The tangle of science: Reliability beyond method, rigour, and objectivity. By Cartwright, Nancy, Jeremy Hardie, Eleonora Montuschi, Matthew Soleiman, and Ann C. Thresher. Oxford University Press. 2022. 249 pages.

The Tangle undertakes to change the focus of philosophy of science towards the *reliability* of scientific *products*. Both parts matter. Reliability is the fact that we sometimes can count on science to do what it's meant to. This is different from demarcation of science from pseudoscience, different from whether science discovers truth about hidden reality or just saves phenomena, different from whether it attains mind-independent objectivity or forever remains a social construction. All of these traditional preoccupations are orthogonal to reliability, and accounts of demarcation, realism or objectivity tell us little about how reliability is achieved. The idea of a product is also crucial. In *The Tangle* the term 'product' covers theoretical and empirical claims, but also instruments, techniques, concepts, procedures. Like products of everyday life, they have purposes, expiry dates, and instructions for use. They are redeployable, but they work better in some environments than in others and only when they are properly linked to other products. It makes no sense to speak of products as true, even approximately, only as fit for purpose. Putting the two together, the authors urge:

> The focus should be not on how to warrant the truth of scientific claims but rather on how to *accredit* this *tangle of work* – or how to gain assurance that the body of work that supports the reliability of a scientific product is up to the job (5).

The tangle is the central setpiece of this book. It refers to the combination of interlinked scientific products, whose deep and varied dependence on each other can secure reliability. At first blush, this idea feels so familiar that you wonder how newsworthy it is. Of course, theories often depend on each other and on observations, and on measurements, and on models, and on experiments. This dependence is sometimes deductive, sometimes analogical. Coherence, webs of beliefs, co-evolution of theory and observations are well known characters in philosophy of science. And no one denies the material reality of instruments and institutions which make these webs possible. Why do we need the tangle in addition?

I warmed up to the concept of tangle gradually and this is because the authors have a cunning strategy. They start with a beguiling image – the floating nest of the African Jacana bird, a wonder of agility and security – a physical tangle that ensures the incubation and birth of delicate creatures. It is hard not to fall in love with the nest. The second move is to review the concepts that philosophers typically use to secure the reliability of science. These 'usual suspects' are scientific method, rigour, and objectivity, and they get one chapter each. These three chapters show that each of these concepts is both silent about the tangle while also crucially presupposing it. Now the reader grows concerned – the tangle is there all along but nobody finds it important enough to acknowledge. No one disagrees that background knowledge matters for confirmation, theory building, and explanation, but everybody is happy to blackbox it. You begin to wonder why. The third move is to unpack the concept of a tangle as philosophers would – with distinctions, definitions, and substance. The tangle isn't just the familiar web of beliefs. It has a structure and normative constraints. These constraints support reliability when we manage to get it. The finale is the illustrations, namely detailed reinterpretations of well-known episodes in history and contemporary science using the language of the tangle. Now there is no escaping. You've intuited the tangle, you've seen how it is occluded, you've unveiled it, and you've pointed at its instances. Case closed. This reader was convinced… mostly. But let me retrace the steps in more detail.

The idea of a scientific method is usually the first stop for anyone, philosopher or scientist, who wants to capture the specialness of science as a path to knowledge. Twentieth-century discussions have centred almost uniquely around it, starting from the inductive logic of the Vienna Circle, to Popper's falsificationism, to modern Bayesian confirmation theory. The conceit of all these projects, as Hilary Putnam noted, is to identify features that make scientific inference legitimate *independent* of the subject matter. To the extent that any theory of scientific method assumes such independence, namely the possibility of formulating context-free rules or procedures that, if followed, result into science, all these theories have failed. Beyond comforting generalities, they said very little about what makes science good when it is. To be action-guiding, a rule needs to be translated into the language of the particular research problem. Otherwise it remains vague and uninformative. Reject falsified hypotheses! Which ones are those exactly? These queries are only answerable using information about the area of scientific investigation in question (John Norton's material theory of induction makes this point too). Such incompleteness does not bother all defenders of scientific method and the authors' main interlocutor here is Elliott Sober. Sober accepts the necessity of appealing to context but thinks that incomplete methods are still methods. To Cartwright et al, this won't do. As soon as you have admitted that any rule (or any epistemic virtue such as simplicity) has to be precisified in context, you have helped yourself to a tangle all the while minimising its role.

