GOODMAN'S PROBLEM AND SCIENTIFIC METHODOLOGY*

Nelson Goodman's "new riddle of induction"[1] shows that there is more than one way to project from data, in a way consistent with traditional "formal" constraints on such inductions and projections. From a data set of emeralds in which all are found to be green, we can infer (fallibly) that all emeralds are green, but we can also apparently infer that all emeralds are *grue*, where an object is grue if and only if it has been observed before now and is green, or has never previously been observed and is blue. The riddle asks: What makes the "green" induction better than the "grue" one?

I. INTRODUCTION

Half a century after Goodman, a huge variety of solutions has been proposed. There is no consensus on which solution is best, but many of the solutions have a feature in common. They assume that the two inductions share all the features that might be addressed by a general discussion of good methods of projection *within* science. Here I have in mind the kind of discussion that might be found in a statistics textbook, or a book describing the methods of data analysis used in some particular scientific field.

The two rival inductions are usually presented very schematically, of course: all observed *Fs* are *G*, so probably all *F*s are *G*. But it is often thought that no real difference is made if we fill in more detail, and make the cases more scientifically life-like. We might dress up the two inductions in the garb of estimations from samples using statistical techniques, but that will not affect the basic problem. We will still have a set of emerald observations that is equally compatible with a normal-looking "green" projection and with an abnormal-looking "grue" projection.

So in a sense, the philosophical literature has often assumed that the statistics and data analysis textbooks used in science really need a kind of "chapter zero." This chapter zero would lay down constraints

* Earlier versions of this material were presented to colloquia and conferences at Stanford University, Australian National University, University of California/Irvine, and University of California/Berkeley. I am especially grateful to Aldo Antonelli, Branden Fitelson, Brian Ellis, Ben Escoto, Alan Hayek, John MacFarlane, Simon May, Rajani Rajan, and Brian Skyrms for comments and correspondence.

[1] Goodman's most famous discussion of his problem is in his *Fact, Fiction, and Forecast* (Cambridge: Harvard, 1955). For a collection surveying solutions proposed in the last fifty years, see Douglas Stalker, ed., *Grue!* (Chicago: Open Court, 1994).

0022-362X/03/0011/573–90

on inference that are in a sense prior to the constraints discussed in the book itself. These constraints would impose restrictions on which predicates belong in data analysis at all. Some predicates, like "grue," are just *nonprojectible*, and should be ruled out or at least downgraded ahead of time, as the basis for inferences from data. Suggested bases for this restriction vary widely (from the metaphysical to the conventional), but the basic idea is very common. Of course, philosophers assume that in practice the enculturation of a young scientist will informally impart most of the content of the absent chapter zero. There is usually no need to make chapter zero explicit, except when doing philosophy.[2]

I will argue against this approach to Goodman's problem. I will also, roughly speaking, propose a new solution, but this solution draws very heavily on the work of others. The solution is also old in another sense; it rests on an appeal to some standard ideas in scientific data analysis. Further, part of my "solution" consists in making a distinction between different aspects of the problem, which require different kinds of treatment. What we have come to call "Goodman's problem" is something of a mixture. Philosophical discussions of confirmation have assimilated issues and problems that have very different properties. This finding has general consequences for the philosophical project of giving a "theory of induction." These claims about the disunity of Goodman's problem will be introduced at the end of the article, however.

As a preliminary, I should also note that some Bayesian analyses of Goodman's problem are important exceptions to my generalization about how the literature has handled Goodman's problem. Here I have in mind some versions of the idea that the solution involves a difference between the *prior probabilities* of the two hypotheses about emerald color.[3] I will not argue directly against this approach, and will say little about Bayesianism in this article. However, the core of my proposals could be embedded within a Bayesian framework.

II. THE JPB PROPOSAL

This section and the next will develop a proposed solution to Goodman's problem. Although this solution is based on concepts used

[2] John Pollock has expressed the alleged need for this supplement to standard statistical theory explicitly: "It seems likely that…[a] strong projectibility constraint should be imposed upon all familiar patterns of statistical inference. The need for such a constraint seems to have been ignored in statistics"—"The Projectibility Constraint," in Stalker, ed., p. 141.

[3] See, for example, Colin Howson and Peter Urbach, *Scientific Reasoning: The Bayesian Approach* (Chicago: Open Court, 1989), chapter 4; and Elliott Sober, "No Model, No Inference: A Bayesian Primer on the Grue Problem," in Stalker, ed., pp. 225–40.

within scientific data analysis, my discussion will be very informal. In fact, my point of departure is a proposed solution, nearly thirty years old, that is a paradigm of an "analytic philosophy" approach to the problem, one that uses no tools from science or technical philosophy of science. This is Frank Jackson's proposal, introduced in 1975 in this JOURNAL and revised in 1980 with Robert Pargetter.[4]

Jackson argues that to deal with Goodman's problem we must explicitly attend to the fact that certain objects, the ones we are using as the basis for our induction, have the property of *having been observed* (or *examined*, or *being within our sample*). I will use the symbol "O" for this property. Different treatments of induction, and different inductive arguments, will involve slightly different O properties. In this article, I will abstract away from those differences as much as I can (though see Jackson, *op. cit.*, for more detail). So we start with a representation of inductive arguments that is something like this:

(1) All Fs that are O are G.

