THE C. BOYDEN GRAY
Center for the Study
of the Administrative State
ANTONIN SCALIA LAW SCHOOL • GEORGE MASON UNIVERSITY

The Role of Judgment and Deliberation in Science-Based Policy

M. Anthony Mills

CSAS Working Paper 21-16

Facts, Science, and Expertise in the Administrative State

GEORGE MASON UNIVERSITY
ANTONIN SCALIA LAW SCHOOL

The Role of Judgment and Deliberation in Science-Based Policy

M. Anthony Mills[1]

I. Introduction

It has become a commonplace that public policy ought to be based on scientific evidence. So commonsensical does the idea seem that one is tempted to ask, "What on earth was [it] based on before?", as philosopher John Worrall put it in a slightly different context.[2] And yet, a moment's reflection reveals that the situation is rather more complex.

Public policy in a modern representative democracy such as ours is inevitably responsive to a host of pressures: popular opinion, constituent demands, practical constraints, local custom, institutional dynamics, interest-group activity, party loyalty, political ideology, and value disagreements of all sorts. For many proponents, science-based policy offers hope that political decision-making might be insulated from such pressures, and placed on a firm, objective foundation. By basing public policy on scientific evidence, in other words, we may minimize our political disagreements and thus arrive at optimal solutions to our shared problems.

Indeed, those in the public square often speak as though the correct policy interventions flow from our scientific knowledge with something resembling deductive certainty, leaving little, if any, room for doubt about the proper course of action. Any disagreements about the most effective means to address a given problem or how to balance the inevitable tradeoffs the proposed interventions entail may then be rejected as not simply incorrect but also unscientific, even anti-scientific. Such a view presupposes a belief about what counts as sound political decision-making. But it also presupposes (albeit often implicitly) a conception of the nature of scientific evidence and its role in such decision-making.

According to this view, scientific knowledge arises from the application of reason to empirical observations according to rule-like procedures that leave little, if any, room for human judgment. Scientific conclusions may be fallible, but only because of human bias or incomplete data. When successful, science offers a kind of repository of neutral evidence, insulated from the uncertainties, ignorance, and value disputes that beset our politics. Accordingly, "following the science" allows us to constrain, if not eliminate, the role of judgment—and thus of deliberation—in political decision-making. Judgment and deliberation thus come to be seen as unnecessary, at best. At worst, they are treated as obstacles to the implementation of sound policy.

The political rhetoric surrounding the coronavirus pandemic offers a case in point. A typical characterization of this crisis holds that political polarization, institutional dysfunction, and disinformation are preventing the implementation of "science-based" measures that would otherwise enable us to vanquish Covid-19. There is, of course, a lot of truth in this diagnosis—polarization, institutional dysfunction, and disinformation are real and corrosive forces that have stymied efforts to manage the crisis effectively. Yet, frequently coupled with this diagnosis is a further assumption: that were it not for "anti-scientific" attitudes exploited by demagogues and

[1] Resident scholar at the American Enterprise Institute; senior fellow at Pepperdine University's School of Public Policy. I am grateful to Phil Wallach, Gracie Mills, and the participants in a research roundtable at the C. Boyden Gray Center for the Study of the Administrative State for helpful comments on prior drafts.

[2] John Worrall, "*What* Evidence in Evidenced-Based Medicine?" Philosophy of Science, 69, Sept 2002, pp. 316–330, p. 316. Worrall was referring not to policy but to evidenced-based medicine.

political inefficiencies imparted by an outmoded political system, scientific evidence would translate into incontestable and effective policy solutions. However psychologically comforting or politically useful this notion may be, it misrepresents the nature of scientific knowledge and its role in practical decision-making—especially during a large-scale and multifaceted crisis such as the one we currently face.

Unsurprisingly, perhaps, such rhetoric has incited a populist backlash, which foments suspicion of scientific, medical, and public health institutions and resistance to policies informed by them. Pandemic skeptics exploit the uncertainties surrounding our scientific understanding of the virus and point to human error and expert disagreement as grounds for distrust in public authorities. They go further still and question the motives of scientific, medical, and public health experts, accusing them of bias and corruption. Such attitudes fuel and are fueled by disinformation and polarization, thus engendering more accusations of "anti-science" by the other side. What goes unnoticed, however, is that both sides of this demoralizing dialectic share the same picture of scientific evidence and its role in political decision-making.

Consider that coronavirus skeptics do not simply disagree with public health interventions, but demur that they are unscientific. For instance, skeptics claim that there is no evidence that "lockdown" policies work or that masks are effective—not simply that these measures unduly restrict individual liberty or are outweighed by their economic or other costs—or insist that mortality figures and hospitalization rates are based on unreliable or even falsified data. Like their political opponents, skeptics take it for granted that, were it functioning in our decision-making processes as it should, scientific evidence would leave no room for doubt about the proper course of action. In other words, both sides—in different ways and for quite different reasons—agree that our public decision-making is insufficiently scientific. But, in fact, it is our public understanding of science that is insufficient.

Scientific evidence is indeed vital to public policy. The pandemic has made this undeniable, if it was not already obvious enough. But science does not offer a repository of neutral evidence that arrives, ready-made, onto the political scene. On the contrary, scientific knowledge is an achievement, the result of a complex process in which the judgment of scientific experts—call it *expert judgment*—plays a decisive role. Utilizing such knowledge to make policy decisions is even more complex, requiring not only expert judgment but also the judgment of those non-experts—call it *non-expert judgment*—whose experience, knowledge, or know-how is also needed to deliberate well about the best course of action.[3] It follows that judgment and deliberation are not secondary, lesser processes, that we must rely on when integrating scientific evidence into the policymaking process. Rather, judgment and deliberation are essential to this process, in part because they are essential to science itself. Failure to appreciate this fact risks

[3] For the purposes of this paper, to say that someone exercises "non-expert judgment" does not necessarily mean that this person is not an expert, only that he or she is not an expert *in the scientific field at issue*. Thus, although a nuclear physicist and a molecular biologist are both scientific experts, the biologist is a non-expert when it comes to the field of nuclear physics and vice versa. This distinction becomes particularly important in practical contexts. For instance—to anticipate an example discussed below—if what is at issue is how to assess the data or conclusions of nuclear physics, then a nuclear physicist, rather than a molecular biologist, would possess the relevant type of expertise and, presumably, expert judgment. However, a decision about whether to build a nuclear power plant may require the judgment of nuclear physicists but also the judgment of "non-experts," including engineers, architects, environmental scientists, lawyers, and economists as well as elected officials and representatives from the local community. Some of these "non-experts" are of course experts in their own right, even scientific experts. But they are, presumably, "non-experts" when it comes to the field of nuclear physics. Nevertheless, their expertise—whether scientific or not or credentialed or not—is needed, in conjunction with the expertise of the nuclear physicists, to deliberate well about the problem at hand.

engendering unrealistic expectations about what scientific knowledge can accomplish in practical decision-making, thus inviting not only disappointment, distrust, and skepticism, but also bad policy.

In what follows, I will make a case for this alternative account of scientific knowledge by examining the role that expert judgment plays in scientific reasoning.[4] I will then consider what implications this account has for how we understand practical decision-making informed by scientific knowledge. I conclude by suggesting that integrating scientific evidence into public policy is by nature deliberative, a reciprocal process in which both expert and non-expert judgments must play roles, and which requires that both experts and non-experts act with prudence.

It should be emphasized at the outset that this is intended as a second-order analysis, a philosophical examination of the nature of scientific reasoning and practical decision-making based on scientific evidence. I will make no attempt to assess the science surrounding COVID-19 (or any other science, for that matter), or the efficacy of particular policy interventions. Many of the examples used in the analysis are taken from scientific domains far removed from our current crisis. Although the lessons drawn from them are applicable to this and other crises—and indeed to policymaking generally. My hope is that by reflecting a little more deeply on the nature of scientific expertise, we can better understand the relationship between scientific knowledge and public policy, including how to use such knowledge to improve our policy decisions and the processes by which we arrive at them.

II. The Role of Expert Judgment in Scientific Reasoning

A. The Scientist's Goals and Methods

To see how expert judgment plays a role in science consider first how a scientist goes about formulating a research agenda. Which problems are worth pursuing and which are not? When should a particular line of inquiry be abandoned as fruitless? A researcher's choice about research goals may be shaped by a number of factors, including personal, professional, financial, practical, and ethical considerations. Is the research valuable—either in its own right or because of its potential applications? Is funding available for it? Is there a likely chance of success in the given timeframe? Is the experimental design practicable—is it ethical? The type of judgment needed to make these decisions is not epistemic, although it will surely be informed by the scientist's experience, knowledge, and familiarity with the state of the field in addition to practical considerations. Such non-epistemic judgments play an indirect role in science: they are necessary preconditions for research, although they do not directly influence scientific reasoning.