Another worry about scientific method is that the very idea of it harms science. Actionable methods are many, not one, and which rule applies and in what formulation depends on the situation. Even if a rule is widely applicable, to reify it is as **the** method is unearned privilege and the very act of granting it such privilege changes science – it directs resources to further and further application of this particular method, even if better ones are available. Randomised controlled trials, as we shall see shortly, have that effect. Note how different this critique is from Feyerabend's famous dismissal of method. He gave us famous examples of scientists being opportunistic and getting away with it, while Cartwright et al point to the fact that much of the justificatory action in science lies beneath the method.

Finally, methods of inference, even if flexible and plural, are inevitably biased towards propositional products such as theories and explanations. This narrowness hurts philosophy of science. Our field has long obsessed about underdetermination and the pessimistic meta-induction as threats to realism but ignored the full range of products that science produces. On this list are:

> models, measurement definitions, procedures, and instruments; concept development and validation; data collection, analysis, and curation; experimental and non-experimental studies; statistical techniques; methods of approximation; case study; narratives; etc.; etc.; etc.(3)

Any of these can be critical to the reliability of scientific products, but many of them are better evaluated by concepts such as fitness for purpose, dependability, or reliability, rather than truth or level of confirmation. Reliability is better because most science is not even a candidate for truth. To be reliable is to be reliable *for something*, so purposes and context become central rather than being considerations on the side.

So we should speak of methods rather than method, and when we do, we should not treat them as the key to the success of science, but just one ingredient.

The next target is a less familiar concept to philosophers of science, namely rigour. It is nevertheless an entirely appropriate target because rigour is so much part of the *lore* in public discussions. Everywhere you look, organisations tasked with evidenced-based policy tout rigorous design, rigorous testing, rigorous procedures. Apart from having the frankly gendered connotations of hardness, seriousness, and inflexibility, what exactly does rigour mean? In the world of reports issued by research councils and think tanks, rigour evokes quantitative methods, statistical analysis, gold standards. Cartwright et al want to do better. They start with a good old conceptual analysis of 'rigour' as it features in these discussions. Then they show that a prime example of rigorous methods in science, namely the randomised control trial, only accomplishes something far too narrow for what evidence-based practise requires (see shortly). If that is what rigour is, then it is pitifully insufficient. Just like method, rigour is a filler word that promises more than it delivers.

The heart of the authors' conceptual analysis is a distinction between formal rigour and substantive rigour. The content of the former is captured in this extended quote:

> a. Nontrivial—not just looking to see.
> b. Auditable—you can ask for an account of how in a particular case the process was followed and get a comprehensive answer.
> c. Rule-based—you can write out what the process is that you have to follow.
> d. Precise, as to
>> i. The rules—no 'do what seems reasonable at this point'.
>> ii. The inputs—no 'add a pinch of sugar'.
>> iii. The output—no 'yep, it works'.
>
> e. Integrable—the rules tell you how to combine the inputs and/or outputs.
> f. There is a valid argument (i.e. the conclusion follows from the premises) showing that if all input assumptions and data are correct and you followed the rules correctly, then the output should be as it is supposed to be.(57)

All these features of formal rigour are undoubtedly attractive. They ensure certainty of the output if the rules are followed. The problem is that few rules are like that. Vagueness and ambiguity abound and, in any case, such purely procedural rigour does not ensure the soundness of inputs. This is the point of the credibility revolution in econometrics – to ensure or make likely that the variables that feature in the equations are actually causally connected. More than formal rigour is required, and the authors introduce a stronger sense: *rigour with certifiable inputs*. This captures the idea that a research method has to produce evidence that substantially bears on the problem at hand. To do so it has to solve ‘a problem that threatens the reliability of a product you are going to put to use’ (60).

This section of the book (2.3-2.6) uses Cartwright’s earlier work on the costs and benefits of RCTs, but now framing the story as a pursuit and a failure to secure rigour with certifiable inputs. The formal properties of an RCT ensure an unbiased estimate of the average treatment effect of an intervention on an outcome variable in the population studied. That’s a valuable achievement, but a limited one if you consider how much more evidence is needed to warrant such an intervention at the level of policy. Once you think of rigour in the stronger sense you can see the problem with calling randomised controlled trials (RCTs) rigorous. Two shortcomings stand out.

