival of the earlier style, focused on procedures. But it is not foundationalist, and includes a kind of "model-building" orientation. In the case of induction, there are continuities between model-building that has a philosophical level of ambition and abstraction, and model-building of a more practical kind within science. To some extent, problems in the epistemology of induction are exchanged for problems about how model-like theoretical constructs relate to the world and our situation in it. But there is net progress. Reichenbach is a particularly important precursor of the approach I defend.[1]

2. Goodman's Problem and Naturalness Constraints

Suppose we are answering a "how many Fs are G?" question. We note how many F's we have seen that are G, and extrapolate. We feel better about it if we have seen many F's rather than a few.

Is this approach justified? Not always. We confront Goodman's "new riddle of induction" (Goodman 1955, Stalker 1994). Though in many cases such extrapolations are surely rational, the predicates used in our argument apparently can't be just anything, if

---

[1] Some might say at this point that other views about induction that might have once been "standard" are being replaced by a Bayesian view. The availability of a Bayesian approach does not affect the main arguments in this paper. I see Bayesianism, in this context, as something like what Strevens (2004) calls an "inductive framework," as opposed to an "inductive logic." The attitude to the relation between past and future experience exhibited by a Bayesian agent is determined by the agent's assigment of likelihoods – these contain the agent's opinions about what depends on what, and whether past events can make particular future ones more or less probable. It is compatible with Bayesianism that these assignments of likelihoods be inductively skeptical, or reflect counter-inductive attitudes towards experience. A Bayesian framework may be used to represent many different views about the relation between observed and unobserved, including the views in this paper.

the inference is to be a good one. This can be illustrated by introducing a "grue" predicate. Something is *grue* if and only if it either has been previously observed and is green, or has not been previously observed and is blue. If all observed emeralds are in fact green, we find we can use this data as the basis for two conflicting inferences about the future: the familiar inference leading us to expect newly observed emeralds to be green, and also an inference leading us to expect them to be blue. After all, the previously observed emeralds, being green, are also all grue. If we extrapolate grueness to the presently unobserved emeralds, this leads us to expect that those ones will turn out to be blue. We have encountered an apparent collapse of inductive methods: with suitable choice of grue-like predicates (in *F* and/or *G* position), just about anything in our past observations can be used to support just about any hypothesis about the future.

Goodman gave other examples of bad inductions, using less exotic language. One uses the predicate "is a third son": if we find the first few people in this room are third sons, that does not give us reason to think that everyone in the room is a third son. What seems needed is an extra constraint. Perhaps inductive arguments have to use predicates that pick out "natural" properties or kinds (Quine 1969, Lewis 1983). Many different bases for such a constraint have been offered, ranging from the conventional to the metaphysical, and the idea of a naturalness restriction on predicates and sets has spread into many other parts of philosophy, always with Goodman's "grue," and the problems it causes for induction, as the motivating example.

A paper that went against this trend was Frank Jackson's "Grue" (1975). He argued that in a good induction the predicates used *can* be just anything, provided that an additional premise is true. The extra premise he called the "counterfactual condition." An induction, for Jackson, takes information about objects that have *three* features – they are observed, they are *F*, and are *G* – and draws a conclusion about objects that are unobserved and *F*. A good induction looks like this:

J1. All *F'*s which are observed are *G*
J2. If those *F'*s had not been observed, they would still have been *G*.
Therefore (fallibly):
J3. All *F'*s are *G*.

The extra premise is true for *F*=emerald and *G*=green, but not when *G*=grue. To say this relies on background knowledge, but background knowledge that, according to Jackson, is reasonably available and does not beg the question. Most importantly, there is no restriction on predicates *per se*. The predicates used can be anything at all, provided that premise J2 is true. It is *hard* for premise J2 to be true with some predicates in the *G* position, given normal choices for the "observed", "past", or "sampled" predicate. But naturalness is not an issue, and linguistic "entrenchment" in Goodman's sense is not an issue; those constraints drop out of the story.

Jackson's argument is the beginning of the answer to Goodman's problem. It is not the whole answer. This is for both internal, detailed reasons and more general ones. First, the proposal does not handle all grue-like predicates (especially "emerose" cases). Some were handled in an modified formulation given by Jackson and Pargetter in 1980, and others were not. These problems are discussed in an earlier paper of my own (2004). Using an idea due to Alexis Burgess we can make more progress. In this discussion I will ignore those additional cases and assume we only have to deal with the original grue problem, where Jackson's 1975 proposal seems to work. But we can also ask a more external question: what is the *status* of Jackson's proposal? If it is right, what makes it right? And what does it then achieve for us?