(2) All Fs (whether or not they are O) are G.

Goodman's problem shows that not all inductions that have the above form are good ones. Here we use the following definition of "grue."

Grue $=_{\text{df}}$ (green & O) or (blue & not O)[5]

Substituting "emerald" for F and "grue" for G, we create a bad argument. But Jackson in 1975 suggested that if we add one more premise (which does not involve a projectibility constraint on predicates themselves), the resulting form will always make for a good inductive argument. We need to add the "counterfactual condition." Rephrasing his proposal slightly:

(1) All Fs that are O are G.
(2) If those Fs were not O, they would still have been G.

(3) All Fs are G.

This condition is met for G = green and not for G = grue.

[4] See Jackson, "Grue," this JOURNAL, LXXII, 5 (March 13, 1975): 113–31 (reprinted in Stalker, ed.); and Jackson and Pargetter, "Confirmation and the Nomological," *Canadian Journal of Philosophy*, X (1980): 415–28.

[5] This definition of "grue" is based on Jackson's 1975 discussion. Jackson also discusses various other definitions of "grue" in the literature, and shows why several of them (especially those that require particular grue emeralds to change from green to blue) fail to generate real philosophical problems about induction. Jackson's most precise definition of "grue" makes allowance for the fact that particular objects can change color. Below, T denotes some particular date, like January 1, 2020, while t is a variable. Grue at $t =_{\text{df}}$ (green and t & observed by T) or (blue at t & not observed by T).

Jackson's introduction of counterfactuals would be rejected as a nonsolution by Goodman himself. For Goodman, the concepts of law, disposition, natural kind, and counterfactual dependence form an interdefined network and are all dubious when interpreted in a strongly realist way. In response to worries of this kind, Jackson emphasizes that what we need here is knowledge of counterfactuals about the particular objects in the sample, not about emeralds in general.[6] I urge the reader to accept the use of these counterfactuals for a moment, and let the rest of the picture unfold. We return to the issue later.

Jackson's 1975 proposal works well for the case of "grue." But consider a different class of problem cases, "emerose" cases.[7]

Emerose$_1$ $=_{\text{df}}$ (emerald & O) or (rose & not O)

Now consider this argument:

(1) All emerose$_1$s that are O are green.
(2) If those emerose$_1$s were not O, they would still have been green.

(3) All emerose$_1$s are green.

Again we seem to have true premises and a bad induction. For a family of reasons, including reasons of this general kind, Jackson refined his proposal in his 1980 article with Pargetter. Their modification was to treat the F term and the G term similarly. I again rephrase their proposal.

(1) All Fs that are O are G.
(2) If those Fs were not O, they would still have been F and G.

(3) All Fs are G.

This will block the emerose$_1$ case. But consider a new "emerose" predicate.

Emerose$_2$ $=_{\text{df}}$ (emerald) or (rose & not O)

Now we do not take *any* emeralds out of the class of emeroses. This restores the problem for Jackson and Pargetter. If we substitute "em-

[6] Laura Schroeter (in discussion) suggested that Jackson's specific counterfactual claims might be resisted. Suppose we accept Goodman's arguments that similarity judgments (like claims about patterns and regularities) are language-dependent. But counterfactual claims, on many views, are dependent on similarity judgments. Then a speaker of a language in which "grue" is linguistically normal might 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. This objection raises interesting issues but I will not pursue it further here.

[7] See Goodman, p. 74, footnote 10.

erose$_2$” for *F*, then the second premise in the induction is true. But the induction is bad.[8]

A reply to my emerose$_2$ argument was proposed by Alexis Burgess (while an undergraduate, in a midterm).[9] Here is a modified version of his suggestion for the proper form of an inductive argument:

(JPB) (Jackson-Pargetter-Burgess)

(1) All *F*s that are *O* are *G*.
(2) There exists no property *C* (other than *G*, and weakenings and strengthenings of *G*) such that:
  (2.1) All *F*s that are *O* are *C*,
  (2.2) Some *F*s are not *C*, and
  (2.3) The *F*s that are *O* are *G* because they are *C*.

---

(3) All *F*s are *G*.