Things get more complicated when we consider methodological issues. What methods are called for by the problem at hand? What kind of test or experimental design is needed? Would it be better to perform a randomized control trial or an observational study? Can the hypothesis be tested directly against the data or are computer simulations necessary? What kind of mathematical techniques are most appropriate? What's the best way to weigh the relative risks of false positives or false negatives in your test results? What kind of statistical model is best

[4] The sources for my account of scientific reasoning, and the role of judgment therein, are many and various, but include, especially: Henri Poincaré (1902); Émile Boutroux (1926); Pierre Duhem (1906); Émile Meyerson (1908, 1921); Michael Polanyi (1958, 1962); Thomas Kuhn (1973); Jerome Ravetz (1973); and, more recently, Nancy Cartwright (2009, 2012, 2019 as well as Cartwright and Hardie 2012 and Munro, Cartwright, Montuschi, and Hardie 2016); Heather Douglas (2000, 2008, 2012, as well as Douglas and Magnus 2013); and Harry Collins (2010, 2014 as well as Collins and Evans 2002, 2007).

suited to the target? When is it necessary to upgrade the experimental or testing apparatus and how is this assessment to be done?

Expert judgment is called for in all of these cases. As in formulating a research agenda, such judgment may be guided by non-epistemic considerations, including feasibility or ethics. For instance, no one can perform an experiment directly using black holes, so computer simulations may be needed; the latest diagnostic test may be too expensive or excessively sensitive for a given purpose; data may not yet be available on rates of reinfection for a new pathogen, limiting the range of useful statistical models; it may be impractical or unethical to include relevant subjects in a certain kind of test, etc. But such judgments also have an epistemic component.

For instance, a decision to use a stochastic rather than a deterministic model to study a disease outbreak may be informed by empirical knowledge about the behavior of the target population or grounded in background theory or both. Even the practical limitations on experimenting with black holes stem in part from what we know, scientifically, about black holes. Or imagine you are a clinical chemist operating a laboratory.[5] In this capacity, you may be called upon to make any number of methodological decisions in the course of your work that will influence the outcome of your research, such as whether and when to upgrade your testing apparatus. As Zweig and Campbell point out, such decisions "may require some judgment about diagnostic performance."[6]

Say that new research suggests an automated assay is more accurate than the more conventional one you have on hand. You decide to test the accuracy of the assays for yourself directly. What's the best way to go about doing this? You might try comparing the test results of the two assays. But should you take the existing assay as a baseline for the assessment of the results or the new one? Or should you instead test the accuracy of both against an independent "gold standard"—i.e., a test considered by the professional community to be the best available and so to provide a baseline for judging diagnostic performance?[7] What if such a test is too expensive or time-consuming to conduct? Or what if no gold standard is available?

These types of decisions need not be—and typically are not—paralyzing or irresolvable. On the contrary, they are the stuff of day-to-day scientific practice. The point, however, is that making such decisions requires the exercise of expert judgment. Such decisions may be aided by formal methods, heuristics, and professional best practices as well as informed by past experience, well-established theory, and empirical observations. Thus, the expert's judgments are neither arbitrary nor subjective, nor are they devoid of epistemic content. But they are judgments nonetheless.

B. Collecting and Interpreting Data

The methodological considerations outlined above suggest that expert judgment does not only play an indirect role in science, e.g., in formulating research goals. Expert judgment can also play a more direct role in scientific reasoning.[8] This point may be brought out more fully by considering something as routine to scientific work as collecting and interpreting data.

[5] This example comes from Zweig and Campbell 1993.

[6] Ibid., p. 563.

[7] See, e.g., Portney and Watkins 2015.

[8] See Douglas 2000 for the distinction between "direct" and "indirect" roles for judgment in science, and Steel and Whyte 2012 and Elliott and McKaughan 2014 for critical appraisals. Douglas's distinction comes into play in the context of the appropriate roles for epistemic and non-epistemic value judgments, whereas the focus of this paper is on the broader notion of "expert judgment," whether epistemic or non-epistemic. Moreover, the argument presented

i. Seeing as an Expert

Consider an example adapted from Pierre Duhem.[9] Imagine that you are basically scientifically illiterate and are visiting a physics laboratory to observe what science looks like in action. As you enter the lab, you notice a table strewn with instruments and materials, some of which you are able to identify—a battery, some wire wrapped in silk, vessels filled with a silvery substance, coils, and what appears to be a small iron bar with a mirror on it. This, you presume, is an experimental apparatus. Next you see the physicist plunge the metallic end of a rod into some holes; the iron bar begins to oscillate causing the mirror attached to it to reflect a beam of light onto a celluloid ruler. The physicist carefully observes the vibrating spot of light caused by the oscillating piece of iron and takes scrupulous notes.

As Duhem points out, if you ask the physicist what he is doing, he will not say, "I'm studying the oscillations of the piece of iron carrying this mirror"—even though that is precisely what *you* observed him doing.[10] Instead, he will say something like, "I'm measuring the electrical resistance of a coil."[11] It will not be readily apparent to you how his account matches up to your observations: you have not observed any electrical resistance (and don't really know what resistance is, anyway) and can only guess at how the equipment used by the physicist conveys this meaning to him. But if you press further and ask the physicist to explain what he means, he may reply that "your questions would require some very long explanations" or "recommend that you take a course in electricity."[12] His point, however dismissively expressed, is that to grasp the significance of the experiment requires that you have some prior acquaintance with the relevant science—e.g., an understanding of electromagnetism, and the kind of instruments used to measure electrical current and resistance. To perceive the experimental results as *scientific data*, in other words, requires that you be able to see as the physicist does—that you be able to see as an expert.

This example illustrates the fact that a scientific experiment requires not only careful observations but also a translation of these observations into a theoretical framework using a quantitative and symbolic language that is scientifically meaningful. Hence, as Duhem points out, an experimental result is not simply a "group of concrete facts," such as you observe in the physics laboratory, but rather the "formulation of a judgment interrelating certain abstract and symbolic ideas which theories alone correlate with the facts really observed."[13] This is what is going on when the physicist tells you that he is measuring electric resistance, rather than watching a vibrating point of light produced by the oscillating piece of iron. Because of the theoretical interpretation he brings to bear on the experiment, the physicist—the expert—is able to see phenomena to which you are, in some very real sense, blind.

here about expert judgment does not depend on an airtight distinction between direct and indirect roles, as the discussion of expert judgment in scientific methodology should make plain. See also Douglas 2008 for an excellent analysis of the role of values in expert reasoning generally.

[9] See Duhem 1906, p. 145. Duhem was a physicist and historian and philosopher of science from the early twentieth century. Note that his example could easily be updated to include digital rather than analog tools, without altering the epistemological point. But the use of analog tools is useful for illustrative purposes.

[10] Ibid.

[11] Ibid.

[12] Ibid.

[13] Ibid., p. 147.

We can, somewhat artificially, separate this process of theoretical translation into two stages.[14] The first stage consists in collecting data, for example, by taking readings off an instrument, such as an ohmmeter or galvanometer. The second consists of using these data to infer something about the phenomenon of interest, e.g., about electric resistance or electric currents or whatever. Expert judgment plays a role in both stages.

ii. Expert Judgment in the Collection of Data

Return to the physicist in the laboratory using an instrument to measure electrical resistance. Let's say also that the experiment he is performing has never been carried out before. How does he know if his instrument is giving him reliable readings?[15]

Some amount of error in his data is of course inevitable, since neither he nor his instrument is perfect. But how does he decide which errors are tolerable and which are problematic? How does he differentiate, for instance, between systematic error and those introduced by unknown sources (say, magnetic interference)? The physicist is unable to use the outcome of the experiment as a standard against which to assess the reliability of his data (as a student might when reproducing a well-known experiment), since the outcome is not yet known. His assessment that the apparatus is working "'well enough'" is, as Ravetz points out, a matter of "judgment ... based on his experience of that particular piece of equipment, in all its particularity."[16]

Here, again, such judgment is neither arbitrary nor subjective. Besides past experience, the physicist will rely on best practices, rules of thumb, and theoretical knowledge of the experimental context—including both practical and theoretical knowledge of the apparatus itself. Part of what it means to be an expert practitioner is to have cultivated the knowledge, habits, and instincts needed to assess one's instruments effectively in a given situation so as to reliably collect empirical data.[17]

Stay with our example. Imagine that you are no longer scientifically illiterate, but are now moderately familiar with the field of electromagnetism. (Say you have enrolled in college-level physics courses.) You will comprehend in broad strokes what the physicist in the laboratory is doing: you know what electrical current and resistance are, how to model them mathematically, what kinds of instruments can be used to measure them, etc. In other words, you are able to see as the physicist does, at least to some degree. But you still are not in a position to assess confidently if and when the data are reliable enough—to know, for example, if when the instrument registers no reading that indicates a flaw in the instrument or the absence of the phenomenon it is supposed to be measuring. You do not yet have sufficient knowledge of and experience with the relevant instruments and scientific theories. What you lack that the physicist possesses is expert judgment—something that is acquired not just through formal education but through experience and honed through practice.[18]

Scientific data, then, are not raw observations but rather the end result of a theoretical procedure in which expert judgment plays a role. And the further transformation of such data

[14] See Ravetz 1973, pp. 76–88.

[15] Ibid., pp. 76–78.

[16] Ibid., p. 77.

[17] It may be, in fact, the laboratory technician, rather than the principal research scientist, who possesses such expertise in most abundance.

[18] Ravetz, following Polanyi, calls this "tacit knowledge." For more on this theme, see Collins 2010.

into information—data that convey something scientifically meaningful about the phenomenon under study—involves yet more expert judgment.[19]

iii. Expert Judgment in the Interpretation of Data

A standard way to begin using data is to plot the data points on a graph and find a curve that fits them. Having done so, you can use the mathematical interpretation of the data, rather than the individual data points themselves, as the basis from which to make inferences or predictions about the phenomenon of interest. There are a variety of well-known techniques for doing this. Yet, mathematically, an indefinite number of curves may be made to fit a given set of (finite) data. How are you to decide which to use?