What Jackson says is that if we put inductions into his form, we see that the good ones are those in which his premises are true, and the bad ones are those where some premise is false. His proposal handles the cases, and it is also the basis for a plausible story about what has gone wrong in the bad inductions. When premise J2 is false, it is *only because those F's were observed* that they were *G*; their *G*-ness is a special case, not something that can be extrapolated. The methodology here is the usual 20th century one: we take logic as far as it goes, and then use intuitions. But suppose someone now asks: if I go through the world applying the Jackson schema and drawing conclusions, will I do well or badly? What will be its consequences? Can you show me that at least in principle, or in *some* relevant range of circumstances, it should lead me to truth rather than error? Nothing about this is said by Jackson, or could be said within his framework. And in fact

we can do more. To see how, let us forget emeralds, and look at the kind of question that would actually be answered using simple extrapolation from a sample.

Suppose you want to know how many teenagers smoke. The obvious way to answer this question is to collect a random sample of teenagers, find the rate of smoking in the sample, and extrapolate to the larger teenage population, in a way guided by statistical measures of likely error. Why should we do this rather than something else? Why not find the proportion in the sample and extrapolate half that proportion, or one minus that proportion? (This second option would be a kind of counter-induction.[2]) *Not* because of an equilibrium between intuitions – or at least, not at this stage in the analysis. Instead, we have a statistical model of why the procedure is in principle a reliable one. The model tells us how samples of different sizes will be distributed, in relation to the actual properties of the population being sampled. It tells us when, and the extent to which, the properties of a sample are reliable indicators of the properties of the underlying population.[3]

I said that the question about teen smokers could be answered using inference from a sample. But there are various ways this method might fail. Maybe you cannot collect a random sample, as the smokers tend to avoid you. Perhaps teenagers will not tell

---

[2] For a finite population of *F*'s it would not make sense to do this in all cases; if you see a rate of 100% *G* in a sample, and generalize that exactly 0% of *F*'s are *G*, you have said something that your own observations show to be false. Other policies nearly as counter-inductive would be possible, though. Further, the conclusion reached by extrapolation from a sample is properly expressed by saying that the true value lies in an interval, whose size is determined by properties of the sample (see note 3). A counterinductivist might do something similar.

[3] If the true proportion of *G's* in the population of *F*'s is p, and the population is sampled randomly with samples of size N, then the number of *G*'s in the sample will be distributed binomially with parameters N and p. This "sampling distribution" will have a mean of pN and a variance of $Np(1-p)$. Let $p*$ be the observed proportion of *G*'s (the number observed divided by N). The sampling distribution of $p*$ will also be binomial, with a mean of p and a variance of $p(1-p)/N$. The observed proportion $p*$ is an unbiased estimate of p. An investigator will also calculate a "confidence interval" around $p*$, which will depend on $p*$ and the sample size. (The likely size of the difference between p and $p*$ depends on p itself, and this is being estimated from $p*$.) A 95% confidence interval might be used, in which case the claim made is that if samples were taken repeatedly and the investigator was to claim each time that the true value p lay within that interval around the observed $p*$, the investigator would be right 95% percent of the time. The philosophical significance of this sense of "reliability" may be questioned, especially for a real-world setting in which only one sample is taken. Here, as indicated in the Introduction, we face problems concerning the relation between an idealized model and real-world cases.

For problems of estimation of the kind discussed in this paper and using the binomial distribution, a 95% interval is often calculated as: $p*\pm1.96\sqrt{[p*(1-p*)/N]}$. This method is based on the fact that the binomial distribution approximates a normal distribution when N is large (the "Wald confidence interval"). Other methods are also used, which may be more appropriate for small values of N and small or large $p*$. (Eg., when $p*$ is zero or one the interval given by the rule above has zero size.)

you the truth. There is also a third, more unlikely possibility. Perhaps being asked the question tends to *make* some teenagers immediately take up smoking. So they truthfully answer that they smoke, but only because they were asked. The process of data-gathering is interfering with the objects you are observing, in a way that makes them an unreliable guide to the unobserved cases. This is not a case of "selection bias"; we can assume that the original collection of teenagers asked about their smoking really was a random sample. Some statisticians call this a "Hawthorne effect," after a famous case in the 1930s.[4] It also has a kinship with the notion of a "confounding variable," although that term is usually applied in the context of causal inference rather than estimation. The term "observation selection effect" is used in some literature to cover various phenomena, including sample bias, the confounding-like phenomenon discussed here, and maybe others. In any case, that special relationship between surveying and smoking would make the method fail. The problem has nothing to do with "non-projectible" predicates; it has to do with the process of collecting our sample and some unwelcome causal relations.