This deals with the emerose$_2$ case, and with the other cases as well. In the case of emerose$_2$, the relevant *C* is being an *emerald*. In our sample of emerose$_2$s there are only emeralds, and we have reason to believe that the greenness of these objects was due to their being emeralds. The emeraldness of the emeroses in our sample is interfering with the induction. The same applies in the case of emerose$_1$. In the case of the original grue problem, *C* is being *observed*. Here Burgess’s proposal reduces to Jackson’s original one.[10]

There are problems of detail with this proposal; it is not clear how best to formulate the second premise. In my formulation I used a “because” in (2.3) rather than an explicit counterfactual. This “because” is meant to include a variety of dependence relations, including both causal and logical ones. I emphasize, following Jackson, that this “because” claim is about the specific objects in our sample, and it should not be interpreted as involving too strong a brand of responsibility. The bracketed restriction on values of *C* can be ignored for now. I use the term *JPB proposal* (Jackson-Pargetter-Burgess proposal)

[8] This is not a case where one sample generates two conflicting inductions, and Jackson did require this in his 1975 article for there to be a philosophical problem, as opposed to a case of the ordinary fallibility of induction. But this is a case where green emeralds can be used to project greenness in all unobserved things, and similarly irrelevant observations can generate conflicting color projections about the same unobserved objects. So induction has collapsed as a way of discriminating good from bad predictions.

[9] Burgess has not published a discussion of this topic of his own (see his manuscript, “It’s Not Easy Being Grue”). His view of the status of the proposal, and its best formulation, differ from the view defended here.

[10] Emerose$_1$ could also be dealt with by requiring via a different version of premise (2.3) that the sampled objects not be *F* (as well as not be *G*) because they were *C*. See the discussion of emerose$_1$ in the next section.

for a family of slightly different formulations that have the same basic structure. Different formulations and technical complications are discussed below in the appendix.

The most important point here is not the details of how to formulate the JPB proposal. What is important is that with Burgess's refinement of Jackson and Pargetter in place, we are in a position to see what is really going on in the grue/emerose family of problems. What Burgess has done is (inadvertently) assimilate the grue problem to a familiar set of issues in scientific methodology.

III. CONFOUNDING VARIABLES, HAWTHORNE EFFECTS, AND BIASED SAMPLES

In statistics and data analysis, one thing that people look for is associations between properties. But not all associations are regarded as genuine, in a sense that allows extrapolation, generalization, or causal interpretation. These associations are sometimes called *spurious*. How can one distinguish spurious associations from genuine ones? Sometimes one cannot, or not until things go wrong. But there are some well-known sources of error that can be guarded against. These problems are treated as defeaters of inferences that are ordinarily sanctioned.

The philosophical problems we have discussed here correspond to known sources of trouble in data analysis; they correspond to two sources of spurious associations. For a moment, it will be helpful to shelve "emerose$_1$" and focus on "grue" and "emerose$_2$."

The problem raised by "grue" is an unusual version of, or a close relative of, the problem of *confounding*. Here I have to stretch existing scientific terminology somewhat. The term "confounding variable" is usually only employed in discussions of causal inference. Judea Pearl formulates the most basic idea here as follows: if we are trying to work out whether there is a causal relation between *X* and *Y*, a confounding variable is another variable *Z* that affects both *X* and *Y*.[11] I prefer the following formulation: if we are trying to work out whether a variable *X* causally affects another variable, *Y*, a confounding variable is another variable *Z* that (i) affects *Y*, (ii) is nonrandomly associated with *X*, but (iii) is not itself causally dependent on *X*.

Suppose we want to know whether smoking causes heart disease. But when we do a survey of the population we find that the overall rate of heart disease is *lower* in the smokers. What does that show? We do not know; it might be due to a confounding variable. The

[11] See Pearl, *Causality: Models, Reasoning, and Inference* (New York: Cambridge, 2000), chapter 6.

smokers might tend to be athletic people who exercise a lot. Then we might find that the overall rate of heart disease is lower in smokers than in nonsmokers, but as a consequence of a confounding variable, exercise. We can correct for the problem by dividing the population into exercisers and nonexercisers, and assessing the association between smoking and heart disease within each. It may turn out that within each group, smoking and heart disease are positively correlated, even though in the whole population the association disappears. (This is a case of "Simpson's paradox.")

So when doing causal inference we must guard against confounding variables. Confounding variables give rise to spurious associations (or spurious absences of association).[12]

The grue problem appears at first to be a very different kind of problem from the smoking/exercise problem. Here we are not doing causal inference, but pure projection. We are not trying to find out whether emeraldness causes greenness or vice versa, but only trying to find out whether all emeralds are green. We would not mind if there was a common cause of emeraldness and greenness, or if one caused the other, so long as all emeralds are green. But if we look at these "pure" projections closely, we find some kinship with the issues arising in causal inference. We should not make a projection from a sample if there seem to be the *wrong kind of dependence relations* between properties of the sampled objects.[13]

In the grue case, we find an association between emeraldness and grueness in our sample. But it is clear that this association has arisen for the wrong reasons. All the objects in the sample are observed things, as well as grue emeralds, and we know that observation *affects the application* of the predicate "grue." It is not that observation physically affects the objects in the sample, but it affects the ways that objects fall into classes expressed by predicates. This is a peculiar kind of dependence relation between features of the sample, which interferes with the projection (a kind of "Cambridge dependence"). It is epistemologically similar to the situation we would have if we had reason

[12] Pearl (*op. cit.*) argues that confounding is an irreducibly causal concept that cannot be expressed in statistical terms, but is also fundamental to data analysis in many sciences. The question of whether or not confounding can be defined in purely statistical terms is orthogonal to my concerns in this article.