First, the prized property of orthogonality (when the treatment effect is probabilistically independent of confounders) isn’t actually secured by the two traditional techniques, namely random assignment and blinding. In reality, local factors, for example differential dropout rates, can undermine orthogonality. Design of the trial alone is never enough to ensure orthogonality in the real world. You also need local causal knowledge. It tells you what to correct for as the trial goes on. And, here's the crucial rub, this local knowledge is not itself derived from rigorous methods.

Second, nothing in the RCT by itself ensures exportability to new populations. True, sometimes the trials happen on giant populations and the effect sizes are strong (think of the COVID vaccine trials). But even then, you need extra knowledge to secure optimism that the effects will hold outside the trial sample. This is why Cartwright (this chapter sometimes veers into ‘I’ rather than ‘we’) resists the term ‘external validity’ – there’s no prospect of genuine validity when it comes to out-of-sample predictions.

This might sound like a pessimistic message, but it’s not. Experienced practitioners often have the requisite knowledge to substantiate orthogonality and exportability. Rather, the message is that RCTs cannot stand on their own and hence their rigour is but a thin veneer. RCTs’ reputation free-rides on the tangle, unfairly appropriating all the credit for successful interventions. A striking figure (2.1, p.78) lays out a “process tracing theory of change”, namely a model that describes in detail the steps by which an intervention can produce an outcome in a real situation. In this case, it’s a mobile phone app designed to improve nutrition and health outcomes in Indonesia. Such a model is an impressive sight – a scientific achievement in its own right – for it undertakes to translate and aggregate all the relevant evidence – rigorous or not – into a stepwise process that takes care at each juncture that all the support factors are in place, that the potential derailers can be controlled, that due safeguards are in operation. Can the right people use the app to input the data? Will they trust its recommendations? If the app says to go see a health professional, is there one close by? And so on. Each piece of causal knowledge in this chain is uncertain and only a tendency anyway, but together they create a tangle that secures reliability.

Another compelling metaphor completes the picture:

> Rigorous results are like the small rigid twigs in a bird's nest. They do some of the job. But a heap of twigs will not stay together in the wind. They must be sculpted, skilfully, with a variety of very different scientific outputs if they are to play a role in supporting the reliability of our scientific products. Just as a bird must weave together a hodgepodge of leaves, grass, moss, mud, saliva, feathers and so on to build a secure nest (82).

Chapter 3 tackles objectivity, another trademark feature of science wheeled out to secure its trustworthiness. Unlike the chapters on scientific method and on rigour, this chapter has a complex message and, in my view, too many moving parts. I found it diffuse and hard-going. It features an aside on how to deal with the so-called Ballung concepts of which objectivity is an example (eliminate them, or precisify, or list the ingredients that make it up). This is important material because so many concepts in social and life sciences are of this nature – multivocal, complex, laden with different senses. But this interlude takes this part of the book into a different direction and at times treads on familiar territory. New literature on objectivity is generally sensible, with almost everyone concluding basically the same thing – objectivity takes on different meanings in different historical and cultural contexts and it is hard to eliminate this feature because the concept serves an inherently vague function, protecting science from threats that cannot be specified in advance. This leads to the truism 'context matters' and the chapter is at its strongest when it unpacks this truism. To be objective in a specific setting requires first unpacking the aims of a scientific product and then assessing with the best available knowledge the best methods for achieving these aims. A key case study is the Vajont Dam disaster in which Italian engineers were unusually put on trial for their failure to be objective. For Cartwright et al, the aims of the dam were not specified at the level such a big project needed and, when geological knowledge was mobilised, it was too general and insufficiently tested on local limestone and with no awareness of inductive risk. Page 125 provides a list of questions that, if posed and properly considered, would have made the dam knowledge more objective and might have avoided the disaster. Answers to such lists of questions cannot be provided independent of the case in question – no tangle, no objectivity. The authors' overall intention is to thicken the substance of objectivity. It is not enough to say that objectivity is relative to aims and context, you must also have an account of how to locate and cross-relate these aims. Fair enough, though by this point in the book the point seemed overlaboured.