What you need is to find the right family of curves, and then to determine which one most accurately captures the target.[20] As a first attempt, you might try a curve that passes through as many of the data points as possible—for the sake of argument, let's say a curve from the family of parabolic equations. However, you know that, no matter how expert you are, there is bound to be "noise" in your data—errors inadvertently introduced by you or your instrument while taking measurements. A curve that fits the data too closely will capture some of these errors—that is, it will fail to distinguish between signal and noise. Such considerations might incline you to choose a simpler curve, instead, say one from the family of linear equations. And yet, if your curve is too simple, you will wind up leaving out both noise *and* signal.

What you have hit on is a tradeoff between goodness-of-fit and simplicity. Increase the complexity of your curve, e.g., by selecting a second-degree rather than a first-degree polynomial, and you may capture more of the data. But you may thereby decrease accuracy. Increase the simplicity of your curve, e.g., by selecting a first-degree polynomial, and you may increase accuracy, although you will capture less of the data. How are you to strike the right balance? Note that you cannot simply compare your preferred curve to the "true" curve behind the data points, since that is precisely what you are trying to determine.

There are a host of formal methods that can aid your decision-making. For instance, the Akaike Information Criterion (AIC) provides a mathematical procedure for determining the right level of complexity (understood in terms of the number of parameters) when choosing a statistical model.[21] As Forster and Sober put it, Akaike's theorem "states a precise rate-of-exchange between these two conflicting considerations [goodness-of-fit and simplicity]; it shows how the one quantity should be traded off against the other."[22] Using the AIC, we can say precisely why and when too much complexity leads to overfitting and why and when simpler models are more likely to be accurate. This method provides a rule-like criterion for selecting which curve best fits the data. But it is only one method among many.

How are you to decide which method to employ in a given situation? Should you use the AIC or some other one, such as the Bayesian Information Criterion or a likelihood-ratio test? Like all formal methods, these rely on certain assumptions, e.g., about what constitutes simplicity, the significance of information loss, etc. The methods themselves do not provide a rule-like criterion for deciding whether and when these assumptions apply.[23] Should you

[19] See Ravetz 1973., p. 83. Ravetz distinguishes between "data" and "information." Douglas and Magnus 2013 make a similar distinction between "data" and "phenomena," following Bogen and Woodward 1988.

[20] See Forster and Sober 1994.

[21] I'm using the term "model" here in the statistical sense of hypotheses that have at least one adjustable parameter (see Sober 2004, p. 16).

[22] Ibid., p. 11.

[23] See Douglas and Magnus 2013 p. 11.

understand simplicity in terms of the number of parameters or some other way? Must you always prefer simplicity? The suitability of your formal method will depend, in part, on your scientific training (some fields might standardly use certain methods, while other fields prefer different ones), your knowledge of the particular situation, and your broader scientific aims. In other words, the use of formal methods does not eliminate the need for expert judgment.

To reiterate a now familiar theme: These types of decisions need not be and usually are not paralyzing. On the contrary, they are part of the everyday practice of science. Yet, making such decisions, and doing so effectively and without excessive deliberation, requires expert judgment.

C. Testing Hypotheses

So far, we have been considering the roles expert judgment plays in the early stages of scientific inquiry—in choosing a research program, a method, and in collecting and interpreting data. We turn now to one of the central aspects of scientific practice: testing hypotheses.

It is popularly believed that scientific inquiry proceeds in a stepwise, rule-like manner: make empirical observations, formulate a hypothesis, and then test the hypothesis by experiment. If the hypothesis is confirmed, accept it as is true; otherwise, reject it as false. The preceding discussion shows why the first part of this characterization is inadequate. Rather than formulating hypotheses on the basis of raw observations, scientists make observations against the backdrop of theories, using complex instruments and mathematical techniques that require expert judgment for their operation. For similar reasons, we shall see that the second part of the above characterization of scientific inquiry is also inadequate. Like collecting and interpreting data, testing hypotheses is a complex theoretical procedure in which expert judgment plays a role.

A traditional way to understand how to test a scientific hypothesis is as follows. First, derive an empirical prediction from the hypothesis, and then design an experiment to test the prediction. Next, run the experiment and observe whether the prediction obtains. If it does not, then the hypothesis may be rejected as false. This procedure follows a rule of logical inference known as modus tollens: If your hypothesis, H, is true, then a given observation, O, will obtain. Now, if O does not obtain, then we can infer that H is false.[24]

Consider a hypothesis proposed by the German scientist Franz Ernst Neumann in the late nineteenth century.[25] Neumann supposed that the vibration of a light ray is parallel to the plane of polarization. Call this hypothesis H_N. Otto Wiener was skeptical of H_N and so decided to subject it to an experimental test. First, he derived an empirical prediction from it: If H_N is true, he reasoned, then if a beam of light reflected at 45 degrees from a mirror is made to interfere with the incident beam polarized perpendicularly to the plane of incidence, dark and light interference bands will appear parallel to the mirror. Call this O_N. Wiener then devised an experiment to check whether O_N obtains. He carried out the experiment, observed no such interference bands, and so concluded that H_N must be false.[26] The vibration of a polarized light ray is not parallel to the plane of polarization.

[24] Schematically:

$H \rightarrow O$
$\sim O$
$\therefore \sim H$ (MT)

[25] This example is also from Duhem 1906, pp. 184–185.

[26] Schematically:

$H_N \rightarrow O_N$

This account of the reasoning behind hypothesis testing—and Wiener's experiment, in particular—overlooks several important complexities. Notice that to check whether a prediction obtains requires running the experiment and interpreting its results. But this, we saw above, involves theoretical translation, i.e., using background theories to interpret or make sense of the instruments, experimental setup, and resulting data. Deriving an empirical prediction from a hypothesis also requires assuming "auxiliary hypotheses," i.e., hypotheses other than the one being tested. For instance, in deriving and testing his prediction, Wiener assumed not only H_N, but also a theory of optics, including, for example, a hypothesis about the angle of incidence, and various hypotheses concerning the behavior of light, including the hypothesis that light intensity is measured by the mean kinetic energy. (Call the last hypothesis H_L.) Moreover, experiments rely on what are called ceteris paribus conditions—such as that no external forces are contaminating the experimental results—and idealizing assumptions, such as Wiener's assumption that the plane mirror used in his experiment is perfectly reflective.[27]

Strictly speaking, then, it is not a single hypothesis that gets tested in an experiment, but rather that hypothesis combined with these other, auxiliary hypotheses and background assumptions.[28] For this reason, if the prediction derived from the hypothesis is not observed, it is logically possible that it is not that hypothesis but rather one of the auxiliary hypotheses that has been refuted.[29] Put differently, observations *underdetermine* which is the incorrect hypothesis. In Wiener's case, by subjecting H_N to experiment, he was, in effect, also subjecting whatever other hypotheses he was taking for granted to experiment as well, including H_L. So, it is logically possible, based on his failure to observe O_N, that it is not H_N but one of these other hypotheses—such as H_L—that should be abandoned. (This, it turns out, is precisely what Henri Poincaré argued in defense of Neumann's hypothesis: that H_N may be preserved by rejecting H_L instead.[30]) It is also logically possible that one of the hypotheses from the theory of optics has been disconfirmed (however unlikely that may be) or that the experimental setup was flawed or that the instruments were not working properly or that some unknown errors were inadvertently introduced in the course of the experimental test.[31]

Such underdetermination of theory by observation has been used by philosophers and sociologists to draw any number of sometimes quite outlandish epistemological and metaphysical consequences, from the irrationality of scientific reasoning to the non-existence of the entities postulated by scientific theory. We can leave such questions to one side. What matters here is that, like the other types of uncertainty described above, underdetermination is part of the ordinary practice of experimental science. As such, it obviously does not undermine the scientific enterprise any more than the other types of uncertainty faced by the scientific expert discussed above—else there would be no such thing as successful hypothesis testing.

$\sim O_N$
$\therefore \sim H_N$ (MT)

[27] For the role and nature of ceteris paribus conditions in science, see, especially, Nancy Cartwright 1983.

[28] Schematically, what we have is not $(H \rightarrow O)$ but $((H_1 \ \& \ H_2 \ \& \ H_3 \ldots \ \& \ H_n) \rightarrow O)$.

[29] Logically, if we have not $(H \rightarrow O)$ but $((H_1 \ \& \ H_2 \ \& \ H_3 \ldots \ \& \ H_n) \rightarrow O)$, then it does not follow that $(\sim O \rightarrow \sim H)$. Rather, what follows is only $(\sim O \rightarrow \sim (H_1 \ \& \ H_2 \ \& \ H_3 \ldots \ \& \ H_n))$, which by De Morgan's Laws becomes $\sim H_1$ or $\sim H_2$ or $\sim H_3 \ldots$ or $\sim H_n$.

[30] Poincaré 1891.

[31] Of course, one can always check whether it is one of these other scenarios which obtains. But then the very same underdetermination considerations will come into play when conducting these tests. That is, one can pick out a different H to test, such as H_L, but one does not thereby isolate that hypothesis from the bundle of other hypotheses any more than Wiener isolated H_N.

