



Explaining the UK Metascience Unit's Investments in AI-Driven Research Evaluation

Dr Benjamin Steyn, Co-Head, UK Metascience Unit

Executive Summary

The UK Metascience Unit is funding a cohort of research projects, each focussed on how AI can be used to measure different aspects of research quality, including originality, rigour, and societal impact. The projects include national and international research collaborations forged at a recent sandpit funding event, involving institutions including ARMA, UK Reproducibility Network, and major publishers including Elsevier, Springer Nature, Taylor and Francis, and Cambridge University Press. These new measures aim to assist, not replace, human reviewers to protect the integrity of the scholarly record in the age of AI, while also giving nations, universities, and funders a new vocabulary of quantitative measures to better articulate and guide strategic direction beyond publication counts and citations. This note explains the rationale behind these investments.

Introduction

There are many occasions where the R&D system seeks to evaluate the quality of research outputs. These include:

- Journal and conference peer review
- Budget allocation and strategy-setting (Funder/National/Regional)
- Funding Programme/project evaluation
- National research assessment
- Academic recruitment and career development
- Global University League Tables
- Within the research process itself, including within automated AI for science loops
- In metascience, where scholars are often testing interventions to improve research

As it stands, society's options for evaluating research in these contexts are typically:

- Human expert peer review
- A traditional set of bibliometric indicators, namely, citations and publication counts.

The Metascience Unit is aware that both of these standard approaches are the subject of important scholarly debate.

Human review of outputs has been criticised by scholars for being subjective¹, prone to many kinds of bias², and for pace and cost³. In the age of AI, some journals and conferences are becoming increasingly overloaded by the number of AI-assisted submissions, such that it is becoming more difficult for human experts to review them all within acceptable timeframes.⁴ Even if human review

¹ Park, IU., Peacey, M. & Munafò, M. Modelling the effects of subjective and objective decision making in scientific peer review. *Nature* **506**, 93–96 (2014). <https://doi.org/10.1038/nature12786>

² Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. <https://doi.org/10.1002/asi.22784>

³ Nguyen VM, Haddaway NR, Gutowsky LF, Wilson AD, Gallagher AJ, Donaldson MR, Hammerschlag N, Cooke SJ. How long is too long in contemporary peer review? Perspectives from authors publishing in conservation biology journals. *PLoS One*. 2015 Aug 12;10(8):e0132557. doi: 10.1371/journal.pone.0132557

⁴ Rees, G., & Wilsdon, J. (2026). Could agentic AI topple grant-funding systems? *Nature*, *652*(8112), 1119–1121. <https://doi.org/10.1038/d41586-026-01297-y>

is considered the ‘gold standard’ for reviewing a single output, it is not always practical for contexts where one is trying to compare across vast portfolios of research, for instance, for those wanting to compare universities in league tables.

Among those seeking to quantitatively evaluate research outputs, the standard measure is the citation. While it is implicitly accepted by most system actors that a citation count constitutes some form of quality signal, the vast majority of metascience scholarship over the past decades has been critical about the use of citations as a singular metric of quality. Some of the well-documented⁵ issues with citations include:

- Time lags: delays, sometimes measures in decades, between publication and upticks in citations
- Echo-chamber dynamics: citability relates to the crowdedness of a sub-field, truly novel work may not have an existing community ready to cite it.
- Matthew Effect: success begets success – initially popular papers’ citations compound over time, pushing citations up at a rate that no longer matches their inherent quality v.s. comparable low-cited papers.
- Negative/neutral citations: a citation is not always an endorsement. People cite for many less flattering reasons, e.g. to critique.
- Bad actors: a litany of bad practice is possible to artificially inflate citations, e.g. self-citation rings.
- Goodhart’s Law: “when a measure becomes a target, it ceases to become a good measure” (particularly when a single metric dominates).

