PROCEEDINGS OF THE ARISTOTELIAN SOCIETY
HOSTING AND PUBLISHING TALKS IN PHILOSOPHY SINCE 1880

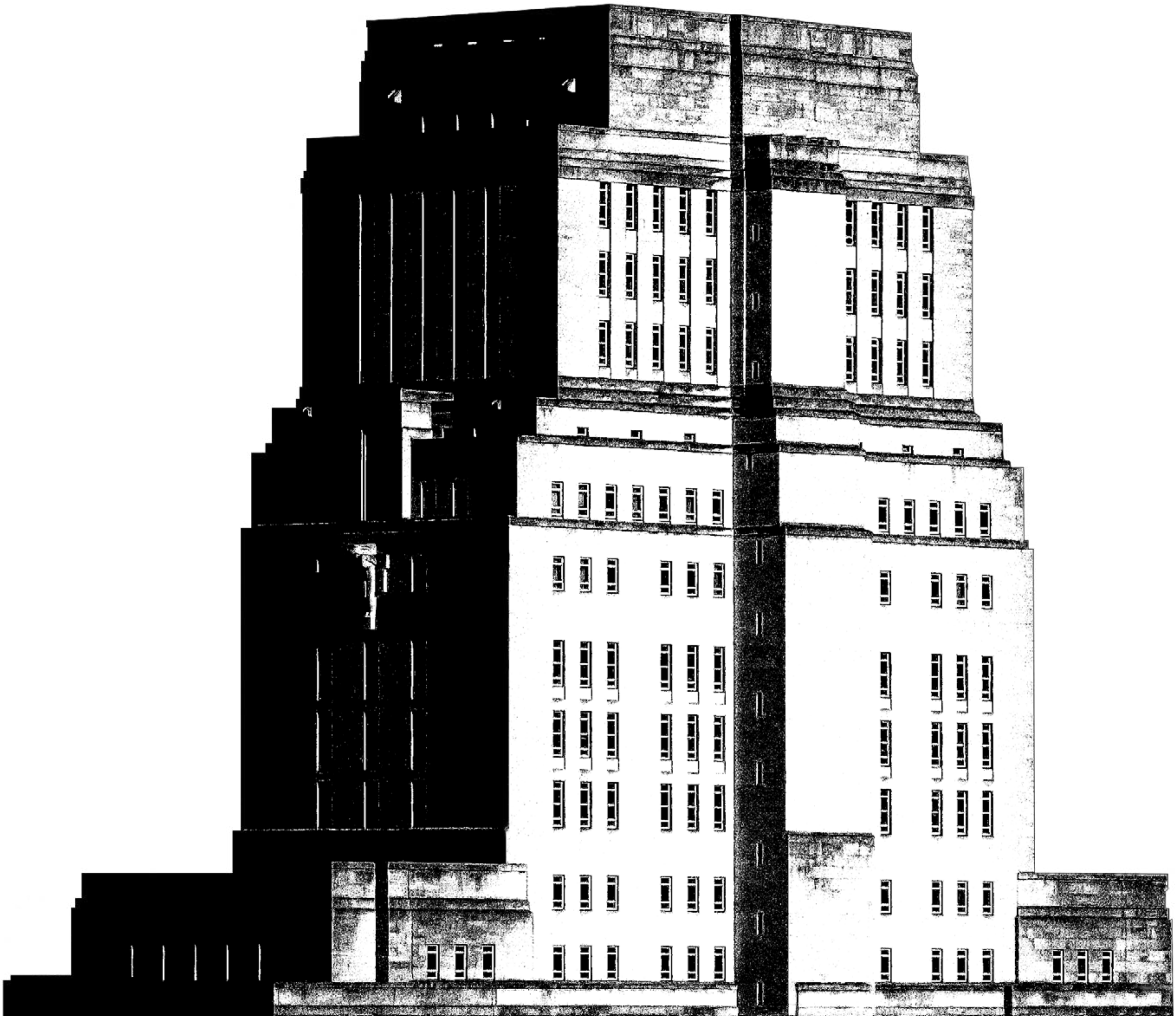
2019–2020
140TH SESSION
VOLUME CXX

CHAIRED BY HELEN STEWARD
EDITED BY GUY LONGWORTH

SENATE HOUSE
UNIVERSITY OF LONDON
THE WOBURN SUITE

Why Trust Science?
Reliability, Particularity and the Tangle of Science

NANCY CARTWRIGHT



PROCEEDINGS OF THE ARISTOTELIAN SOCIETY
140TH SESSION

ISSUE NO. 3
VOLUME CXX
2019–2020

WHY TRUST SCIENCE?
RELIABILITY, PARTICULARITY AND THE TANGLE OF SCIENCE

NANCY CARTWRIGHT
DURHAM UNIVERSITY

MONDAY, 27 APRIL 2020

17.30–19.15

ONLINE VIA ZOOM
Please visit our website for further details

CONTACT
mail@aristoteliansociety.org.uk
www.aristoteliansociety.org.uk

BIOGRAPHY

Nancy Cartwright is a methodologist and philosopher of the natural and human sciences, with special focus on causation, evidence and modelling. Her recent work has been on scientific evidence, objectivity and how to put theory to work. She is a Professor of Philosophy at Durham University and the University of California San Diego, having worked previously at Stanford University and the London School of Economics. Professor Cartwright is a former MacArthur fellow, a fellow of the British Academy and the Academy of Social Sciences, a member of the US National Academy of Sciences, the American Philosophical Society (the oldest honorary academic society in the US), the Academia Europeae and Leopoldina (the German Society for Natural Science). She has won the Hempel Prize for lifetime achievement in philosophy of science and with Elliott Sober, the Lebowitz Prize for Philosophical Achievement of the Phi Beta Kappa Society. She is Tsing Hua Honorary Distinguished Chair Professor in Taiwan and has been awarded honorary doctorates from the University of St Andrews and Southern Methodist University. Her latest books are *Nature, the Artful Modeler* and *Improving Child Safety: deliberation, judgement and empirical research* with Eileen Munro, Jeremy Hardie and Eleonora Montuschi.

EDITORIAL NOTE

The following paper is a draft version that can only be cited or quoted with the author's permission. The final paper will be published in *Proceedings of the Aristotelian Society,* Issue No. 3, Volume CXX (2020). Please visit the Society's website for subscription information: aristoteliansociety.org.uk.

WHY TRUST SCIENCE? RELIABILITY, PARTICULARITY AND THE TANGLE OF SCIENCE