What underdetermination *does* suggest is that there is no rule-like criterion for determining which hypotheses to accept and which to reject; expert judgment must play a role.[32]

Consider that in the course of testing a hypothesis, a scientist may make any number of decisions, sometimes unconsciously, about how to interpret recalcitrant observations. Should such observations be considered counter-instances to the hypothesis or anomalies that have yet to be explained or just noise in the data? Something like reasonableness or prudence comes into play here—Duhem called it good sense.[33] For instance, it would be unreasonable—imprudent—to throw out a battle-tested hypothesis on the basis of a single recalcitrant observation. On the other hand, it would be unreasonable or imprudent to cling obstinately to a hypothesis that has been repeatedly disconfirmed, simply because it is logically possible to reformulate one's background theory to accommodate the experimental data. This is precisely how scientists reason when they condemn a theory or hypothesis as *ad hoc*, even though it covers the same set of facts.[34] As Duhem points out, "pure logic is not the only rule for our judgments; certain opinions which do not fall under the hammer of the principle of contradiction are in any case perfectly unreasonable."[35]

Let's say, for example, that you are a relatively inexperienced scientist who has formulated a brand new hypothesis. You derive a prediction from it and design an experiment to test the prediction. To carry out your test you have to employ a standard set of instruments and experimental techniques, and you have to assume a number of well-established theories and auxiliary hypotheses. You run the experiment, and it appears to disconfirm your hypothesis. Now, even though it is logically consistent, it would be unwise to conclude on the basis of this observation that it was, in fact, not your hypothesis that was falsified but one of the well-established hypotheses you took for granted. Such a conclusion may be perfectly logical, but not therefore reasonable. Good sense would suggest that, at the very least, you should first check your experimental design, test your instruments or your handling of them, and run the experiment again.

To say that judgment plays a role in hypothesis testing does not mean that this process is arbitrary or subjective, any more than the process of choosing or assessing a method or of collecting or using data is arbitrary or subjective. As in these other cases, practical realities and the broader scientific context will constrain the scientist's judgment to some degree—by

[32] Some have suggested that formal methods, including the very same methods used in model selection, can be applied to cases of alleged underdetermination, thus providing a rule-like criterion (see, especially, Sober 2004). This suggestion poses additional complications that cannot be addressed adequately in the current paper. But suffice it to note that the same argument about the use of such formal methods in model selection would apply, *mutatis mutandis*, to hypothesis selection. See Douglas and Magnus 2013, p. 11 and Douglas 2012, p. 147.

[33] Duhem 1906, pp. 216–218. Duhem's term is *bon sens*. Although he borrows the concept from Pascal, the term itself has a rich and variegated history. It is sometimes used to translate the Greek *phronesis* or Latin *prudentia*. Descartes also uses the term (in a quite different sense) in the famous opening pages of his *Discours de la méthode*.

[34] Occam's razor is another principle that may be invoked in such situations. However, as we saw above in the discussion of model selection, there may be reasonable disagreements about what constitutes parsimony or simplicity. Similarly, predictive or explanatory power are often invoked, but these, too, may be given divergent interpretations or weights by different experts in different contexts. The literature on the criteria for theory choice is enormous. For a selection, especially concerning the role of Occam's razor, simplicity, and explanatory and predictive power, see: Poincaré 1902; Duhem 1906, 1908; Meyerson 1908, 1921; Einstein 1918, 1933; Reichenbach 1920; Schlick 1938; Carnap 1950; Quine 1951; Popper 1959; Kuhn 1962; 1983; Lakatos 1970; Van Fraassen 1980; Sober 1999, 2000, 2002, 2004; Forster and Sober 1994; and Douglas and Magnus 2013. For the role of values and non-epistemic factors in theory choice, see, e.g., Rudner 1953; Jeffrey 1956; Levi 1960; Kuhn 1973; McMullin 1976; Shapin and Shaffer 1982; and Douglas 2000.

[35] Duhem 1906, p. 217.

narrowing the range of plausible hypotheses, for example.[36] Formal statistical methods may be of assistance, too. And experience and background knowledge will be indispensable. Yet, however limited by practical and epistemic factors, the role of expert judgment in scientific reasoning nevertheless remains.

III. The Implications of Expert Judgment for Non-experts

The foregoing discussion suggests that expert judgment is essential to scientific reasoning. That is to say, judgment not only shapes science indirectly, but also plays a direct role in some of the most important aspects of the scientific enterprise—including making observations and drawing conclusions from experimental results. This has several implications for understanding the relationship between scientific knowledge and practical decision-making (and policymaking in particular).

First, if scientific conclusions depend on expert judgment, then non-experts who rely on such conclusions must defer to expert judgment, at least when it comes to the truth of the scientific conclusions in question. Consider, again, our example of the physicist experimenting with electricity. The non-expert onlooker is, in some very real sense, incapable of perceiving, much less performing, the experiment being carried about by the physicist. She has little choice but to trust the physicist's interpretation of the experimental evidence.

The non-expert may of course evaluate the expert's judgment in a kind of second-order way. She might, for instance, assess the trustworthiness of the expert—e.g., does she have reason to believe the expert is lying to her? Or she might assess the reliability of his judgment, e.g., by considering whether he possesses the relevant kind of background and experience, or is well-regarded by his colleagues. In doing this, the non-expert is employing the same kind of good sense all of us use on a day-to-day basis, however implicitly, in assessing the reliability of our fellows. What the non-expert is not equipped to do is make the expert judgment in place of the expert—or even to evaluate his judgment scientifically. To make or evaluate a judgment about, say, the best statistical model to use or how to interpret anomalous data points or whether a hypothesis is disconfirmed by observation would require that the non-expert be either an expert herself, or at least something approaching it.[37] As we saw above, even an onlooker versed in the relevant scientific fields might not possess the expert judgment needed to carry out the experiment or evaluate its results.

Note that this is true, to varying degrees, even among scientific experts. Thus, a seasoned particle physicist entering the laboratory of a geneticist will be in a position not totally dissimilar to the non-scientist, even if she partly shares with the geneticist a common background language, including familiarity with experimental and mathematical techniques, statistical methods, etc. Because of that shared background language, the particle physicist may be better positioned to understand or even acquire the relevant kind of expertise than the non-scientist.[38] But unless and

[36] For instance, a Sober 2004, points out, in practice the scientist is often testing two different hypotheses against a shared backdrop of theories and hypotheses.

[37] On this theme see note 39, below.

[38] But not necessarily better positioned. Indeed, a scientific expert's disciplinary background could, in some cases, make it *more* difficult for her to acquire expertise in—or even to understand adequately—another scientific field because of the assumptions or biases she brings to the table by virtue of her training and experience. For instance, a nuclear physicist and an environmental scientist are both scientific experts trained in mathematical and experimental techniques. But they may nevertheless disagree on standards of evidence or methodology, with significant implications for how they assess a given practical problem or what policy recommendations they are likely to provide. In such cases, communication between scientific experts can become more difficult or breakdown

until she has acquired such expertise, the particle physicist is in no position, for example, to adjudicate between rival hypotheses in the latest genetics research. Like the non-scientist, she too must rely on the judgment of the geneticist.[39] This is, in fact, an important aspect of the scientific enterprise: scientific experts must be able to take for granted conclusions established by their colleagues working in both closely related and widely disparate fields. If every scientist had to establish every conclusion that she took for granted in her own research, she would not only have to be an expert in a dizzying array of scientific fields, she would also wind up spending her entire career trying to reinvent the wheel, rather than making new contributions to her field.

Second, although non-experts who rely on scientific conclusions are thus dependent on expert judgment, at least in some sense, it of course does not follow that expert judgment is infallible. On the contrary, expert judgment is fallible in the same measure that all human judgment is fallible. This is stronger than the more familiar claim about the empirical defeasibility of scientific conclusions. It is uncontroversial—even if underappreciated in political contexts—that scientific hypotheses are open to revision on the basis of further empirical observation. Thus, even if a given hypothesis is extremely well-established—let's say it has received repeated and diverse types of confirmation over the course of decades—it is still possible, however unlikely, that one day it will be abandoned on the basis of new empirical findings. The history of science offers an impressive array of such examples. But if expert judgment is necessary to establish scientific conclusions, then it is not just the possibility of disconfirming evidence, but also the reality of human fallibility that could undermine such conclusions.

Third, if expert judgment plays a role in establishing scientific conclusions, it follows that there is room for reasonable disagreement *within* science. For instance, scientific experts may differ on the proper choice of method, about what statistical techniques to employ or how to interpret or evaluate experimental findings. They may also disagree about which hypotheses or theories are confirmed or disconfirmed by experiment. Typically, such disagreement is kept to a relative minimum, at least within a mature field, by disciplinary consensus and shared standards of evidence and evaluation. In extreme cases, disagreements may be so fundamental that even

altogether. This is characteristic of what is sometimes called "post-normal" science (see note 41, below). For an excellent discussion of clashing disciplinary assumptions and related issues, see Daniel Sarewitz 2004. Jonathan Fuller 2020 provides a helpful discussion of such disciplinary disagreements or misunderstandings among clinical epidemiologists and infectious disease epidemiologists in the context of the coronavirus pandemic.