There is a close relation between this phenomenon and the grue problem. The case of the grue emeralds features a *non-causal analogue* of the problematic dependence relations seen in the smoking case. Just as the teens surveyed would not be smokers if we had not asked them about it, the emeralds would not have been grue – would not have counted as grue – if we had not observed them. The process of observation is interfering with the properties we are interested in. The semantics of "grue" turn observation itself into a confounding variable. Or to adapt a piece of metaphysical jargon, if the teenage-smoking problem was like a case of confounding, this is a case of *Cambridge-confounding*.[5] When we try to state a condition that would rule out such problematic dependence relations, and say it in abstract counterfactual terms, we will find ourselves saying what Jackson said in his paper on grue.

Neither Goodman nor Jackson said anything about randomness. They also offered no account of why following arguments that meet their requirements will do us any good. In cases where you *can* collect a random sample, however, there is a model that tells you how and why certain projections will be reliable. The model also tells us that you cannot

---

[4] Hawthorne is a place, not a person. The case is sketched in my 2004 paper.
[5] Here I draw on Peter Geach's notion of a "Cambridge change" (1972).

use random samples to answer grue-questions with grue-observations in the same way you can use samples to answer green-questions with green-observations. The problem arises *as a feature of procedures.* If you had a "non-interfering" way of sampling the emerald population, you *could* estimate the proportion of grue emeralds. How hard this is depends on the exact "grue" predicate used.

The view that emerges diverges sharply from intuitions. Suppose we can randomly sample, and we are keeping an eye out for the confounding role of observation that arises with some predicates. Let us go through the world, sample, and project. The *G'*s that the model allows us to project include *G*=(green or smaller than a fingernail). They also include: *G*=(jadeite or nephrite), *G*=(green and born in Iceland), and *G*=(green or identical to the number 4). This looks all wrong; our intutions revolt. But the model of reliable estimation from from samples allows it. There is no need for the predicates in an inference to be of a kind that can figure in natural laws. Returning to Goodman's "third son" case, there is really no problem here. If there is a rate of third-sonness in this room, and you collect a random sample of people and assay it for third-sonness, you *can* draw reliable inferences, depending on sample size and so on. The same is true on the *F* side – the population or class whose rate of *G*-ness you are interested in. It can be as arbitrary as you like, as long as it can be sampled.

3. A Second Form of Inference

The previous section described one kind of inference that meets the criteria given earlier for being "inductive," and discussed why the inferences are justifiable. This approach is inapplicable to many induction-like inferences, however. Lots of collections we are interested in cannot be randomly sampled. "Random sampling" here means that every member of the population you are drawing conclusions about has the same chance of making its way into the sample. So a collection containing future individuals (future ravens, future DNA molecules, future third sons) cannot be randomly sampled. It surely seems that we can sometimes gain knowledge of generalizations in such cases, however. In these cases, our observed instances are not a sample drawn from a total population, but are more like a subpopulation "attached" to it. The past, in particular, is attached to the future, not drawn from it.

In the sampling cases, the power of randomness is what gives us a "bridge" from observed to unobserved. In the second kind of case, the bridge – when there is one – is very different. If we want to make inferences about a population that cannot be sampled, we must ask: what *kind* of collection is this? Are these objects the products of a common origin? Do they have a common internal structure? What sort of causal relationship is there likely to be between properties we are projecting from and properties we are projecting to? There need not be "laws of nature" overtly on the scene here, but we are basing the inference on some kind of natural connection – some combination of laws, mechanisms, and etiologies.[6] I will say less about these inferences than I did about the first category, and will focus primarily on the contrasts between the two.

In the second category of inference, something like the "naturalness" of kinds and properties is central. It is *so* central, in fact, that a crucial feature of the first category drops out. That is the *number of cases* observed. In inferences from random samples, numbers are epistemically significant. Larger is always better. In the case of inferences of the second sort, based on causal structure and kinds, this is not so. In the purest examples of this sort of investigation, *one* instance of an *F* would be enough, in principle, if you picked the right case and analyzed it well. Ronald Reagan is supposed to have said "once you've seen one redwood, you've seen them all."[7] When something like this is true, it is a powerful basis for inference. In practice, one is usually not enough. Numbers do often play a role, but this role is different in character from what we find in the first category of inference.

---

[6] For discussions of "induction" which bear in different ways on this second kind of inference, see Boyd (1999), Millikan (2000), and Norton (2003).

   Looking further back, Harman (1965) claimed that all good inductions are really cases of inference to the best explanation in disguise. Inference from samples is not like this – you may believe that 20% of teenagers smoke, by inference from a sample, without having any idea why they smoke. There is a relation between explanatory inference and the second category of induction-like inference discussed here. One need not always infer to a particular explanation of the *F-G* association in a sample, however, to infer that unobserved *F*'s are *G*. Background information might make it likely that there is *some* suitable relation between *F* and *G*, making extrapolation reasonable, even though the particular relation is not known.