In evaluating science philosophers tends to focus on *general laws* and on their *truth*. I urge a shift in focus to the *reliability* of the *panoply of outputs* science produces ('the *tangle* of science') and in tandem, from the *general* to the *particular*. Here I give 5 arguments to support this, including: a) many, many scientific outputs (measures, devices, empirical studies, model organisms, …), which must be supposed reliable if we are to warrant our general principles, aren't truth apt; and b) reliability invites the crucial question, 'Reliable for what?' Getting clear the *particular* purpose is essential for effective action and just evaluation.

JUDGING FROM the amount of attention awarded the topic, philosophers take establishing general theories and general laws to be the core of scientific endeavour. And not only philosophers. A National Academy of Science 2008 report urges: '[T]heories are the goal of science' [58]; the National Research Council 2002 includes 'Replicate and *generalize* across studies' [52; ital added] among the 6 principles of inquiry they use to characterize science; and the International Network for Natural Sciences claims: 'Basic research…Seeks *generalization*.'[ital added]. This paper urges us to refocus from the acceptability or truth[1] of general laws to the *reliability* of the *panoply of outputs* the sciences produce (for short, 'the *tangle*[2] of science'); and in the course of this, from the general to the *particular*. Here I present a number of reasons for this shift.

Theory at some level or another always plays a part, but a vast network of other kinds of scientific outputs are required, working as a team, to underwrite the success on any particular scientific endeavour, from designing a satisfaction survey for a new teaching method for your metaphysics module to building the large hadron collider at CERN. This means that philosophy needs to be concerned with *all* the products of science, from theories and models to concepts and measures, studies and experiments, data collection, curation and coding, methods of inference, narratives, devices, technologies, plans and science-informed policies and more. I do not urge attention to these just because every labourer is worthy of their hire but because each successful endeavour in

1 All I say applies whether we look at truth or any of the surrogates (truth-likeness, structural homomorphism, etc.) on offer. Throughout I use 'truth' to cover all these.

2 Tangle is used to suggest more than 'panoply', but that is a topic for elsewhere.

science depends on these being up to the job in the way they are needed for that endeavour to be successful. So we need to figure out how to evaluate their in-situ reliability.