[13] In traditional philosophical discussion, a distinction is drawn between the "observed" and the "unobserved" emeralds. In discussing the grue problem using the tools of scientific data analysis, the distinction that is more significant is that between the emeralds in a sample and the emeralds in the rest of the emerald population—the "sampled" and "unsampled" emeralds. To keep this article readable, I have moved between these distinctions without giving them much explicit attention.

to believe that the collection of the emeralds in our sample had included a process that painted or dyed those emeralds.

The problem with "grue" is similar to a confounding variable problem, as it involves the interfering role of an extra dependence relation that creates a spurious association. I do not claim that this is exactly the same as a classical problem of confounding, but the kinship is quite close. In causal inference, confounding involves a dependence relation that may create a spurious association in an entire population, misleading us about causal structure. In the case of projection (or estimation), the dependence between grueness and observation creates a spurious association in our sample, misleading us about the larger population of emeralds. Causal inference must guard against the possibility that it is not just the smoking that is affecting the health of smokers in our population, but unusual rates of exercise as well. Estimation and projection must guard against cases in which it is not just the emeralds' physical nature that is affecting how color predicates apply to them, but the process of observation as well.

A more specific comparison can also be made between the grue problem and what is known in the social sciences as the "Hawthorne effect." In a case of the Hawthorne effect, the subjects of a social scientific study behave differently because of their interaction with the observer studying them. They may work more efficiently, for example, as was conjectured in the case that generated the concept, a 1920s study of a factory in Hawthorne, Illinois.[14] In a Hawthorne effect case, *O* has a causal impact on other properties of interest in the study; in the grue case, *O* causes trouble via a noncausal dependence.

So my argument is that standard methodological principles in science tell us not to extrapolate grueness from a sample of grue emeralds, because we have reason to believe that the association in the sample is the product of a "bad" dependence relation that behaves similarly to a case of confounding, or a Hawthorne effect. Roughly, the semantics of "grue" turn observation itself into a confounding variable.

The emerose$_2$ problem is different. Here we do not find a bad *dependence* relation between properties of objects in the sample.[15] The emerose$_2$ predicate introduces a *selection bias* in our sample.

If we are to make a good inference using the basic techniques of

[14] The original study (which is now quite controversial) is discussed in E. G. Mayo, *The Human Problems of an Industrial Civilization* (New York: Macmillan, 1933). Mayo's discussion tends not to distinguish between the effects of special kinds of work supervision and the effects of being studied *per se.* I am grateful to Art Owen for bringing the Hawthorne effect to my attention.

[15] That is why the emerose$_2$ problem was not handled by Jackson and Pargetter's 1980 proposal.

statistical analysis of samples, every member of the global population we are interested in should have the same chance of being found in the sample. Often this is not possible in practice, but then we at least should try to ensure that the properties that made some objects likely to end up in the sample are not associated with the property we are making a projection about (property G). But the only objects that can be found in a sample of emerose$_2$s are emeralds. Roses cannot be found within the sample and still be emerose$_2$s, because of the peculiar semantics of the predicate "emerose$_2$."[16] So it is impossible to collect an unbiased sample of emerose$_2$s. And we have good reason to believe that the emerald/rose difference is relevant to inferences about color.

So standard methodological principles tell us not to extrapolate greenness from a sample of green emerose$_2$s to the total population of emerose$_2$s, because we have reason to believe that any such sample will be contaminated by selection bias.

In the grue and emerose$_2$ problems we find two different ways of generating a spurious association: selection bias, and something akin to confounding. These problems are, I emphasize, somewhat different. In the case of grue, we may have gathered a perfectly good sample of emeralds, but these objects have been affected, with respect to how crucial predicates apply to them, by their being observed. In the case of emerose$_2$, we just have a bad sample.[17]

The case of emerose$_1$ is slightly different again. We can exclude the emerose$_1$ projection as a case of selection bias, as with emerose$_2$. But in this case we also find an inappropriate dependence between properties of the sampled objects and the fact of being observed, as in the case of grue. If the emerose$_1$s in our sample had not been observed, they would not even have been emerose$_1$s (assuming they would still have been emeralds in that case). These objects would have been outside the population being projected *to*, as well as being outside the sample being projected from, because of the behavior of the term "emerose$_1$." This is an especially unusual case.