[7] *Eigen's Political and Historical Quotations* (http://www.politicalquotes.org/) makes this attribution, citing the *The New York Times Magazine*, July 4, 1976. An urban legends web page (http://www.snopes.com/) claims that it was attributed to him by Governor Pat Brown, and what Reagan actually said was perhaps relevantly different from the point of view of the epistemology of induction: "if you've looked at a hundred thousand acres or so of trees — you know, a tree is a tree, how many more do you need to look at?"

In simple cases of these inferences based on causal structure and kinds, the *F*'s all have some feature in common which has a causal relation to *G* of a kind that makes generalization possible. What the investigation is really aiming to do is assess some sort of dependence relation, which might be expressed by a conditional, linking *F* and *G*. This need not describe a causal relation *from F* (or an underlying feature of *F* things) *to G*; it might run from *G* to *F* (all redwoods have such-and-such in their DNA) or from a common cause to each of them. Sometimes the causal basis for the dependence might be known, while in others there may just be reason to think there is some basis of the right kind. Either way, if such a dependence can be established, it does not matter whether the class of *F*'s can or cannot be sampled, is small or large, and so on. Seeing a number of cases of *F* can be helpful, because it may shed light on how *F* and *G* are related, and on how the *F-G* connection is affected by variation in circumstances. But you may be able to get this knowledge just as well or better by looking at other things, which are not *F*, and mere repetition of *F*'s which are *G* – mere weight of numbers with respect to the *F-G* association – does no good at all.

One need not be looking for complete uniformity to engage in this second approach to extrapolation. An unpacking of a few representative emeralds, or melanomas, might show that there will be a particular kind of diversity with respect to the properties you are interested in. If you find that emeralds or melanomas are diverse in nomologically comprehensible ways, it will often make sense to undertake separate investigations of each sub-type. But it might be that the investigation stops with the claim that given the make-up of *F*'s, some range of alternative features *G*, *H*,... (etc.) may each be found with particular probabilities.

At this point a terminological issue arises. Is this second kind of inference also "induction"? I said in the introduction that I was reserving the category "induction" for cases in which the number of *F*'s seen is supposed to be epistemically relevant. I said just above that in the second kind of investigation, having large numbers of *F*'s can be helpful. There is certainly a broad sense in which this helpfulness is "epistemic," but I distinguish it from the stronger sense seen in inference from random samples. In inference from samples, support for the conclusion goes *via* weight of numbers; there is no way that a sample of 10 could in principle do what a sample of 100 can do, if only you

could interact with those 10 objects in a less noisy way, and learn more about the relationships between their various features. In the second kind of inference, that *is* the case, and it is also true that many ways of seeing new instances are no help at all. So I regard the role of numbers in the second category of inference as *practical* rather than epistemic.

Some people might want to use the term "induction" differently, including both categories. The label itself does not matter much, so let us set aside the label and focus on the picture. That picture is one in which we can recognize two kinds of inference. The first is generalization from random samples. This form of inference has the following features: sample size matters, randomness matters, and "law-likeness" or "naturalness" does not matter. The second kind of inference is generalization based on causal structure and kinds. In these cases sample size *per se* does not matter, randomness does not matter, but the status of the kinds matters enormously. These two strategies of inference involve distinct "bridges" between observed and unobserved cases: one goes via the power of random sampling, the other via reliable operation of causes and mechanisms. Then we see that the philosopher's concept of induction, especially since Goodman, has often been a *hybrid* of these. Philosophers have supposed that the crucial category of inference is one in which (i) sample size matters, (ii) randomness is not an issue, and (iii) naturalness of kinds does matter, but *weakly*. By "weakly" I mean that naturalness is used only in the exclusion of bad kinds and predicates, clearing the way for the weight of numbers to do its work. This combines elements of the two, but does so in a way that includes *no* bridge from observed to unobserved. The link that exploits sampling is not available, and there is not the right kind of role for kinds and causal mechanisms either. In good inferences based on causes and kinds, scrutiny of the F and G in question is not aimed at merely excluding bogus collections and pseudo-properties. After all, most combinations of a natural F and a natural G do not show any sort of stable association in which a few cases can be used to draw inferences about many. In good inferences based on causes and kinds, there is reason to think that some particular connection between F and G makes projection possible.