The movement for the philosophy of science in practice has long called for broadening our attention beyond the propositional knowledge encoded in general laws to implicit knowledge and practices. Along these lines, Hasok Chang (2012) offers *coherence* of practices as key to scientific successes. This is the kind of work I think we need. But I don't want the message to be lost in a focus on *practices* – where what is underlined is the fact that these are *activities*, things people *do*. Models, measures, data sets and the like are real *products* of science, like the consumer price index (CPI) as a measure of inflation: Is it, as Reiss (2016) questions, reliable as a way to adjust U.S. veteran's benefits to keep them stable year to year?

There has also been a great spate of work on models. There the emphasis has been close to what I urge here, not primarily on the *truth* of models but rather on how *reliable* a scientific model is: Can it do what we want it to? Later I shall also discuss relevant work on measures.

I should note that concern for reliability does not imply no concern for truth. It may be that what we want from a claim is that it be literally true. Or from a model that it accurately represent, say, all the significant causes of some specified phenomenon. We need to assess then, 'How likely is it that the claim or the model can deliver on this demand?' But an almost exclusive focus on truth distracts us from what goes into the hard job of backing up the reliability of all these other kinds of scientific products. Ironically, without good reason in support of the reliability of a host of these other scientific products, we will be on shaky ground in assuming the truth of a scientific claim. An easy place to see this is with the need for good concept validation and reliable measures (see 3. and 5. below for examples). If we don't have these to back up our evidence claims, we do not have firm grounds to support our hypotheses.

I urge a shift in focus from the truth of general principles to the reliability of these different kinds of scientific products for several reasons:

1. It is notoriously difficult to nail down what general claim it is that is supposed to be true
2. A great many general scientific principles are not truth apt as they come – and when rendered so, they are either false, unwarranted or of limited utility
3. The reliability of a huge network of other scientific outputs must be presupposed in order to warrant the acceptability/truth of a general principle
4. Much of what we need to evaluate in science isn't a candidate for truth

5. Reliability invites a crucial question otherwise often overlooked: Reliable for what?

1. That it is often hard to nail down just what claim is being made in scientific contexts is not news. Here I remind you of two familiar sources of the problem: a. 'meaning as use' – the meaning of scientific claims is often (perhaps always) dependent on the network of inferences in which the claim participates, and b. scientific principles are often rendered without quantifiers, as generics, or are supposed to hold only 'ceteris paribus' (CP).

a. From Wittgenstein through Wilfrid Sellars to Robert Brandom, meaning as use has been keenly defended in philosophy. When it comes to science, it is notoriously associated with the views of Kuhn, Feyerabend and Hanson, who underline the problems it raises for sharing and debating "the same" claim across different communities of use; and it sparked a host of work on 'trading zones' and scientific 'pidgin' languages when Peter Galison offered these in remedy to the problems of cross-community debate and consensus.

Think about this in the context of Chang's work (2012). For Chang, claims that play an essential role in the successes of past theories are true. For example, according to mid-eighteenth-century phlogiston theory, substances rich in phlogiston, like metals, are combustible. If a metal is "dephlogisticated," say from burning, phlogiston is released in the form of a flame, and the metal loses its key metallic properties. The phlogiston theory was successful at producing pure air (dephlogisticated air that Joseph Priestley claimed was better for respiration), inflammable air (phlogisticated air) and "calx" (metal that had lost its shine because of lack of phlogiston), as well as metal (made partly from calx). The theory could also be used to make plain water by combining phlogisticated water and dephlogisticated water.

Claims now discarded (like those of phlogiston theory) that were implicated in a variety of successes are true, Chang argues, true because of their pragmatic usefulness. But that does not mean they are only "true, pragmatically speaking". They are true simpliciter, Chang maintains, because the pragmatic theory of truth is the only theory that can be made sense of. The point of this here is that the fact that (if Chang is correct) these claims are true does not mean they can be lifted from the set of source practices in which they were embedded and inserted into current science. They may be true, but if current science does not provide the same resources to make use of them, *they won't be the same claim*. This undercuts the point of trying to establish the truth of claims removed from the tangle of other work that supports them and gives them meaning. And particularly important to note is that, when it comes to confirmation, we should be wary of trusting them when elements in that body of work are flawed – for instance unvalidated concepts are used or insufficiently defended measures.

b. A great many of our useful scientific principles do not come in proper propositional form. So, supposing that only propositions can be true or false, they are not, as they come, candidates for truth or falsehood. *Equations* are a clear example. We may like to think of them as "true" but attempts to render them as propositions generally dramatically diminish their usefulness, or so I argue: See 2. below.