Burgess's generalization of Jackson is useful because it captures, in a rough but very informative way, a common feature of the various

[16] Here, and elsewhere, I assume that the sampled emeralds are the same as the observed ones, to avoid having to define different "grue" predicates to use in slightly different cases.

[17] See also J. Moreland's "On Projecting Grue" for an earlier attempt to deal with the grue problem (though not emerose problems) using the concept of randomness, applied within a Carnapian framework—*Philosophy of Science*, XLIII, 3 (1976): 363–77. Note also that while I use the term "bias" in a specific way here, the term can also be used more generally to refer to any (nonchance) source of spurious association.

kinds of problem discussed above. In any induction, we need to try to ensure that there is not some property of the sampled objects that *makes those objects special* with respect to the question being addressed. This is what "*C*" in the JPB formula is aimed at picking out. If there is some such property, we should not think that what we find in the sampled objects can be extrapolated to the unsampled objects.[18] The grue and emerose predicates devised by philosophers do not raise a new kind of problem for nondeductive inference—instead they raise existing kinds of problem in an unusual way. We must guard against bias and confounding when making inferences from samples—that much we already knew. But most problems of these kinds come from empirical features of the situation, from extra causes we have not controlled for, or flaws in our sampling methods; we must make sure that our emeralds did not come to us via a process that physically alters their color. "Grue" and the "emerose" predicates are unusual because they are terms that *create* their own problems with confounding and bias, when used in inductions. That is why grueness in an emerald sample does not have the same significance that greenness has in such a sample. There is *no* way to control for the "effects" (in scare quotes) of observation on the properties of our sample in these cases. A standard green-emerald induction is a risky thing, but at least our model of the inference does not tell us in advance that it is unreliable.

Before concluding this section, I should revisit the issues concerning counterfactuals and "because" statements in the JPB proposal and my discussion of confounding. For followers of Goodman, any use of these tools in a purported solution of the new riddle produces a subtle circularity in the analysis. I will not address the most general issues of that kind here. The view being proposed does require that we accept and make use of the idea of dependence relations among the properties of individual objects. These can be described in a variety of ways, and can be philosophically analyzed in a variety of ways. No particular theory of counterfactuals is assumed here, for example (though see footnote 6 above). The kind of talk about dependence

[18] A referee suggested that once the family of problems have been linked in this way, they might be further analyzed by linking them to the "reference class" problem familiar from earlier discussions of induction. This possibility can be seen most readily in the case of emerose$_2$, but might be extended to the others. For example, *observed emeralds* are a narrower reference class than *emeralds*, and the peculiarities of the "grue" predicate might prevent us from using the frequency of grueness in the narrower class as a guide to the frequency in the broader class. The Hawthorne effect might be treated the same way. This suggestion raises interesting issues, but would involve recasting the discussion in a different framework and I will not follow it up further here.

relations that is needed for the solution to work is, as I hope this section has shown, a fairly down-to-earth kind that is familiar within science. It is a fairly extreme manifestation of Humeanism to deny that *any* dependence concepts of the kind required are legitimate.[19]

My proposed solution to Goodman's problem is not to be identified with the formula labelled "JPB" above. As we will see in the Appendix, it is hard to get a version of the JPB formula to deal with all the cases in the right way. And formulas like JPB, even when they seem to capture the cases correctly, do no more than capture a set of intuitions about justified inference. They do not say why this kind of transition from observed to unobserved is justified or reliable. Indeed, I do not think that JPB, or anything similar to it, represents a rule or argument form that is justified in all cases. What JPB does is abbreviate some key features of a more detailed treatment of reliable inference from samples, a treatment developed in statistics and various parts of science itself. That approach does more than capture intuitions; it uses a model of the sampling and inference process to show that some inference procedures are reliable (if the assumptions in the model hold). I do not deny that there are foundational problems with these aspects of scientific methodology. But I argue that this body of theory and practice does contain tools that resolve the grue and emerose problems, in the versions discussed above. The JPB formula, in turn, abbreviates the features of that theory that give us the crucial epistemic distinction between "grue" and "green" as they function in projections.

IV. THE LIMITS OF THE SOLUTION AND THE DISUNITY OF THE GRUE PROBLEM

I have argued that the restrictions needed to exclude the grue and emerose inductions already exist, implicitly, within the toolkit of scientific data analysis. However, I have developed this argument by forcing the grue induction into the form of one particular kind of inference found in science: inference from a sample to a larger population from which the sample is drawn, using statistical tools. This is, in many ways, the appropriate scientific case to focus on first. But as we will see, we need to think differently about grue when dealing with some other kinds of inference. Goodman's compact original discussions raise problems for more than one kind of inference, and it is a mistake to impose a single solution on what is really a collection of problems.