So far I have been treating the mistake that has been made about induction as a philosophers' one. But perhaps the error has deeper roots. It may be that humans have

inductive *habits* which correspond more-or-less to the philosophers' picture I am criticizing. That is, we may have habits in which weight of numbers is taken to have a general importance independent of randomness, and a "can't-be-just-anything" principle also modulates how the experience of numbers affects us. Anthropologists have recently become interested in what might be pan-cultural habits of induction, especially as applied to living things (Medin and Atran 1999). It may turn out that these habits have shaped quite specific features of philosophical treatments of induction. Perhaps, but if so, these are just elements of our psychology, and of our "folk epistemology." These habits are likely to have been reliable enough in the practical domain in which they were developed – or rather, they are likely to have achieved a reasonably good balance with respect to reliability of various kinds, cost, and speed (Gigerenzer and Todd 1999). To show that is to show a kind of in-principle reliability, but a very local kind. This would not, for example, show that these inductive habits are reliable within science and other contemporary epistemic endeavours.

I will finish this section with a historical note. Given the arguments I have made, it is interesting to look back at an exchange between Hans Reichenbach and John Dewey in Dewey's "Schilpp Volume" (1939). Reichenbach modeled all nondeductive inference on a kind of statistical estimation (not the same kind as that discussed here), and hence saw the number of observed cases as crucial. Dewey spurned traditional concepts of induction, especially with respect to the role of weight of numbers. He thought that in actual generalization in science, everything hangs on the scientist's ability to find individuals which are representative of their kind. If we can do this, then one individual is often enough. The hard work goes into saying why a particular case should be representative. Reichenbach argued that Dewey did not appreciate the role of probability, and the significance of the possibility of convergence on a limiting value through repeated observations. For Reichenbach the key to projection lies there. I say that both were were on the right track with respect to understanding some inferences in science. But both were too inclined (ironically) to project, treating one kind of case as the key to all.

4. A Nominalist Challenge

This section discusses an objection to part of the argument above. The objection was raised (in discussion) by both Laura Schroeter and Ira Schall. Like Jackson, I claimed there is an asymmetry between green and grue inductions that has to do with the fact that observation "affects" the objects sampled with respect to *G*-ness in the grue case, not in the green case. But is the asymmetry real? Goodman noted, after all, that "blue" and "green" can be defined in terms of "grue" and "bleen," as well as vice versa. He also argued that judgments of similarity are language dependent. Counterfactual claims, on many views, are dependent on similarity judgments. Putting these points together, it might seem that if we imagine someone starting out with a language that takes "grue" and "bleen" as basic, the asymmetry for them would go away or reverse. Such a person would apparently be entitled to insist that if our particular observed emeralds had not been observed they would still have been grue, so they would also have had to have been blue. Then, as Goodman held, for linguistically different agents different inductions will be acceptable.

To assess this argument, assume we have two agents who each have a home language, normal English or grue-English, but also speak the other language. Their aim is to learn about emerald color by inspecting a random sample. The grue-English speaker picks up the first emerald, and notes that it is both green and grue. Then he asks: "If this thing had not been in the sample, would it still have been grue?" He is a native grue-speaker, but he also knows our words. So he can note to himself that if this thing had not been in the sample, then in order for it to have been grue, it would have had to have been blue. So he is wondering whether *this individual thing in front of him* would have been blue if it had not been one of the emeralds that happened to make its way into the sample – if we had picked up another instead.

Suppose he says that it *would* still have been grue if unobserved, and hence blue. We ask him what the basis for this claim is. He might answer that two grue things are more similar, *ceteris paribus*, than two green things. He goes on to say that if we ask what things would have been like in a situation in which this emerald had never been observed, that situation would have been one in which this emerald was as similar to its actual state as the assumed difference from actuality permits. So it would have been grue,

and hence blue.

If someone says this, we do reach a kind of standoff. But it is different from the standoff that is usually seen as taking hold. The stand-off we reach is one that concerns each member of a collection of individual observed things, and how the agreed-on properties of each thing are causally and counterfactually related. We can pick up each one, and discuss its chemistry, the processes that formed it, and its history of interaction with us. For the grue-English speaker to say that each of the sampled emeralds would still have been grue if unobserved, he has to not just use a different *categorization* of things from us, but *also* has to have a large collection of very strange beliefs about chemistry, history, and light.

The situation is again like the teen smoking case. Suppose you believe that when teenagers are asked about their smoking, many of them immediately take it up. This is a strange belief about a causal dependence. It is not just a strange way of categorizing things. When a model of sampling is applied in an actual-world case, its application will depend on factual beliefs about other matters. Agents with different beliefs on factual matters will put the same model to different use, and may make divergent projections. *Linguistic* difference is not enough to do this, and neither is a different sense of similarity. Disagreements about causal and counterfactual dependence will do it, but that is not surprising. The analogy between the grue emeralds and the fickle teen non-smokers helps us to see this.