Generics also raise problems. What claim is made in sentences of the form 'X's φ', which are common across the sciences, like 'Neurons transmit signals electrically', 'Democracies do not go to war with democracies', 'People respond to incentives', or 'Limestone reveals its flaws on its surface'. We are inclined to judge some of these as "true", others "false". But so far there do not seem to be semantics available for them that are well suited across these scientific contexts even if we allow different semantics for different contexts.

Laws in the form of 'X – ceteris paribus' – ones with CP clauses attached either implicitly or explicitly – are also notorious troublemakers. Here we do have a semantics on offer that purports to circumvent their troubles. The standard problem is that we use 'ceteris paribus' or some such fudge expression when we think X fails under certain circumstances and holds in others, but we don't know what these circumstances are. The claim then is ill formed, we don't really know what is being maintained.

Michael Strevens (2012) proposes a solution. He considers as an example, 'CP, Printing money causes inflation', which he supposes to be true. But, he notes, 'Under a number of different circumstances, printing money may not affect inflation ([e.g.] the extra currency is hoarded in mattresses rather than spent…)…Because these circumstances are rather diverse, an attempt to specify the economic regularity with any degree of precision will be a daunting undertaking, requiring presumably many clauses, subclauses, parentheses, and footnotes.' [653] But these long-drawn-out lists, he argues, are not the way to do it. Rather we say merely, 'Ceteris paribus, printing money causes inflation.' Strevens then offers a semantics for CP claims that does not assume we are able to specify these great many clauses, subclauses, parentheses and footnotes.

Stevens supposes that for these CP claims, there is an underlying mechanism that, operating undisturbed, generates the regularity described in the claim. The CP clause refers to this mechanism and its undisturbed operation. He maintains that we can succeed in referring to this mechanism without knowing its structure (in his terminology, the structure is 'opaque' to us). So: supposing Strevens is right, given the existence of the *Printing* *money mechanism* and our ability to refer to it with our CP clause, 'CP, Printing money causes inflation' is true.

Strevens also notes that economic mechanisms may not be opaque: 'When economists propose a hypothesis such as *Printing money leads to inflation*, they are able to describe to some extent, if not completely, how the mechanism works; such descriptions of course play an important role in picking out the intended targets of inquiry.' [672] This mention of 'intended targets' points to my second problem with a focus on truth for general principles.

2. Once claims are sufficiently well formulated to be genuine candidates for truth/falsity, they can't do much of the work we put them to in science. Consider CP principles as nailed down by Strevens's semantics. These will cover precious little of the 'intended targets of inquiry'. That's because most mechanisms that underwrite CP principles never work undisturbed. Rather they work in real settings, where much else affects outcomes. Printing money, for instance, always happens along with many other fiscal and monetary policies and a host of other economic activities; so, money may be printed and yet inflation not ensue. Yet the principle, 'CP, Printing money causes inflation' can be crucial to understanding and modelling what does happen in these situations. These complex economic situations are the 'intended targets' for the principle.

We don't generally use these kinds of CP principles as claims from which to derive other claims. We use them in tandem with much else to build models and predictions. We don't need them, then, to be propositions that are candidates for truth. Rather, we need a network of other scientific products to underwrite their reliability for the purposes to which we put them.

In Cartwright (2020) I defend a similar claim about many of our favourite equations, including those in physics. Even if I am mistaken about these prized products of physics, we must not fall into thinking that physics constitutes science and thus we need not concern ourselves with the kinds of principles used elsewhere.

3. We tend to focus on evidence as the basic warrant for general principles. But, as argued repeatedly in philosophy of science,[3] there is no fact of the matter about whether fact e is evidence for hypothesis h independent of background assumptions. Consider Maria Elena Di Bucchianico's (2009) story of two warring camps in high temperature superconductivity. One took the explanatory mechanism to be phonons; the other, magnetic modes. In 2002, new methods rigorously showed a "kink" in the dispersion curve

3 Recent examples arguing this include Helen Longino 1990, John Norton 2003, Nancy Cartwright 2013.

of reflected photoelectrons. Both camps agreed that the data were correct. But they had wildly different interpretations of it due to the great number of differing assumptions they were also committed to. Each of the two warring camps claimed this evidence supported their theory and was incompatible with the opposition's.