[19] One theme of Pearl's book is a defense of the scientific respectability of an anti-Humean, causal realist position, which is friendly even to counterfactuals. His treatment of confounding as a causal concept is part of that defense. My discussion here is not committed to the particular version of anti-Humeanism that Pearl defends.

I will begin this part of the discussion by looking at some issues connecting the grue problem with laws and causation. The solution discussed in earlier sections of this article asserts no connection between the confirmation of generalizations and "lawlikeness." If all emeralds are "accidentally" green, this makes no difference to the story told above. It also does not matter whether or not the emeralds form a "natural kind." What matters is that the emeralds form a collection of objects that can be randomly sampled.

It does make a difference if the mechanisms responsible for color in the sampled emeralds are *different* from the mechanisms responsible for color in the unsampled emeralds. If we believe there is diversity in mechanisms that correlates with membership in the sample, we have a problem of the "confounding" kind discussed earlier. Whatever the facts are concerning the link between F and G in the sample, we need to assume that the same factors are operating in the unsampled part of the population, for a reliable inference to be made about the total population using the simple statistical model. What is making our inference procedure work in this case is the power of random sampling, as described by probability theory—the tendency for random samples to resemble the larger populations from which they are drawn. That is what is giving us the ability to move from the observed to the unobserved.

Although the distinction between natural kinds and mere collections, and between lawlike and "accidental" generalizations, does not matter in this context, the power of random sampling to connect the observed to the unobserved does depend on stringent conditions. We do not have to push very hard in order to push up against the limits of the set of inferences that can be treated that way, and hence to the limits of the proposal I made about Goodman's problem in the previous sections.

Think once again about emeroses. In the case of emerose$_2$s, only emeralds can be in our sample and still be emerose$_2$s. So there cannot be a good random sample of emerose$_2$s. But this points us towards a more general problem: *future* individuals cannot ever be in a present-day sample. Emeralds (or smokers, or ravens) that do not yet exist cannot find their way into our present-day samples. This shows us how quickly we reach the limits of the class of inferences that can be modeled and justified using the idea of random sampling alone. We can make an inference from a sample to a larger population without worrying about what sort of class, collection, or kind the larger population constitutes. But when the population from which we draw our sample is really a mere *sub*population of a larger population that we wish to project to, then we need substantive assumptions about similarity between the subpopulation and the larger population. Otherwise, we have no reason to believe our sample is a good basis for projection to the larger population.

The epistemic gap between random sample and total population is fundamentally different from this gap between a subpopulation and a larger population "attached" to it. If we want to make inferences about a larger population that cannot be properly sampled, we must ask: What *kind* of collection is this? Are these objects all the products of a common type of origin? Do they have a common internal structure? What sort of causal or nomic relationship is there likely to be between properties we are projecting from and properties we are projecting to? Here we reach the kinds of issues usually discussed by philosophers in relation to "natural kinds."

So now we must ask a fresh set of questions about grueness. Once again we imagine having before us a pile of emeralds, but now we cannot see this pile of emeralds as drawn from a total population by random sampling. Instead, our pile of emeralds is like a little subpopulation attached to the rest of the emeralds (or a sample of such a subpopulation). Once again, we ask why we should not project grueness to the rest of the emeralds. In this context, Goodman's argument is raising a different set of issues from those discussed earlier. Now it is a very important fact about grue that it is an odd-looking property in relation to our background knowledge for the case at hand.

Here we do run into inconvenient features of the standard example. Minerologists seem to treat the greenness of emeralds as a defining characteristic of the kind: emeralds are beryl crystals made green by chromium impurities. So let us stipulate that the kind being investigated is beryl crystals containing chromium impurities. The focus of our investigation will then be the way a beryl crystal with chromium responds to light. As the grammar of my previous sentence suggests, in this investigation it is only practical considerations that make a sample of a hundred emeralds superior to a sample of one. The mere addition of new emeralds to our data set does not, as it does if we are making an estimate from a random sample, make much difference to the situation.[20]

[20] The debate between John Dewey and Hans Reichenbach in Dewey's "Schilpp volume" includes an interesting exchange on this issue. Reichenbach modeled all nondeductive inference on statistical estimation, and hence saw the role of sample size in generating convergence on a true value as crucial. Dewey had a different model of nondeductive inference, in which all hangs on the ability to find an individual that is representative of its kind; if we can do this, then one individual is enough. I am suggesting that both Dewey and Reichenbach were on the right track with respect to understanding some inferences in science, but both were too inclined to general-

I will not try to give a detailed account of exactly which factors bear on this form of the grue problem. Clearly though, the large literature on natural kinds is relevant here.[21] Some of the other tools employed in more standard discussions of grue—parsimony, prior probability—may also be important. I would favor a multi-factored treatment of this aspect of the problem. The special features of "grue" that Jackson focused on still play a role. When we make inferences about emeralds from those we have seen, it is relevant to ask whether there are any likely effects on our emeralds of the processes by which they made their way into our possession. But a great deal of attention must be paid to something that has no role in the JPB approach—what sort of kind *emerald* is, and the network of properties (especially intrinsic, structural properties) characteristic of this kind.