[39] In some cases, a person, such as a scientific expert lacking expertise in an adjacent subspecialty or even a non-scientist with the relevant background knowledge and experience, may develop enough familiarity with a certain area of expertise that she is able to keep up with the literature or to follow the intricacies of a particular experiment or intra-expert disagreement. This is what Harry Collins and Robert Evans 2007 refers to as "interactional expertise." Yet, this still differs from "contributory expertise"—the kind of expertise possessed by practitioners who are advancing the state of their field by making substantive contributions, including by conducting experiments and interpreting experimental findings, advancing and refuting hypotheses, and the like. As Collins and Evans point out, "interactional expertise" characterizes the situation of many scientific experts who are project leaders on research teams with dozens of scientists spanning fields and subfields. The concept of interactional expertise has important implications for thinking about the relationship between experts and non-experts in general and scientific experts and political decision-makers, in particular. Consider that science advisors or professional staff working in federal agencies often possess such "interactional" (as opposed "contributory") expertise, typically in combination with other types of expertise, whether in management, law, or policymaking itself. Such "interactional" experts often play important translational roles in science-based decisions or in rendering such decisions comprehensible to non-technical audiences, whether elected officials, members of the public, or the courts. Though important, these considerations go beyond the purview of this paper. (For relevant discussions, see Whyte and Crease 2010; Douglas 2012; Collins 2014, and notes 80 and 81, below.)

such shared standards are thrown into question. This is what the historian of science Thomas Kuhn called a crisis in science.[40] However, in a well-established field, the kinds of uncertainties that make expert judgment unavoidable and expert disagreement possible tend not to hamper the ordinary process of inquiry.[41] This, of course, does not mean that there is no human error in mature sciences (although, presumably, individual experience, knowledge, and know-how as well as shared professional and disciplinary standards and best practices should help reduce the likelihood of error). Nor does it mean that mature sciences are immune to revision on the basis of future empirical evidence. Mature sciences are not infallible, either.

Most of the time, these features of scientific expertise pose no significant challenges to non-experts. Consider scientists interpreting experimental results in particle physics or evaluating competing interpretations of quantum theory or appraising two rival theories in cosmology, such as the steady-state theory and the Big Bang theory. Few would dispute or decry the fact that only specialists are in a position to make the relevant determinations in these cases.[42] Similarly, the fact that experts can err or disagree in making these determinations might seem both obvious and uncontroversial. Thus Fred Hoyle and other partisans of the steady-state model disagreed with those who defended the Big Bang theory and turned out to be in error. This was seen by the public as a victory for the Big Bang theory, not a commentary on the reliability or trustworthiness of cosmologists.[43]

Many such scientific controversies have little, if any, non-epistemic consequences for anyone outside the relevant scholarly communities. The situation becomes considerably more complex when scientific knowledge is applied in circumstances that have do consequences beyond the laboratory.

A. A Medical Example

Suppose you have reason to believe you have been infected with a relatively rare disease. You go to your doctor, presenting symptoms consistent with such a diagnosis. Your doctor decides to administer a diagnostic test. Naturally, there is judgment involved in both your decision to seek medical care and your doctor's decision to administer the test, including perhaps what sort of test is most appropriate and how the test gets administered. What of the test result itself? Here, one might think, we have hit on an expert conclusion in the form of a binary test result that follows from a rule-like procedure. But the situation is not so simple.

[40] See Kuhn 1962.

[41] This is what Kuhn calls "normal science" (ibid.). In political contexts, we do not always have the luxury of drawing on "mature" scientific fields, where there is a well-established consensus. As Collins and Evans point out, the "pace of politics is faster than the pace of scientific consensus formation" (Collins and Evans 2002, p. 241). Thus, political problems often require that decisions be made in conditions of radical uncertainty, e.g., when there is no expert consensus or there is insufficient data or there are conflicting methodological values and standards. This kind of situation is sometimes described as "post-normal" science (see Funtowicz and Ravetz 1993). I concur with Collins and Evans that this situation is not necessarily the rule, and that a broader taxonomy may be useful for thinking about the types of scientific knowledge that are needed in political decision-making. However, as we shall see below, the uncertainties involved in integrating scientific knowledge into practical and especially political decision-making are significant enough even in cases where our scientific knowledge is quite robust and an expert consensus is available. (But see note 71, below.)

[42] An important exception to this point may be Darwinian evolution, although this is likely due to the outsized influence (perceived or real) that this area of scientific research has on religious belief.

[43] Note that although the discovery of the microwave background radiation was widely seen to have dealt a death-blow to the steady-state model, proponents of that model did not think so. On the contrary, steady-state theorists had their own—albeit ad hoc—way of understanding and accommodating this bit of evidence.

Of course, no diagnostic test is perfect. This means that for any given test result, there is some possibility that the result is erroneous.[44] For instance, the test could return a positive result even though you do not have the disease, or it could return a negative result even though you *do* have the disease. Since the test is dichotomous, there are, in fact, four possible outcomes: A true positive result, a false positive result, a true negative result, or a false negative result. How does one know, in a given case, which outcome has occurred?

Test performance is estimated quantitatively using the concepts of "sensitivity" and "specificity."[45] Sensitivity measures the test's ability to identify patients who do, in fact, have the disease—the rate of true positives. Specificity, by contrast, measures the test's ability to identify patients who do not, in fact, have the disease—the rate of true negatives. For instance, a test with high sensitivity, say 95 percent, will correctly identify almost all of the patients who have the disease—95 percent of them. It will "miss" only 5 percent of them—these are false negatives. A test with 95 percent specificity would correctly identify 95 percent of those who do not have the disease in question. But it would return false positives for 5 percent of them. Together, these two measures allow one to estimate the probability that a given test result is erroneous.

There is usually a trade-off between these two measures. A test that is better able to identify true positives (higher sensitivity) will typically have more false positives, whereas a test that is better able to identify true negatives (higher specificity) will typically have more false negatives. How should one strike the proper balance between the two? What is needed is a decision threshold, which sorts the outcomes into binary classifications, so that results over a given threshold are classified as "positive" and those falling under it are classified as "negative." There are a variety of formal methods that may be used for this. Among the more well-known is the receiver operating characteristic curve (or ROC curve), which plots the true negative rate and the false negative rate in order to establish an optimum tradeoff between false-positive and false-negative results.[46]

But besides the selection of formal methods for establishing a decision threshold, the expert must also make a decision about how to weigh the costs of making incorrect or correct classifications. This, of course, requires judgment—a judgment about how to assign such costs. And, as Zweig and Campbell note, "assessing or assigning cost to false-positive or false-negative classifications is complex. This can be expressed in terms of financial costs or health costs and can be viewed from the perspective of the patient, the care providers, the insurers, dependents, society, etc. … [S]ome judgment about the relative costs of false results should be made when selecting rationally an operating decision threshold."[47] While the expert may rely on epistemic considerations and formal techniques in making this decision, he cannot avoid making a value judgment—one that will influence the outcome of the test.[48]

[44] See Qian, Refsnider, Moore, et al. 2020.

[45] See Ranganathan and Aggarwal 2018a.

[46] See Zweig and Campbell 1993, p. 561. The ROC curve was developed by the U.S. Army during World War II for using radar systems to detect enemy aircraft. See also Qian, Refsnider, Moore, et al. 2020 and Zweig, Ashwood, Galen, et al., 1995.

[47] Ibid., p. 572.

[48] This is what is at issue in the current debate over what kinds of coronavirus tests are most appropriate in different settings. (See Apoorva Mandavilli, "Your Coronavirus Test Is Positive. Maybe It Shouldn't Be," *The New York Times* September 17, 2020.) Note that this question turns on technical considerations (e.g., concerning the sensitivity and specificity of diagnostic tests, the "cycle threshold" of PCR tests, which tests are most reliable for which purposes), practical challenges (e.g., how to manufacture and distribute the appropriate tests at scale, quality control in non-clinical settings), political factors (e.g., who should be responsible for procuring the tests, the federal

Note that sensitivity and specificity are functions of the true state of affairs—e.g., of the number of true positives and true negatives. With them, one can calculate the probability that the test will yield a positive result, *given* that a patient has the disease, or that a test result will yield a negative result, *given* that a patient does not have the disease. But, of course, the true state of affairs is often not known. How, then, does one assess test performance? In the ideal case, there is a gold standard against which the test may be measured. But gold standards are not always available. And, of course, gold standards themselves typically provide only the best available estimates of the true state of affairs (or "ground truth").[49] So what's the best way to proceed? The situation is parallel to that of the clinical chemist working in the laboratory from our example above. In both cases, some amount of expert judgment will be necessary to assess diagnostic performance, including how best to evaluate the testing apparatus and to determine whether and how to upgrade it. Such judgment will, of course, be informed by both theoretical and empirical knowledge, formal methods and heuristics, as well as professional best practices (including what, if anything, constitutes a gold standard).

Sensitivity and specificity enable one to evaluate test performance—they provide information about the test itself. But the doctor interpreting your test results wants to know more than this. He wants to know the probability that *you* have the disease, given your test results. To assess this—what is called the "posterior probability" or "post-test probability"—the doctor must also estimate the probability that you have the disease independently of or prior to the test result. This is called the "prior probability" or "pre-test probability."[50] With these two probabilities in hand—the probability of a positive or negative test result given the true state of affairs and the prior probability that you have the disease—it is then possible, using a well-known statistical technique, to calculate the probability that you have the disease based on your test results.[51]

But how is the prior probability estimated? The prior probability is typically understood in medical contexts such as the one we're considering in terms of the prevalence of the disease in a given population. Low disease prevalence translates into a low prior probability and high disease prevalence into a high prior probability. For instance, let's say that a disease is extremely rare in a given population. Then if a patient from that population presents with symptoms consistent with that disease, there is nevertheless a low (prior) probability that she does, in fact, have the disease. Of course, she *may* have it, and for that reason the doctor may administer a test to check.[52] But the prior probability—the pre-test probability—is used to calculate the

government or states?), and value judgments, namely how to weigh the relative costs of false positives and false negatives.