5. A Discussion of Cases

So far I have discussed two ways in which observed cases can reliably be used as a guide to unobserved, and contrasted both with how philosophers usually think of "induction." In this section I look at how the strategies discussed above relate to some actual cases of generalization.

In actual epistemic practice, especially in science, we see a mix of the two approaches, plus much more. We see a "mix" in two senses. First, there are fairly clear examples of work done according to each strategy. Second, the answering of a single question may draw, explicitly or implicitly, on both.

I listed some examples at the start of the paper. Let us start with teen smoking. Here, random sampling is used. Collecting samples is feasible, and its limitations don't matter much; we don't expect a generalization to cover future or very distant cases. The class of teenagers is so locally diverse and causally complicated that it would be fruitless to mount a nomological or mechanistic assault on the problem. In Massachusetts in 2007, about 18% of high school students smoked.[8] This figure was based on a random selection of about 3000 students.

Close to the other end of the scale, we have the charge on the electron. This was found to be roughly - 1.6 x 10^{-19} coulombs by Robert Millikan in the oil-drop experiment (Millikan 1911, 1917; Franklin 1997). At the time of Millikan's work it was not agreed that there was a fundamental unit of charge for an electron. Some researchers, such as Felix Ehrenhaft, held or suspected that the charge may vary continuously. So Millikan was not just setting a parameter which everyone agreed must hold universally. Millikan suspended tiny individual oil drops by balancing them between forces due to gravity and an electric field. He then turned the field on and off to see how fast the drops responded. He used that measurement to calculate the total charge on each drop, and found it always to be a multiple of a particular number. This, he argued, was the charge on the electron.

There was no question of his sample being a random one. It was not even a random sample of electrons in Illinois. And Millikan was not looking to support his estimate by sheer weight of numbers. He published results from 58 drops (excluding at least 49 others, for reasons that have been queried). His aim was to get a few well-behaved cases that would show the phenomenon clearly and permit a measurement. In the context of background knowledge, the consistent finding of multiples of one number was taken to establish uniformity – not just for Illinois, but for the universe.

This case is very far from the teen smoking one, but it is not the simplest possible case of a "seen one, seen them all" inference. Millikan was not able to get at individual electrons, and he needed to collect a number of cases to make the argument that there was a natural unit being detected. A more overtly mixed case is that of "Chargaff's rules" regarding the composition of DNA. The most important of these states that the amount of

---

[8] More exactly, about 18% had smoked in the previous 30 days. My source is: *Health and Risk Behaviors of Massachusetts Youth, 2007: The Report.*
<http://www.doe.mass.edu/cnp/hprograms/yrbs/2007YRBS.pdf>

C (cytosine) is the same as the amount of G (guanine), and the amount of T (thymine) is the same as the amount of A (adenine), in all DNA. Chargaff did this work before the structure of DNA had been discovered by Watson and Crick, and there was no obvious reason why the relation should hold. Chargaff began publishing versions of this finding in the late 1940s, and followed up in the early 1950s. In his early papers he gave results for 9 kinds of organisms, scattered through most major groups, adding a few more later. In comparison to Millikan, Chargaff was looking for some coverage of the total field of living organisms – he was looking outside the analogue of Illinois. His list included bacteria, yeast, wheat, sea urchin, and human. He took it that equal proportions in one organism, even in *all* his sample, could be a special case or an accident. In his articles at the time he was quite cautious: "It is noteworthy – whether this is more than accidental, cannot yet be said – that in all desoxypentose nucleic acids examined thus far the molar ratios of total purines and total pyrimidines, and also of adenine to thymine and of guanine to cytosine, were not far from 1" (1951, p. 206).[9] Only with Watson and Crick's work, a few years later, did it become clear that the rule *had* to hold universally, because of the structure of the DNA molecule itself.[10]

How reasonable would it have been to extrapolate on the basis of Chargaff's data alone? To do this, it seems that one would need to argue that *if* there was variation in DNA composition, Chargaff's sample ought to have revealed it. There seem to be two ways in which this might be done. One is to mount an argument from features of the biochemistry of life. No argument of this kind was available in Chargaff's time. The other might be to argue that though the sample was not taken in a truly random way, it might reasonably be taken to have some of the features of a random sample. It is as if someone chose teenagers for a smoking survey who have exactly 5 letters in their first name. This is not a random sample of American teenagers, but it seems likely to be similar to a random sample in some ways, and not biased with respect to the smoking question. Claims of this kind often seem very plausible, but they have a lot of problems around them. What we would need to argue is that the relationship between Chargaff's sample and the total set of organisms on earth is one that has *some useable similarity* to the