But background assumptions are only one among the huge network of scientific products that need to be reliable for a general principle to be warranted. This network usually includes much that is implicitly assumed in evaluating a principle. But often items that are new or might be thought to be missing are discussed explicitly. Consider the discussion by Jennifer Skeem and Christopher Lowenkamp (2016) about whether the Post Conviction Risk Assessment [PCRA] algorithm is generally an accurate and racially unbiased predictor of recidivism. Among the vast amount of work reported in this paper, I note two for illustration:

- Worries about one kind of validity (inter-rater) for the procedures for assigning the input information for the algorithm: 'The PCRA has been shown to be reliable and valid. Specifically, officers must complete a training and certification process to administer the PCRA. The certification process has been shown to yield high rates of inter-rater agreement in scoring…'. [17]

- Worries about test bias: '[There is] little evidence of test bias for the PCRA—the instrument strongly predicts arrest for both Black and White offenders and a given score has essentially the same meaning—i.e., same probability of recidivism—across groups.'[2]

Here I am not endorsing that the work presented in that paper and elsewhere supports that the PCRA is reliable for predicting recidivism in some (generally not well-enough specified) populations of offenders. Rather I want to point out two among a host of components that are needed to support its reliability that should raise red flags if missing.

Sharon Crasnow's (Forthcoming) discussion of the V-dem measure of democracy is another case where worries about validity are brought to the fore. Crasnow notes that V-dem is very attentive to test bias. For instance, they use predominantly in-country experts for coding and they explain, 'Multiple experts (usually 5 or more) code each variable'. [16]

So, we may be impressed by a new result – say a wonderfully precise novel prediction born out in a carefully conducted experiment – and take it to confirm a general law. But whether the result is relevant to the law depends on a host of other assumptions being true, other experiments having been well conducted, a host of concepts being true to the world and their measures being sound, well-constructed and carried out well and so forth.

4. Many of the scientific products we need to assess are not "truth apt". Science creates a huge variety of different kinds of outputs that play different roles in different contexts. Each needs to be able to do the job at hand if we are to rely on its use, and much scientific effort is devoted to ensuring this. These various scientific outputs make a motley assortment. Here are just some in no special order and at no special level of description:

- Theories
- Laws
- Local claims – descriptive and predictive
- Bridging principles
- Models
- Methods – innumerably many across the sciences
- A huge variety of theoretical and practical practices
- Concept development and validation
- Measures
- Evaluations
- Devices
- Model organisms
- Statistical analyses and other applications of mathematics (approximations, …)
- Data curation
- Production
- Preservation
- Classification
- Dissemination
- Narratives

…

Truth is the wrong dimension along which to evaluate the bulk of these. Whether truth apt or not, for all of these we want to know: Are they likely to do what we want of them?

5. Reliability immediately invites the question: Reliable for what end? The specification of the end is always important in evaluation, whether it is claims or devices we are evaluating, something that is easy to overlook in judging general principles as true/acceptable. General principles are put to a variety of different uses in a variety of different contexts where different bodies of background elements are in place. They will generally do a good job in some of these but not others, in part because of the issues raised above about meaning-as-use. So, the stark judgment "true/acceptable versus not true/acceptable" will lead us astray much of the time.

The importance of being clear just what purposes are intended has been well rehearsed in the modelling literature. Is the model intended to provide understanding? To provide accurate predictions? Predictions about what? Should it depict the significant causes of some phenomenon of interest? Perhaps it is supposed to isolate a single cause to study its peculiar effect. Or are we going to probe the model to learn about the world? Just how do we plan to probe it and what do we expect to learn?

It is also brought to the fore in philosophy of science work on measures. Consider the representational theory of measurement (RTM) developed by Suppes and Luce and given a simple articulation recently by psychologist Norman Bradburn and me (2011). According to RTM, a good measure needs three components plus a defence that the three are appropriate to each other.

- a *characterization* of the concept to be measured
- a formal *representation* of it (as in a table of indicators or an index)
- procedures for assigning values to the items measured.