[49] See Portney and Watkins 2015 and Cardoso, Pereira, et al., 2014.

[50] See, for example, Zweig and Campbell 1993; Qian, Refsnider, Moore, et al. 2020; and Zweig, Ashwood, Galen, et al., 1995.

[51] Using Bayes' theorem ($\Pr(A \mid B) = \Pr(B \mid A)\Pr(A)/\Pr(B)$), one can calculate the post-test probability, thus: $\Pr(D+ \mid T+) = \Pr(D+)\Pr(T+ \mid D+)/\Pr(T+)$ and $\Pr(D- \mid T-) = \Pr(D-)\Pr(T- \mid D-) /\Pr(T-)$, where D+ represents the presence of the disease and D– represents the absence of the disease; T+ represents a positive test result and T– represents a negative test result; Pr(D+) and Pr(D–) represent the prior probability, or the probability of the disease's being present or absent, independently of the test result. These two equations then become: $\Pr(D+ \mid T+) = \Pr(D+)\Pr(T+ \mid D+) / \Pr(D+)\Pr(T+ \mid D+) + \Pr(D-)\Pr(T+ \mid D-)$ and $\Pr(D- \mid T-) = \Pr(D-)\Pr(T- \mid D-) / \Pr(D-)\Pr(T- \mid D-) + \Pr(D+)\Pr(T- \mid D=)$, respectively, and, given the sensitivity and specificity of the test, yield the desired probability. See, e.g., van der Helm and Hische 1979; Khamis 1990; Diamond 1999; Zweig and Campbell 1993; and see Qian, Refsnider, Moore, et al. 2020.

[52] The doctor could also choose *not* to administer a test. This might be for practical reasons, including the affordability or availability of the test or the amount of time needed to administer it (e.g., as compared with the urgency of the potential medical intervention), or reasons having to do with both the characteristics of the disease or the test apparatus. For instance, depending on the sensitivity and specificity of the test, if the prior probability is low

probability that she does have the disease given the test outcome—the post-test probability. Failure to appreciate the prior probability, e.g., basing the probability estimate solely on the symptoms of a particular patient or her test outcome, is an example of what is sometimes called the "base-rate fallacy" or "base rate neglect."[53]

We're assuming in our example that the disease for which you're being tested is relatively rare. So this would suggest that the prior probability that you have the disease is relatively low, even though your symptoms are consistent with a positive diagnosis. But what if the prevalence of the disease is unusually high in your town or neighborhood or your place of work? Let's say you work in a hospital where there are a lot of patients being treated for this particular disease. Or what if the disease is more prevalent among certain demographics? Such factors may increase or decrease the prior probability. And since the post-test probability is partly a function of the prior probability, the doctor's decision about how to weigh such factors will influence how he interprets your test results.[54]

How does the doctor know which factors to consider and how to weigh them? Let's say your test result is positive. If the doctor has estimated a low prior probability, based on the low prevalence of the disease in the general population, the post-test probability that you have the disease will be lower than if he had estimated a higher prior probability, based on the high prevalence of the disease in your place of work, say. But which is the correct prior probability? A patient can be included in any number of different populations or "reference classes."

For instance, you might legitimately be included in the class whose members are all residents of the United States; the class whose members are all residents of Maryland; the class whose members are all residents of Montgomery County; the class whose members are all residents of Bethesda; or the class whose members are all residents of the Kenwood neighborhood of Bethesda. Or you might be classified as a member of the subgroup of a local population that works in a particular environment or a member of a particular demographic, say men over age 65. These reference classes may be concatenated and refined. Thus, you might be included in the class of men over 65 who have high blood pressure and live in Kenwood and drive luxury SUVs. Depending on which reference class one chooses, different prior probabilities may be used, and so different post-test probabilities may result. But probabilistic

enough, the test results could be just as likely (or possibly even more likely) to return a false positive or false negative. In other words, test results could prove unreliable enough that a given diagnostic test would not aid but rather hinder the doctor's judgment. (See note 54, below.) What is relevant for present purposes is that expert judgment comes into play when deciding if and how to administer a diagnostic test, including which kind of test. And this judgment will be informed not only by familiarity with and knowledge of the testing apparatus and statistical techniques but also by what is known about the disease and the patient, including the patient's medical history.

[53] See Maya Bar-Hillel 1979 for a classic account. But see Koehler 1996 for a more critical appraisal, which takes up the complications posed by the reference-class problem, discussed below.

[54] If the prior probability is high or low enough, it can skew the results. For instance, say one hundred patients go in to get tested for a disease. The test sensitivity is 80 percent and specificity is 99 percent. If the prior probability is estimated to be 1 percent, based on the disease prevalence, then for every true positive the test will return a false positive, meaning that 50 percent of the positive test results will be erroneous (BMJ 2020;369:m1808). This is the logic behind clinical epidemiologist Carl Heneghan's claim that 50 percent of Covid diagnoses may be false positives ("How Many Covid Diagnoses are False Positives?", *The Spectator* July 20, 2020. For a critical response, see Chris York, "No, 90% Of Coronavirus Tests Are Not 'False Positives' And This is Why," *Huffington Post*, September 23, 2020).

methods do not tell you which reference class to choose. This is known as the "reference-class problem."[55]

A standard response is to suggest picking the narrowest or smallest reference class for which there are reliable statistics.[56] The problem with this suggestion is that there may be multiple reference classes of comparable sizes for which reliable statistics are available—and indeed multiple ways to interpret reference-class size itself.[57] Another proposal is to use "relevance" as the criterion for selecting the proper reference class.[58] However, like "narrowness" and "reliability," "relevance" will be relative to the circumstances at hand and potentially subject to disagreement.[59] In other words, even if such criteria as "narrowness," "reliability," or "relevance" are made sufficiently clear, they remain "context-dependent" and "sensitive to pragmatic considerations."[60] Others have proposed using formal methods to solve the reference-class problem, from model selection criteria to data mining techniques.[61] Such methods may be quite helpful, but they do not tell you whether or not to apply such methods—or which one to apply—in a given situation.

The reference-class problem does not necessarily show that probability estimates are unreliable, only that there is no rule-like criterion for making the correct one.[62] Once again, expert judgment is needed, this time to apply probabilistic methods correctly to the situation at hand. Such judgment may be aided by formal techniques and professional best practices. And it may also be informed by past experience, theoretical knowledge, e.g., about what demographics are most likely to be afflicted by the disease in question; empirical data, e.g., about the prevalence of the disease in different populations; and first-hand knowledge about you—including demographic characteristics, medical history, or facts about where you live and work.

For instance, let's say the disease in question is infectious and is more common among men over 65 with high blood pressure. Let's also say that you are a 68-year old man who works in a hospital where patients with the disease are being treated. There may be room for reasonable disagreement among experts about how to calculate the relevant probabilities. But it would obviously be a mistake to estimate the prior probability that you have the disease based on the low prevalence of the disease in the general population.[63] Thus, while disputes over the correct

[55] The reference-class problem is usually said to have been discovered by John Venn 1876, although the term was coined by Hans Reichenbach 1949. See Alan Hájek 2007 for a recent discussion of the philosophical implications of the reference-class problem. Although the reference-class problem is often said to pose a problem for frequentist interpretations of probability, in particular, Hájek shows that the problem nevertheless arises in various guises in other—e.g., Bayesian and propensity—interpretations of probability as well.

[56] See Venn 1876 and Reichenbach 1949.

[57] See Hájek 2007. Of course, if reference-class size is understood in terms of the number of members in the class, the smallest reference class is the one that includes only one member, namely you.

[58] See L. J. Cohen 1979, 1981a, and 1981b.

[59] See Koehler 1996, p. 11.

[60] Hájek 2007 p. 568.

[61] See Cheng 2009 for how model selection criteria might be of help, and a critical response in Franklin 2010. Franklin 2011 draws on recent data mining techniques, while Abbas-Aghababazadeh, Alvo, and Bickel 2018 propose an "adaptive reference class method" using a "bootstrap approach to estimate the optimal reference class."

[62] Hájek 2007 makes a more general philosophical argument that the reference-class problem shows that there is no such thing as absolute probability, only relative probability. Koehler argues that the reference-class problem (among other considerations) provides good reason to be skeptical of typical characterizations of the base-rate fallacy. We can leave such issues to one side here.