This brings us to the central chapter 4, which promises to define the tangle. This chapter opens with a section on conventional confirmation theories, such as hypothetico-deductivism or probabilistic theories. The goal of these theories is to work out the conditions under which a hypothesis H is confirmed by a piece of evidence E. These conditions are supposed to be formal and yet capture scientific reasoning at a due level of abstraction. Philosophers debate the precise shape of these conditions and the interpretation of probability when H and E are uncertain. Cartwright et al show just how much hides behind the simple naked sentences such as 'E implies H' or 'E makes H more likely'. An enormous amount of background knowledge goes into even the simple act of presenting facts as evidence E, let alone establishing a relationship between H and E. You need to find the right language to express H and E. Suppose H is that a given programme of education will improve infant nutrition, and E is its apparent success in Tamil Nadu but apparent failure in Bangladesh. It makes all the difference whether you phrase H in terms of educating mothers versus educating decision-makers, since in Bangladesh mothers aren't typically in that position. Is this E sufficiently varied and robust to bear on H? Lengthy is the task of collating diverse pieces of observation from field experiments, ethnographies, interviews, etc. Then you need to ensure the right auxiliary assumptions, because E will never imply H without them. Assigning a probability to H given E or E given H is another mammoth task and so on. None of these obvious facts would be denied by confirmation theorists. Rather, these colleagues do not find them relevant to the task of confirmation theory. This behind-the-scenes work is assumed to happen in some way or another, but it is not within the purview of confirmation theory. Nevertheless, it is confirmation theory that is supposed to represent scientific inference! So, there is an implicit judgement of philosophical priority. It is the formal relation between H and E that matters to philosophy, not the epistemic task that is prior to that relation. For Cartwright et al, traditional confirmation theories aren't wrong, just frustratingly blind to the ground on which they stand, and that ground is the tangle. I share the authors' astonishment at this seeming misallocation of resources in the epistemology of science (although to me this discussion instead belongs in chapter 1 on scientific method).

But the authors' real challenge is to say something more about the tangle except that it is a complex interrelation between a variety of scientific products. They take a while to get to this point, spending, in my view, too much time on examples and related ideas in the literature. The variety of examples, from archaeology to high-energy physics to political science and public health is incredibly refreshing, but also uneven. The lengthy discussion of an index of democracies called V-Dem did not strike a chord with me, since it's not clear what sort of reliability it exemplifies. Such an index is by its nature deeply controversial and the only ratings everybody agrees on is that Sweden should count as one and North Korea shouldn't. Still, the point is to demonstrate that securing the reliability of any epistemic judgement, be it democratic status of a country or success of a nutrition program, needs multiple layers of constraints and multiple mutually supporting ingredients. Just like the pieces of a Meccano building set – another gripping metaphor – the different components of the tangle can be redeployed in multiple contexts and different users will treat them differently depending on their disciplinary background and goals. Any scientific product that lives up to its promise does so because of an overlapping system of mechanisms in which the ingredients work together to pre-empt obstacles to its operation.

The idea seems right but still feels metaphorical – it's hard to separate the tangle from the image of the nest with a variety of sticks and feathers that hold the eggs afloat as they mature. Yet moving beyond the metaphor is necessary if the tangle is to become a tractable concept in the epistemology of science. This is where the authors bring in some normative constraints. The tangle has to be *virtuous*. A virtuous tangle would typically be: 1) rich, in the sense of consisting of many and diverse products, 2) entangled, in that 'the pieces relate to each other and to the product/aim pair' (157), and 3) long-tailed, namely the components feature in other faraway tangles and use theoretical as well as empirical elements. These three constraints go some way to allay concerns that a tangle is merely a coherent web of beliefs, for a web of belief can be narrow, can use only beliefs from a similar source, and can look like a bicycle wheel where all beliefs connect with the centre rather than with each other. Such a tangle would not secure reliability, because it's circular and closed on itself. It lacks secure constraints. A virtuous tangle uses science in a more voracious and systematic way.