So the proper treatment of grue in a purely statistical inference is different from its proper treatment in an inquiry into the causal and nomic properties of a putative natural kind. This disunity is to be expected; we have here two rather different kinds of inference found in science. The two kinds of inquiry are often connected, and they may sometimes combine so closely that they are hard to disentangle in a particular case. But they do raise different epistemological issues.

We see an indication of all this in the JPB formula itself. We ask: Is there a particular known factor *C* that explains the *G*-ness of the sampled *F*s? If no, we proceed with the projection, though of course we might be wrong. If the answer is yes, there are two options. We might believe that *C* is not found in the unsampled *F*s. In that case, *C* is a kind of confounding factor; we cannot proceed with the inference. But if factor *C is* found in all the unsampled *F*s, then it seems that we have something close to a *deductive* argument for the conclusion that all *F*s are *G*. We do not need to rely on our random sample any more. Further questions do arise; maybe we cannot rely on *C* to have the same effects in the unsampled cases...? But to say this is to embark on an entirely different kind of investigation from the statistical inference we started with.

These considerations have consequences for the way that philosophy has formulated and addressed the problems surrounding "induction." In this section, I have emphasized the differences between two

ize. See P.A. Schilpp and L.E. Hahn, eds., *The Philosophy of John Dewey* (Library of Living Philosophers, La Salle, IL: Open Court, 1939).

[21] For the bearing of natural kinds on induction, see especially W.V. Quine, "Natural Kinds" (in *Ontological Relativity and Other Essays* (New York: Columbia, 1969), reprinted in Stalker, ed., pp. ???–???); and also Hilary Kornblith, *Inductive Inference and Its Natural Ground* (Cambridge: MIT, 1993).

distinct kinds of scientific investigation and inference. One is statistical inference from samples. The other is inference about the structures and mechanisms responsible for the clustering of properties in a kind. Clearly these are not the only two kinds of investigation in science, but they are part of the story.

Statistical inference from samples is a natural tool for investigating the relation between two variables when the causal structure relating them is complex, mixed, and imperfectly known. Once we start to develop a reasonably unified story about what sort of causes operate in *F*s, and how these relate to the property *G*, it will be natural to move to a different kind of investigation. Then the multiplication of instances does not much matter. Inferences are based not on the power of random sampling, but on tracing specific dependences and causal paths.[22] We may introduce a new set of categories to organize the domain under discussion, too, when we switch methods; we may no longer treat "the *F*s" as the right class for analysis.

The philosophical concept of "induction," especially in the years after Goodman's discussion, seems often to be a mixture of these two separate kinds of scientific inference. Science contains one kind of inference where sample size *is* important and "naturalness" of kinds is *not*. In these inferences, it is the power of random sampling that gives us our link between observed and unobserved. Science also contains a second kind of inference in which sample size is *not* very important, but the status of the kinds under discussion *is*. Roughly speaking, it is the causal reliability of structures and mechanisms that gives us our link between the observed and unobserved. (Once again, I do not claim that these exhaust the inference patterns found in science.) The philosophical concept of induction in much recent discussion, however, *includes* the idea that the number of observed cases matters, does *not* include an explicit role for randomness of sampling, and *includes* a role for the "naturalness" of kinds. The result is a hybrid form of inference in which it is supposed to matter both how many *F*s you have seen *and* whether *F*s are a natural kind. But the reliable operation of inner mechanisms common to a natural kind and the randomness of sampling are *two distinct routes to projection*; two distinct ways in which the observed cases can connect us to the unobserved.

[22] In biology, the much-debated statistical concept of *heritability* is an example (see R.C. Lewontin, "The Analysis of Variance and the Analysis of Cause," *American Journal of Human Genetics*, XXVI (1974): 400–11). Heritability is slowly becoming less important to genetics, as particular causal pathways linking genes and phenotypic traits are uncovered in more detail.

I will summarize my main points. Jackson argued in 1975 that all predicates are projectible in principle, so long as inductions respect his "counterfactual condition." He argued that giving a theory of the projectibility of predicates is the wrong way to address Goodman's problem. The first part of my article is in partial agreement with Jackson's claims. In the context of statistical inference from the properties of a sample to the properties of a larger population, any predicate can be projected from a random sample, provided that the inference is not affected by biased sampling, confounding (in my broad sense of the term), or similar problems. However, grue-like predicates tend to create, via their peculiar semantics, problems of just these kinds.

Goodman's problem also arises for a quite different kind of inference, which should not be conflated with the genuinely statistical cases. These are inferences about the characteristic features of a natural kind, generated by the causal and nomological mechanisms operating in that kind. The treatment of Goodman's problem in this second class of cases does still involve considerations of the sort Jackson described, but includes much more besides. It is a mistake to conflate these two distinct kinds of inference by positing a single category of "inductive" arguments, in which both sample size and "naturalness" matter.