[63] Likelihood ratios are often used to calculate the post-test probability when the pre-test probability cannot be assumed to be equivalent to disease prevalence. (The positive likelihood ratio is the conditional probability of a positive test result given that the patient has the disease divided by the conditional probability of a positive test result

reference class can and do arise—sometimes with significant consequences[64]—in many cases, good judgment, based on knowledge, experience, and professional standards, and practical constraints will limit the range of plausible choices.[65]

In sum, while the outcome of a diagnostic test such as the one discussed above is binary, the results are anything but "raw data." On the contrary, they depend on the proper use and assessment of the test apparatus as well as a complex process of statistical interpretation—all of which requires expert judgment. In this sense, the medical example discussed here is similar to the previous examples drawn from "pure science," whether physics and chemistry or cosmology. As in those examples, the non-expert is dependent on the judgment of the expert, in some sense. (For instance, you are probably not in a position to interpret the results of your test yourself.) And, of course, the expert's judgment is fallible—he could be wrong. There could also be disagreement among experts (e.g., conflicting diagnoses or conflicting recommendations for courses of treatment). But there is at least one crucial difference. Unlike the judgment of the clinical chemist, the physicist experimenting with electricity, or the cosmologist debating the correct model of the universe, the judgment of your doctor has significant non-epistemic consequences, e.g., for you and your health.

As we shall see, however, it is not just the significance of expert judgment that grows as science moves outside the laboratory; the role of expert judgment itself grows, as does the need for a wider array of judgments and types of judgment by experts and non-experts alike.

B. Deciding That vs. Deciding To

The philosopher Ian Hacking once observed that there is a difference between determining which hypothesis is best supported by the evidence and deciding which hypothesis to accept for a given purpose. That's because "deciding that something is the case differs from deciding to do something."[66] For instance, let's say the doctor in our previous example determines that you do, in fact, have the disease. He decides, in other words, that the hypothesis that you have the disease is best supported by the evidence. You must then decide what to do about it.

Note, however, that the doctor does not simply provide you with neutral evidence and then leave you to decide. The distinction between "deciding that" and "deciding to" does not map neatly onto the distinction between the doctor's role and yours. The situation is more complex than that. To be sure, the doctor does assesses the evidence, e.g., based on your symptoms and the test results and using statistical methods. But having made that determination, he also makes a decision about what to do next: Redo the test, administer a different kind of test, or recommend a course of treatment. He might prescribe this medication or recommend that

given that the patient does not have the disease or Pr(T+ | D+)/Pr(T+ | D–); the negative likelihood ratio is the conditional probability of a negative test result given that the patient has the disease divided by the conditional probability of a negative test result given that the patient does not have the disease or Pr(T– | D+)/Pr(T– | D–). See, e.g., Ranganathan and Aggarwal 2018b. These may be related to sensitivity and specificity by: LR+ = sensitivity/1 – specificity and LR– = 1 – sensitivity/specificity.) Note that expert judgment still comes into play here, not only in deciding whether to use likelihood ratios but also in choosing how to calculate the post-test probability based on these ratios.

[64] For recent discussions of the legal significance, see Colyvan, Regan, and Ferson 2001, as well as Allen and Pardo 2007. For a critical discussion see Cheng 2009 and responses in Franklin 2010 and 2011. The debate is long-standing. See Koehler 1996, n. 16 for references.

[65] In this sense, Cheng 2009 is right to say that the reference-class problem may admit of practical solutions, even if the problem remains formally unresolved or unresolvable.

[66] Hacking 1965, p. 29, see also p. 27.

medical intervention. In other words, his expert advice concerns not only what is the case but also what to do about it. Nor is his assessment of the evidence entirely neutral. As we saw above, interpreting test results requires that the expert make a value judgment about how to trade off sensitivity and specificity and thus how to weigh the costs of false negatives and false positives.

Of course, ultimately, since it is your health that hangs in the balance, it is you who must decide how best to proceed. The doctor's recommendation can and should inform or even constrain your decision. But your dependence on him is far from absolute. Besides the fact that the doctor could be wrong, you may have information that he lacks—concerning your financial or family circumstances or your moral values—that may also influence your decision. In other words, the doctor's advice—including both his assessment of the evidence and his recommendation about what to do in light of the evidence—can and should factor into your decision about what to do. But it may not be the sole factor or, perhaps, the most significant one. The correct determination about what to do does not flow from the expert's advice with anything resembling deductive certainty. (Not even the expert's determination about what *is* the case flows from the evidence with *deductive* certainty.)

In political contexts, the situation is even more complex. Consider the case of a toxicologist offering advice on what constitutes a safe level of toxicity for a chemical, or a nuclear physicist offering advice on the feasibility of a controlled nuclear reaction, or an epidemiologist offering advice on whether to implement public health measures.[67] In these cases, as in our medical example, expert judgment will play a crucial role. And, again, the expert's judgment will not simply be a matter of assessing neutral evidence, but will include value considerations, e.g., about how to measure and weigh risks, including the risks of false positives and false negatives. Ultimately, the expert's judgment will issue in practical recommendations, e.g., about what policies are or are not best supported by the scientific evidence.

Thus, as in the medical example, it is not simply a matter of accepting the truth of a scientific claim that only an expert is well-positioned to confirm, as when a non-expert accepts the Big Bang theory over its rivals. The policymaker is called upon not merely to accept an expert's judgment about which hypothesis is best supported by the evidence or even to rely on that expert's recommendation about what to do. Rather, the policymaker must to decide what to do, *given* the expert's advice. Regulate or not? Grant a permit to build a nuclear power plant or not? Impose public health measures or not? The expert's recommendation can and should inform or even constrain the policymaker's choice—or range of choices—in these cases. But her decision about what to do must weigh factors that go well beyond toxicology, nuclear physics, or epidemiology.

Like the doctor's recommendation, the expert judgments at issue—and the fallibility of such judgments—have clear non-epistemic consequences on the world well beyond the laboratory. In contrast to the medical example, however, the potential consequences of these decisions—whether or not to regulate a chemical, build a nuclear power plant, or implement public health measures—go well beyond any one person's health.[68] Indeed, they may have wide-ranging societal or economic or public health or even political or ethical implications. Accordingly, while such decisions depend on expert knowledge, e.g., of the toxicologist, the physicist, the epidemiologist, they also involve a host of other factors, both epistemic and non-

[67] See Rudner 1953 and Douglas 2000.

[68] Of course, the medical example could be turned into a large-scale political challenge, for instance, by considering a global pandemic. In such a situation, questions concerning diagnostic testing that might ordinarily impact only test patients directly become matters of national, even global concern. See, e.g. notes 48 and 54, above.

epistemic, from estimating the likely societal or economic or public health impacts to feasibility, cost, or other practical factors to political or ethical considerations. In other words, to make such a decision effectively requires soliciting advice and information from a wide range of people and sources, expert and non-exert alike. To do that, it is necessary not only to be able to interpret and weigh such information and advice, but also to know where to look in the first place. What is needed is what moral philosophers call the virtue of prudence.

IV. Deliberating with Experts

Prudence is, as philosopher Herbert McCabe pithily puts it, the "virtue which disposes us to think well about what to do."[69] And thinking well about what to do usually requires seeking the counsel of others. This, in turn, requires knowing whom to consult and how to weigh such advice. In this sense, prudence is also "the virtue that disposes us to deliberate well."[70] Part of deliberating well is knowing with whom to deliberate.[71]

For instance, it would be imprudent for you not to consult your doctor when making a decision about whether or how to treat yourself for a given disease. In a similar way, it would be imprudent for a policymaker not to consult a scientist with relevant expertise when making a decision that requires scientific knowledge. But it would be imprudent to consult a scientist or a doctor who lacks any relevant expertise (an entomologist rather than a nuclear physicist, say, or a podiatrist rather than an oncologist) or who has an obvious conflict of interest or is a known liar or quack.

It also takes prudence to know how wide the sphere of deliberation should be. In our medical example, the appropriate sphere of deliberation is relatively narrow. Besides your doctor, you might consult your family or perhaps your priest, or decide to get a second medical opinion. But, presumably, it would be unreasonable to refrain from making your decision unless and until you had consulted your third grade social studies teacher or your car mechanic. By contrast, in most cases where a policymaker is called on to make a decision with a scientific or technical component, the sphere of deliberation will be considerably wider.

For instance, imagine you're a policymaker faced with the decision of whether and how to regulate a given chemical.[72] You decide to consult more than one expert in toxicity to get a range of opinions. For simplicity, let's say you consult expert A and expert B, and they are equally credible and trustworthy. But they disagree about what toxicity level poses a threat to public health. Specifically, expert A gives an estimate for dangerous toxicity levels that is lower than expert B's estimate.[73] If you accept A's advice, you risk overestimating the harm posed by the chemical and thus overregulating. While this would pose no direct harm to public health, it might very well have harmful economic consequences, for instance on the industries that use this chemical and the jobs they support. By contrast, if you rely on expert B's advice, you risk underestimating the harm posed by the chemical and thus under-regulating its use. This would leave the relevant industries unharmed but might be detrimental to public health.

[69] McCabe 2002, p. 342.

[70] Ibid., p. 343.

[71] The account of deliberation given here is broadly Aristotelian. For reasons of economy, I must leave to one side how this account may be extended to encompass or otherwise accommodate situations in which there is outright political or methodological conflict (e.g., as in "post-normal" science). For present purposes, it will suffice to note that even in the comparatively "nice" cases considered here, scientific evidence is rarely, if ever, dispositive.

[72] For an excellent discussion of the role of value judgments in such disputes, see Douglas 2000.

[73] See Douglas 2000, pp. 573–577, for an illuminating discussion of this issue.