Even so, no virtuous tangle can guarantee reliability. This is because a tangle is not just an epistemic concept, but also a metaphysical one. From nature's point of view, the tangle is a collection of all the facts about the world that bring about a reliable operation of a scientific product. Whereas from the scientist's point of view, the tangle is the set of the elements that, given her available knowledge and aims, creates tight constraints that avoid errors. If nature doesn't play ball, then the scientist can only hope that the tangle she creates is secure. So, assuming nature will sometimes play ball – this book offers no solution to the problem of induction – building virtuous tangles is conducive to reliability. When you fail, you can rethink where your tangle applies, you can add more constraints to it using new scientific products, you can hedge your bets.

The final chapters illustrate the tangle with episodes that, while canonical in history and philosophy of science, look differently now that the language of tangle is in place. The authors revisit how the boiling point of water was established, how penicillin was isolated, how democratic peace is explained (not canonical but well-chosen), and finally how gravitational waves were detected. This is wonderful material that reads effortlessly and creates the warm glow of understanding, coupled with conviction that the epistemology of science should never again be allowed to relegate the tangle to an unessential background factor. But let me skip those details and move to what I see as unfinished business in the tangle's story.

Some things about this book are radical and refreshing. It attempts to reorient our whole field. It is unusually collaborative, featuring five authors of diverse career stages and backgrounds, and citing work that's often undeservedly ignored. It is full of funky metaphors and imagery. In other ways, however, *The Tangle* is quite old-fashioned. It retains the faith that the story of science's reliability is a combination of epistemology and metaphysics, rather than a story of intersubjective agreement and social institutions. The nod to social epistemology in the introduction notwithstanding, these authors treat reliability as a matter of a good match between the epistemic activities of scientists and nature. It is also not a localist picture, despite what you might expect given Cartwright's history of dethroning laws and theories in her previous work. It is crucial to these authors that scientific products have generality that enables redeployment and reuse. The tangle is a product that applies in a particular context, but its ingredients must sit on the shelves of science supermarkets, so to speak, and be in demand for customers of all walks of life.

These commitments are defensible but surprisingly optimistic. Today's science faces steep obstacles to the reliability of its products. Some obstacles are external. Our tasks are just too daunting – manage pandemics, contain climate change, etc. – and our politics are too much of a mess to take advantage of good science. But other obstacles are from within. Our scientists operate in unhealthy institutions, encouraging them to chase status, flashy headlines, and prestige. Scientists are human and they cut corners, sometimes shamefully so. Some fields enjoy more prestige than they deserve, while others are "soft" and hence forced to defend their worth and scramble for resources. There are huge inequalities as to who does science, where, and on what topic.

All these arise as structural problems from the current organisation of universities, research institutes, and think tanks. Some of us look at them and despair, ending up with pessimistic verdicts as to the alleged ability of science to self-correct or build good tangles. The authors of this book have more faith. But they cannot sustain this optimism without attending to those institutions that sometimes manage to prevent scientific failures. These institutions are a crucial part of the tangle.

Take one striking omission. *The Tangle* does not mention, not even in the index, the problem of replication. Nor the growing recognition that disturbingly many published results are not reproducible and may be dodgy in other more worrying ways, likely because modern science incentivises behaviour conducive to this. Isn't that a major challenge to the reliability of science today?

You might think that the focus on tangle justifies side-lining replication. Those who worry about reproducibility imagine that reliability is a property of scientific claims *on their own*, with their own personal effect size and their accompanying p-values. If the locus of reliability is the tangle, the authors might respond, this is misguided. I agree. It's a mistake, to take one example, to invest as much effort as we do now into incredibly demanding exercises of causal and statistical inference at the expense of field methods and case studies with more sensitivity to specific contexts. But I am not convinced that the tangle gives the authors a license to ignore problems with reproducibility. Products in the tangle are supposed to be redeployable, they need to have some validity, and it is hard not to think this validity requires basic virtues that the replication crisis exposed as lacking. Besides, failures of reproducibility are indicators of deeper structural problems in the way we fund, publish, and organise science.

So mine is a verdict of a critical ally. Let's embrace the language of the tangle and tell richer stories, optimistic and otherwise, about when science succeeds, and about what institutions, not just epistemology and metaphysics, enable this.

Anna Alexandrova
University of Cambridge, UK
aa686@cam.ac.uk