<hr>

[9] See also Chargaff et al. (1950) p. 757. Chargaff's handling of the finding is reviewed in Manchester (2008).
[10] Not quite universally: ssDNA viruses, which have a single-stranded form of DNA, are exceptions.

configuration assumed in a model of random sampling. I think that Chargaff was right to be cautious, even though he probably came to regret it. He was not given a share in the Nobel Prize with Watson and Crick, to his great frustration, even though Maurice Wilkins, whose role was fairly minor, was.[11] People in later years have wondered why Chargaff did not make more of the regularity he had found, and push its significance harder (Manchester 2008). If he had done so, urging that he had uncovered a general fact that was surely a clue to the structure of the DNA molecule, he might well have won the Nobel.

I have been discussing success stories in this section. We should look at some bad inductions as well as good ones. Here the best-known example is a suitable one: the swans. It is no surprise that the white-swan generalization (if people actually made it) turned out to be false. The sample seen by Europeans was not random (as swans in distant parts of the world were very unlikely to be observed) and there was no support to such a uniformity given by known biological mechanisms (or unknown ones, for that matter). One can be unlucky with a good method, of course, but the swans case was not like that. There was no reason to think the generalization was true. No good bridge from observed to unobserved was exploited or even available.

6. Conclusion

The approach to the problem of induction taken in this paper has been to ask: how can inferences from observed cases to generalizations be reliable? "Reliability" can be understood in many ways (Goldman 1986). Rather than going into the details of this concept, I have asked very coarse-grained questions. What *sort of relation* might be available for an agent to exploit? What bridges between observed and unobserved can make reliable inference possible? In the case of a familiar philosophical sense of induction, I argue that there is no bridge. Induction in that sense, again, is supposed to be a kind of inference in which weight of numbers matters even without randomness, and the predicates involved are constrained but only to rule out pseudo-patterns in which the weight of numbers does not have its usual role. I agree that this kind of inference might be an intuitively attractive one, but there is no reason why it should tend to work, at least

---

[11] Rosalind Franklin, who many think deserved a share in the prize, had died.

in worlds like ours. Two relatives of this form of inference can be reliable in principle, however. These are inferences about populations from random samples, and inferences about unobserved cases based on mechanisms, laws, and etiologies common to a natural kind.

Science and other parts of epistemic practice often combine or mix these two methods. Sometimes one or the other is used overtly, but sometimes they are used more implicitly, even inadvertently. A researcher might have a rather incoherent epistemic ideology, while it is possible for an observer of their work to say: here the method used in fact approximated an X-based one... here it approximated a Y-based one.

The role of approximation seems particularly important here, and poorly understood from a philosophical point of view. In the case of inference from samples, the situation seems like this. If a physical set-up really does meet the requirements for random sampling, then reliable inferences can be drawn. But the requirement that every member of the population has the same chance of making its way into the sample is very strong. Or rather, it is strong if it makes sense at all, and it is not even clear that this application of a physical concept of probability makes sense. So philosophers often try to do without randomness altogether. (An example is seen in the Williams-Stove justification of induction, especially as defended in Campbell and Franklin, 2004).[12] That, I argue, will not work. It is better to try to make sense of the idea that many actual-world set-ups involve a useable approximation to random sampling. That is surely what people designing smoking surveys rely on, and the same can also be said in some cases where an investigator does not realize why what he is doing is in fact likely to work. The task we then have is to philosophically understand these partial matches between abstract probabilistic models and physical configurations.

In cases where a real relation between observed and unobserved is being exploited without the agent realizing this, the beliefs formed have a complicated relation to ordinary patterns of epistemic assessment. In one sense the agent is "justified" in generalizing, and in another sense he is not. Our habits of epistemic assessment form a complicated soup. The kind of justification discussed in this paper is not the only one that can be recognized in everyday talk, and not the only one relevant to epistemology.

---

[12] See also Williams (1947) and Stove (1986).

For some philosophers, this paper will also not have made much contact with what they see as the fundamental problem: inductive skepticism. Why was Millikan, Chargaff, or any Massachusetts public health official entitled to assume that the universe will not change fundamentally from one moment to the next? Why were they entitled to think their offices would still be there the next morning, that sea urchins would not explode, and that teenagers would continue to have lungs to smoke with? Nothing has been said here about questions of that kind, and for some philosophers, *these* are the real problems about induction, not questions, like the ones in this paper, that we ask against a background in which many familiar things are assumed to behave normally. In reply, I accept that more dramatically skeptical questions about induction are worth asking. The skeptical challenge is often posed in a way that relies on a particular model of the universe, however – the Humean model of a "loose and separate" world, or a mosaic of "distinct existences" (Hume 1748, Lewis 1986). Within that sort of model, inductive skepticism is an acute problem, but this is just one toy model of the universe, useful for posing certain questions, not a representation of how we must take things to be, or even a default assumption that must be shown to be wrong before we can think differently. And taking the skeptical challenges to be worthy of discussion does not prevent us from asking other questions about how observed cases can point beyond themselves to unobserved ones, questions I have tried to make progress on here.