Sometimes the arguments that the components mesh properly are in the form of formal theorems – e.g. *representation* theorems, like the von Neuman-Morgenstern theorem that provides a formal representation of the concept "utility". Usually they are informal. The point of these arguments is to show that the representation can do the job of representing the concept characterized and that the procedures are reliable for ascertaining what values that concept takes in measured systems.

Consider the capabilities account of wellbeing, developed, in different ways, by Amartya Sen and Martha Nussbaum (1993). Sen characterizes well-being as constituted by, informally put, the set of lives worth living available to an individual; Nussbaum specifies 10 spheres of human experience that everyone should be above a minimal threshold on. Both stress that the values involved are diverse and cannot generally be ranked in importance or traded against one another. For Sen, there may be no fact of the matter for two individuals as to whose available lives are better; Nussbaum calls improvement in one sphere that is below threshold to advance another sphere, a 'tragic trade-off'.

Both insist that the concept of capabilitarian wellbeing does not lend itself to providing total orderings across individuals or populations. Yet many attempts to measure it do just that, such as the Alkire and Foster Capability Deprivation Measure and the Krishnakumar "improved" Human Development Index that Travis Chamberlain (2020) critiques. Chamberlain argues that, despite their care and sophistication, there is no good argument in sight that the procedures specified for these will find out about what they are supposed to. Those procedures are not warranted as reliable for assessing the capabilitarian wellbeing of individuals or populations we wish to measure.

Beyond these worries about whether the procedures and representation offered will serve a concept as it is characterized, the measurement literature is equally alert to the important issue of whether the concept characterized and its related measure will serve the purpose the concept is intended for. The consumer price index (CPI) as currently measured may serve reasonably well the purpose of estimating the average increase in the price of the designated basket of goods across all the places the goods are available for purchase in the US but, as Julian Reiss (2008) suggests, not at ensuring that veteran's benefits can secure the same standard of living from year to year, because veterans living on benefits often have little access to large suburban outlets where prices are cheaper and whose prices bring down the CPI.

We also make *relative* judgments about reliability with respect to how fit for purpose a measure is. Will a poverty measure that counts numbers below an absolute threshold (say $25,750 for a four-person household, as in the US in 2019) or below a relative one (say household income below 60% of the average, as in the UK now) reveal the amount of suffering as well as a "depth of poverty" measure that weights individuals/households according to how far below threshold they are, those farther down getting more weight? And of course, relative reliability judgments are in no way confined to measures; we make them about all of the various outputs that the sciences produce.

Getting clear just what reliability claim we are trying to evaluate is not always easy though. Often it is an iterative process, honing the jobs we expect a scientific product to do as we refine the body of support that it can do the jobs delineated, and vice versa. Sometimes it is only after the fact that we realize we were focusing on one purpose but implicitly assuming others would be served as well. Consider the Vajont damn disaster, discussed by Pierluigi Barrotta and Eleonora Montuschi (2018), where an entire town in the Dolomites was wiped away by a gigantic wave of water because a massive limestone landslide fell into the reservoir, the dam resisted the impact and the overflowing water flooded the entire valley. Engineers had focused on whether a dam built as planned would stand against a range of onslaughts and also on whether the surrounding stone would support it. It seems that it was implicitly supposed, wrongly, that a yes answer implied that human lives in the area would be safe in the face of those onslaughts.

The engineers also relied on a well-supported general principle to tell them about local rock: From the chief engineer, '…the rocks [of the Veneto region] are generally very good [. . .]. Overall, limestone is honest because it reveals its flaws on its surface' [??] (– note the generic form of this principle!). In-depth geological studies were considered unnecessary because the rocks of the area did not raise visible concern. Tragically, there was evidence available at the time that this principle was not true of the area rocks. There was clear local knowledge of large limestone landslides up the valley. Barrotta and Montuschi fault the engineers both for neglecting this local knowledge and for not doing a geological study. The engineers were not warranted in taking the general principle to be locally reliable without more investigation given that lives were at stake.

This example underlines how important is the focus on the particular.[4] Recall the National Academy of Science's claim, '[T]heories are the goal of science'. I argue we should turn this back to front, to formulate a goal that is at least as important: the demand that what we want from science is one particular success after another after another after another. Theories and general principles are among the tools that help us achieve this. They are key ingredients in doing so efficiently. They are, in this light, means, not ends. But only means as part of a huge network of other scientific work that provides interpretation for them, validates and measures their concepts, turns facts and study-results into evidence for them, warrants those facts and validates the study designs, builds principles that bridge from their abstract concepts to more concrete locally-relevant ones, shows how to combine them with other relevant knowledge, etc., etc.