Deciding what to do here requires weighing the potential harms of underestimating or overestimating risks. Clearly, it would be imprudent to ignore the advice of experts A and B in making this decision. Note that, as in the medical example above, these experts may well have useful advice not only about the evidence but also how to weigh the relevant risks.[74] But it would be no less imprudent to exclude from deliberation those non-experts who possess knowledge about the potential economic effects, including not only economists but also those familiar with the relevant industries, say, or about the broader impact the potential economic or public health effects might have on a given local community. Some of this knowledge will be technical, requiring formal education or training, but some of it will be professional or experiential or even local knowledge.

Consulting the relevant non-experts in making this decision is not simply a matter of fairness, e.g., canvassing the opinions of those who are likely to be impacted. Good practical reasoning involves deliberating well, which means knowing with whom to deliberate. In this case, prudence dictates consulting the relevant experts (e.g., chemists, toxicologists). But it also precludes consulting *only* such experts. To deliberate well you must also consult those non-experts whose knowledge or know-how is relevant to the decision at hand.[75] Scientific evidence cannot tell you how to do that, even when there is no dispute among experts about the scientific evidence.

To illustrate this last point consider yet another example, taken from an entirely different policy domain. Imagine, now, that you are a school administrator called upon to make a decision about whether or not to introduce some classroom innovation to your school.[76] Imagine, further, that there is a well-conducted randomized control trial (RCT) that shows this innovation yields better educational outcomes on average. RCTs can offer compelling evidence of causal efficacy. And they are widely considered to provide a kind of "gold standard" in both evidence-based medicine and evidence-based policy.[77] So it would appear you have solid scientific evidence that your intervention will be effective. But, of course, the situation is not so simple.

When successful, RCTs can establish the causal efficacy of a given intervention by estimating the average treatment effect. Since you are not a scientific expert, you are not really in a position to evaluate whether the RCT in question was well conducted. You have little choice but to defer to experts on this question. But that is not your question, anyway. Your question is

[74] Indeed, the question at issue is essentially a question about how to weigh the relative costs of false positive vs. false negatives. See Douglas 2000.

[75] Expertise, in this sense, should not be understood therefore as a simple binary classification, much less a socioeconomic one—as in the popular contrast between "the experts" and "the public." Consider that toxicologists, nuclear physicists, environmental scientists, economists, engineers, architects, car mechanics, accountants, lawyers, policymakers, and indeed, musicians, literary critics, and farmers all possess forms of expertise. Some of these forms of expertise—such as toxicology or nuclear physics or environmental science—are "scientific" in nature whereas others—such as architecture or policymaking or literary criticism or farming—are not (at least according to the standard understanding of the term "scientific"). Similarly, some forms of expertise—such as toxicology or nuclear physics or economics or law or literary criticism or engineering or architecture—require formal education or credentials—whereas others—such as policymaking or farming—may be acquired primarily through experience or apprenticeship. The relevance of these divergent types of expertise for practical decision-making will depend on the problem at hand. Although the focus of this paper has been on *scientific* expertise, this, of course, does not imply that there are no other forms of expertise (see note 3, above).

[76] This example is due to Deaton and Cartwright 2018.

[77] This is not the only reason RCTs are considered a gold standard. They are also touted for their ability to limit sources of bias. See, e.g., Chalmers, Smith, Blackburn, et al. 1981, and Moher, Hopewell, Schulz, et al. 2010. See Worrall 2002 and Cartwright 2010 and Deaton and Cartwright 2018 for critical appraisals of the use of RCTs as a gold standard and the related "hierarchy of evidence."

what to do even if you have a well-conducted RCT in hand. Assuming that the RCT *is* well conducted, it provides evidence that the intervention works *somewhere*. But what you want to know is whether it will work *here*, in your school.[78] And that requires, first, establishing whether or not your school is relevantly similar to the experimental population used in the RCT.[79]

To make such a comparison, you will have to consult a wide range of sources, including, again, both experts and non-experts. For instance, you will need to know something about the comparative sizes, demographics, and socio-economic makeups of the relevant populations as well as institutional facts, e.g., about the resources and capacities of the schools. This may require drawing not only on statistical information but also your own knowledge and experience as well as the professional and local knowledge of teachers, parents, and other administrators.[80] As in the toxicity example above, weighing these diverse factors and knowing whom to consult requires both good judgment and sound deliberation. Of course, even if you are able to reach a conclusion about whether or not the policy could be effective in your school, you still have to determine whether you have sufficient means, know-how, and "buy in" to try it—and how best to go about doing so. Implementation requires yet more judgment and deliberation.

V. Conclusion

Science is indispensable for public policy. But scientific evidence, no matter how robust, can never replace, only inform, judgment and deliberation. And this is, in part, because scientific expertise itself depends on such judgment and deliberation.[81] It follows that using scientific

[78] The distinction between "it works somewhere" and "it works here" is due to Cartwright. (See, for example, Cartwright 2012, p. 975.)

[79] Or, more accurately, that subpopulation of the experimental population for which the intervention was effective. Unless, of course, the experimental population was truly representative of the target population. But as Cartwright 2010 observes, this is surely the exception rather than the rule, especially in social policy contexts (p. 67).

[80] For two classic studies of the role of non-credentialed expertise in science-based policymaking, see Wynne 1996 and Epstein 1996. (See also Collins and Pinch 1998 and 2005 as well as Whyte and Crease 2010 for more recent discussions.) Too often, however, these and similar case studies are cited in the scholarly literature as evidence of the constructive role that non-credentialed expertise can play in science-based policy, without recognizing that these "lay" experts—such as Wynne's Cumbrian sheep farmers—can be and often are experts in their own right, even if their expertise is neither scientific nor credentialed. This implicitly narrow construal of expertise perhaps explains the tendency to utilize such examples to blur or even altogether abolish the boundary between "experts" and non-experts, rather than recognizing scientific expertise as one important form of expertise among many. (However, see Collins and Evans 2007, especially chapter 2, for a balanced discussion of these issues and a helpful taxonomy of expertise.) Both Cartwright and Hardie and Munro, Cartwright, Montuschi, and Hardie 2016 contain insightful discussions of the relationship between scientific and professional expertise, including the importance of judgment and deliberation. Such professional expertise (whether credentialed or not) is especially important in policy contexts, and is deserving of separate treatment (see note, 81, below). For helpful discussions of policy expertise and its relationship to scientific expertise in the context of executive agencies, see, e.g., Shapiro 2015 and Adler 2020 (forthcoming); in the context of congressional agencies, see, e.g., Bimber 1992; Morgan and Peha 2003; Blair 2013; Graves and Mills 2019; and Mills 2020.

[81] To say that judgment and deliberation are essential to both scientific expertise and political decision-making is not to say that there is no distinction between the two. On the contrary, we have seen that there is a rather stark difference between the expert capable of judging well and the non-expert who must rely on such judgment. The converse is also true. That is, there is a stark difference between the type of expertise needed to govern well and the type of expertise needed to be a good research scientist or good scientific advisor. Part of that difference turns on the outsized role of judgment in deliberation in the political sphere and the size and complexity of the challenges faced therein. And part of that difference turns on the differing goals, standards, and methods of these two forms of expertise. However, the focus here has been on the role of judgment within science and its implications for political decision-making. An adequate consideration of statecraft in general would go well beyond the limits of this paper. (See Wallach 2018 for an illuminating discussion of related themes.)

knowledge for practical purposes is never a matter of inputting neutral, ready-made evidence, but rather of relying on expert judgment.

When it comes to ordinary research science, e.g., if we want to know about the atomic structure of matter, the chemical composition of the stars, the age of the universe, or whatever, our dependence on expert judgment is relatively uncontroversial. But when it comes to expert judgments that have non-epistemic consequences, such as a medical diagnosis or advice about the safety of certain industrial practices or the need for certain public health interventions, the issue of deference can become fraught, and may give rise to unease, distrust, or resentment. Of course, these are precisely the types of expert judgments—rather than, say, those concerning which cosmological model to accept—that tend to be most relevant for policy decisions.

Making such decisions requires deliberation. And deliberating well requires knowing with whom to deliberate. Most of the time, when there is uncertainty about how to proceed within science—for instance, when experts disagree about which hypothesis is best supported by evidence—deliberation will naturally exclude not only non-scientists but also most scientists who are not familiar with or engaged in the relevant subfield.[82] In other words, the sphere of deliberation will be small and relatively homogenous. By contrast, making policy decisions informed by scientific evidence requires consulting not only the relevant experts but also those non-experts whose knowledge, experience, and know-how are needed for effective decision-making. In these circumstances, the sphere of deliberation may be quite large and heterogeneous.

Deliberation is a reciprocal process, with mutual obligations. Non-experts must be willing to trust expert judgment, although not blindly or uncritically. Experts, for their part, must be willing to make recommendations while being open and honest about the possibility of error, the role of judgment, and the nature and extent of disagreement within their field. Both must be realistic about what scientific evidence can achieve and the level of certainty it can attain. Failure to do so will only exacerbate unease, inflame distrust, and breed resentment.

Above all, experts and non-experts alike must exercise good judgment. This is most important for policymakers, whose job it is, finally, to decide on the proper course of action. Doing so requires deliberation—not only consulting experts but also knowing how to weigh their judgments and whom else to consult when deliberating about technical policy decisions and their potential implications. Scientific evidence is essential to this process, but rarely dispositive. Science, in other words, illuminates rather than eliminates the need for judgment and deliberation—especially when it comes to "science-based" policy.

[82] This is what Collins and Evans 2002 refer to as a "core set."