<center>*   *   *</center>

References

Boyd, R. (1999). "Homeostasis, Species, and Higher Taxa." In *Species, New interdisciplinary essays*, (R. Wilson, ed.) Cambridge, MA: Bradford/MIT Press, pp. 141-185

Campbell, J. and J. Franklin (2004). "Randomness and Induction," *Synthese* 138: 79-99.

Carnap, R. (1952). *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Dewey, J. (1939) (1939), "Experience, Knowledge and Value: A Rejoinder", in P. A. Schilpp and L. E. Hahn (eds.), *The Philosophy of John Dewey. (Library of Living Philosophers)*. La Salle: Open Court, 517–608.

Franklin, A. (1997). "Millikan's Oil-Drop Experiments," *The Chemical Educator* 2: 1-14

Chargaff, E., S. Zamenhof, and C. Greene (1950). "Composition of Human Desoxypentose Nucleic Acid." *Nature* 165: 756–757

Chargaff, E. (1951). "Chemical Specificity of Nucleic Acids and Mechanism of their Enzymatic Degradation." *Experientia* 6: 201-240.

Geach, P. (1972). *Logic Matters.* Berkeley: University of California Press.

Gigerenzer, G. Todd, P. and ABC Research Group (1999). *Simple Heuristics that Make Us Smart.* Oxford: Oxford University Press.

Godfrey-Smith, P. (2004). "Goodman's Problem and Scientific Methodology." *Journal of Philosophy* 100 (2003): 573-590

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge MA: Harvard University Press.

Goodman, N. (1955). *Fact, Fiction and Forecast.* Cambridge, MA: Harvard University Press.

Harman, G. (1965). "The Inference to the Best Explanation." *Philosophical Review* 74:88-95.

Hume, D. (1748/1993) *An Enquiry Concerning Human Understanding.* Indianapolis, IN: Hackett.

Jackson, F. (1975). "Grue." *Journal of Philosophy* 72: 113-131

Jackson, F. and R. Pargetter (1980). "Confirmation and the Nomological." *Canadian Journal of Philosophy* 10: 415-428.

Kelly, K. (2004). "Why Probability Does Not Capture the Logic of Scientific Justification," in C. Hitchcock, (ed.), *Contemporary Debates in the Philosophy of Science*, London: Blackwell, 2004.

Lewis, D. (1983). "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61: 343-377.

Lewis, D. (1986). "Preface" in *Philosophical Papers, Volume 2*. Oxford: Oxford University Press, 1986.

Manchester, K. (2008). "Erwin Chargaff and his 'Rules' for the Base Composition of DNA: Why Did he Fail to See the Possibility of Complementarity?" *Trends in Biochemical Sciences* 33: 65-69.

Medin, D. and Atran, S. (eds.) (1999). *Folkbiology*. Cambridge, MA: MIT Press.

Millikan, R. G. (2000). *On Clear and Confused Ideas*. Cambride: Cambridge University Press.

Millikan, R. A. (1911). "The Isolation of an Ion, A Precision Measurement of Its Charge, and the Correction of Stokes's Law" *Physical Review* 32: 349-398.

Millikan, R. A. (1917). *The Electron*; Chicago: University of Chicago Press.

Norton, J. (2003). "A Material Theory of Induction." *Philosophy of Science* 70: 647 – 670

Quine, W. V. (1969). "Natural Kinds," In *Ontological Relativity and Other Essays*. New York: Columbia University Press, pp. 114-138.

Rawls, J. (1971). *A Theory of Justice*. Cambridge MA: Harvard University Press.

Reichenbach, Hans (1939), "Dewey's Theory of Science", in P. A. Schilpp and L. E. Hahn (eds.) *The Philosophy of John Dewey*. *(Library of Living Philosophers)*. La Salle: Open Court, 159–192.

Stalker, D. (ed.) (1994). *Grue: The New Riddle of Induction*. La Salle: Open Court

Stove, D. C. (1986). *The Rationality of Induction*. Oxford: Oxford University Press.

Strevens, M. (2004). "Bayesian Confirmation Theory: Inductive Logic, or Mere Inductive Framework?" *Synthese* 141: 365-379

Williams, D. C. (1947). *The Ground of Induction*. Cambridge: Harvard University Press.