PETER GODFREY-SMITH
Research School of Social Sciences/Australian National University
Harvard University

APPENDIX: FORMULATIONS OF THE JPB PROPOSAL

I will briefly discuss different formulations of the JPB proposal. The issue is not critically important, because I regard the JPB proposal not as itself the solution to (part of) Goodman's problem, but as an abbreviation, in the language of philosophical discussions of induction, of a treatment found in statistics and science. So I am uncertain about the significance of technical problems with JPB.

Here is my initial formulation of the JPB proposal.

(JPB)
(1) All *F*s that are *O* are *G*.
(2) There exists no property *C* (other than *G*, and weakenings and strengthenings of *G*) such that:
(2.1.) All *F*s that are *O* are *C*,
(2.2.) Some *F*s are not *C*, and
(2.3.) The *F*s that are *O* are *G* because they are *C*.

═══════════════════════════════

(3) All *F*s are *G*.

The main question is how best to formulate premise (2), especially clause (2.3). I used a "because" formulation. Other possibilities include using an "only because" formulation, using an explicit counterfactual, and using conditional probabilities. One source of trouble is that premise (2) can be made so strong that it results in the argument form being either deductively valid, or close to it, and hence a bad representation of good "inductive" arguments.

That problem is what motivates, for example, the restriction against weakenings and strengthenings of *G*, as permissible values of *C*. The following objection was given (in slightly different form) by Aldo Antonelli (personal communication). If we allow *C* to be a weakening of *G*, then, at least under many versions and construals of premise (2.3), the JPB argument form is deductively valid. This problem is clearest for versions of (2.3) that use an explicit counterfactual. Suppose we used the following:

(2.3*) If the *F*s that are *O* had not been *C*, they would not have been *G*.

Then if we are allowed to consider *C*s that are weakenings of *G*, one candidate for *C* is $O \vee G$. Clauses (2.1) and (2.3*) are true for that choice of *C*. But then for premise (2) as a whole to be true, clause (2.2) must be false for that *C*. Clause (2.2) says that some *F*s are not *C*, so in the case considered here, this will only be false if all *F*s are $O \vee G$. But given that we also know that the *F*s which are *O* are *G*, all the *F*s (observed or unobserved) must be *G*. The JPB premises deductively imply the truth of the conclusion.

So if premise (2) is formulated in a way that uses or implies a counterfactual of that kind, we need a restriction on values of *C*. Similar problems arise if we allow candidate *C*s that are strengthenings, rather than weakenings, of *G*. (Consider $C = F \,\&\, G$.) In the main text I used a "because" formulation that is easy to grasp in many cases, but which is also less precise and may have an uncertain relation to the counterfactuals. In any case, it should be intuitively clear why the need for a restriction arises. The point of the JPB formula is to block inductions in cases where some extra dependence relation is affecting the *G*-ness of the objects in our sample. But various near-relatives of *G* (not to mention *G* itself) will have close logical and counterfactual connections with *G*. These near-relatives can act as "pseudo-explainers" of the *G*-ness of the sampled objects. We do not want these to be considered as possible defeaters of an induction, so they are ruled out as relevant values of *C* in the JPB formula.

Another family of problem cases is due to John MacFarlane and Branden Fitelson (personal communication). Let $C = M_1 \vee M_2 ... \vee M_n$, where this is a disjunction of all the very specific mechanisms responsible for the *G*-ness of the *F*s in the sample. The threat is that reasonable-looking inductions will be disallowed because this sort of disjunction will always (or very often) provide a *C* that passes the test in the second premise of the JPB formula. I suggest, however, that if a *C* of that kind is put forward in an abstract, stipulative way, without adding more specific information about the case, there is no reason to believe that (2.3) is true. If the sampled *F*s had not been subject to the operation of $M_1 \vee M_2 ... \vee M_n$, who knows what they have been like? But this discussion of candidate *C* properties that involve causal mechanisms also takes us back to issues raised in the final section of the main text. Once we start to develop detailed individual explanations for why *F*s in the sample are *G*, then all hangs on whether *F*s outside the sample are subject to the same explanatory factors as those inside. If we have confidence that the explanation for the *F*s inside the sample being *G* holds elsewhere, then we do not need random sampling to support the conclusion that all *F*s are *G*; we can make a direct argument from the operation of these explanatory factors. But if the *F*s outside the sample are subject to different influences from those inside, in this respect, then we should not project from our sample at all.

A distinct problem is how strong the dependence relation expressed in (2.3) should be. It might be advisable to use a formulation that is explicitly weaker than the "because" I used. A "might" counterfactual is one possibility, but it may instead be better to move to an explicitly probabilistic formulation.