I do not mean to suggest that these matters are not tended to. Producing, policing and evaluating all the ingredients that ensure reliability at the point of action is business as usual in science and science-based policy, engineering and technology. Doing it right is a matter of good science. Perhaps that is why we philosophers have not been so engaged with it. But then, accepting the 'right' general laws is also a matter of good science and we philosophers have a great deal to say about how that should be done, and why.

IN CONCLUSION

I clearly went beyond what I have defended in saying, just above, that general principles are means not ends. They may well be both, and they may be means to some very general ends – like understanding or representing a "true" law of nature. In which case, following my line of argument here, I would hope to see a good characterization of what that end is and good arguments in each case that the principle is able to achieve it (some of which may already be available in philosophy of science).

4 How particular? Some purposes are local, like Vajont-dam planning, others, more general. We expect many scientific outputs to serve a given purpose across a range of cases, as with employing the CPI year after year to measure average increase in price. In this case, we need to provide arguments for their general reliability for that purpose, and we need to be alert that what holds generally may fail in any particular case.

What I have defended is that general principles are just one of a vast panoply of outputs the sciences produce. All of these need to be reliable in situ for each purpose we put them to. We are not warranted in expecting the purpose to be achieved without warrant that these are all, together, up to the job they are needed for in securing that purpose. I have also presupposed that theory at some level will be part of the means employed for almost every purpose in science. In which case we should expect the theoretical principles employed to work as they are required to in that case. It is not enough that they are "generally acceptable", as illustrated in the Vajont dam disaster.[5]

Durham University
Department of Philosophy
Stockton Road
Durham DH1 3LE
United Kingdom

5 Again, even though you might think this does not hold for physics principles – the consequences drawn from the good ones are always reliable -- this is far from true for what is by far the bulk of science and science that we regularly use to get around in the world.

REFERENCES

Barrotta, Pierluigi, and Eleonora Montuschi.2018. "The dam project: who are the experts?." Science and Democracy: Controversies and conflicts 13 (2018): 17.

Cartwright, Nancy, and Norman Bradburn. 2011. "A theory of measurement." National Academies Press.

Cartwright, Nancy. 2019. *Nature, the artful modeler: Lectures on laws, science, how nature arranges the world and how we can arrange it better*. Vol. 23. Open Court Publishing.

Cartwright, Nancy. (2013). Evidence, Argument and Prediction. In V. Karakostas, and D. Dieks (Eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings 2*.

Chamberlain, Travis. 2020. The Capabilities Approach to Well-being. MS, Philosophy, UCSD.

Chang, Hasok. 2012. Is water H2O?: Evidence, realism and pluralism. Vol. 293. Springer Science & Business Media.

Crasnow, Sharon. Forthcoming. "Measuring Democracy". MS.

Di Bucchianico, Maria Elena. 2009. "Modelling high temperature superconductivity: A philosophical inquiry in theory, experiment and dissent." PhD diss., London School of Economics and Political Science (United Kingdom).

International Network for Natural Sciences. "Types of Scientific Research". http://www.innspub.net/types-of-scientific-research/ .

Longino, H. *Science as Social Knowledge*, Princeton: Princeton University Press.

National Academies. 2008. Institute of Medicine. *Science, evolution and creationism*. National Academies Press.

National Research Council. 2002.*Scientific research in education*. National Academies Press.

Norton, J. D. 2003. "A material theory of induction". *Philosophy of Science*, 70(4), 647–670. https://doi.org/10.1086/378858.

Nussbaum, Martha and Amartya Sen. 1993. eds. *The quality of life*. Oxford University Pres,.

Reiss, Julian. 2016. Error in economics: towards a more evidence–based methodology. Routledge.

Skeem, Jennifer L., and Christopher T. Lowenkamp. 2016. "Risk, race, and recidivism: Predictive bias and disparate impact." *Criminology* 54, no. 4 (2016): 680-712.

Strevens, Michael. 2012. "Ceteris paribus hedges: Causal voodoo that works." *The Journal of philosophy* 109, no. 11 (2012): 652-675.