Scientometrics

Scientometrics, the academic field aiming to devise better measures to evaluate research, is asking: *How can we improve on this state of affairs?*

To start, consider that it is very difficult to define, let alone measure, the direct impact of a piece of research within and beyond academia. The main output of research activity, knowledge, is not a countable thing. Knowledge can be instantiated through papers, technologies, and policies, but it is indeterminate how much knowledge (if any) there is in any given instance of each. Unlike evaluating a typical public project like a new health or climate investment, research impacts are uncertain, intangible, long-term, interdependent, small, and indirect. Research activity is not limited to one vector of human values e.g. health or climate change. Rather, it may affect them all.

The increasing availability of scientific output data and metadata, combined with advances in artificial intelligence and machine learning abilities in analysing the content of papers themselves, is creating the conditions for a paradigm shift. These developments offer the opportunity to develop new measures of research quality that go beyond traditional citation-based metrics, to more specifically measures of the content of papers themselves, including things like research rigour, novelty, and potential for societal impact of various kinds.

Any new proposed indicator should be met initially with scepticism. A core part of the scientometrics research effort is not only to devise new indicators, but to test and validate them, for instance, against human expert judgement at scale. Once the community are satisfied the new indicator measures the phenomenon it purports to, there are still further questions to answer. Is the indicator easily gamed? Is it biased? Does it work across different disciplines? Does it generate adverse

⁵ Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. HEFCE. <https://doi.org/10.13140/RG.2.1.4929.1363>

incentives which cannot be offset by other indicators? The critical social reflection and discourse around new indicators becomes as crucial as the measures themselves.

The concept of ‘novelty’ has historically been an area of significant focus for scientometrics. This year, the UK Metascience Unit and partners ran a global challenge prize⁶ to find the best automated novelty indicator, as judged against a large purpose-built dataset of human expert judgements. The winning indicator showed a high correlation with human experts, implying that AI can play a role in detecting novel science.

The Metascience Unit and partners want to cultivate a pipeline of new indicators moving through these stages of validation and critical reflection, with the end result hopefully being a responsible and validated ‘basket of indicators’ which allow for a new quantitative vocabulary of research quality. While we aim to develop indicators of commonly discussed virtues of research (novelty, rigour, impact etc.), crucially, we do not stipulate that any particular weighting of these indicators is correct. That is invariably going to be a value judgement that different actors will make for different purposes. For some, it will be an open question whether the properties measured are intrinsically valuable or merely instruments to some further ends. Take, for instance, the concept of ‘interdisciplinarity’. It is something many people say they want more of. Is this sought as an end in itself, which should be numerically rewarded? Or just as a means to some further ends (which can be articulated in other measures)? If interdisciplinarity reliably generates better research, e.g. in terms of novelty, citations, or commercial potential, then arguably the incentive is already baked into a system rewarding those things.

LLMs can also make more holistic judgements of research quality. Mike Thelwall at the University of Sheffield, a Metascience Unit grantee, has shown [high correlations](#) between LLMs and REF Panel scores. However, we expect that a carefully considered basket of task-validated modular indicators should be better performing, and better aligned, than simply asking off-the-shelf LLMs to review papers according to some underspecified criteria. Over time, these modular approaches may be incorporated within broader models, so the distinction may blur.

We should be cautious here. In the age of automated research, making value judgements about what research is good is perhaps the most important task that ought to remain human-driven. With a ‘basket of indicators’ approach, that human value judgement is made explicit in the decisions made about weighting various indicators. Simply feeding a paper to an LLM for review, without detailed instruction, risks abdicating responsibility and accountability. Research virtues are amorphous, multi-faceted, and ambiguous when stated in words. New metrics give us a common vocabulary to ground these nebulous concepts in data.

The Metascience Unit’s Scientometrics Investments

The Metascience Unit’s aim is to fund and bring together a community of researchers and system actors to build this basket of indicators together, while also engaging in critical reflection on the endeavour.

Many of the research teams funded here were created at the recent UKRI Metascience Scientometrics Sandpit Event.⁷ Each project team involves multiple universities, and often involve institutions including ARMA, UK Reproducibility Network, and major publishers including Elsevier, Springer Nature, Taylor and Francis, and Cambridge University Press.

⁶ Steyn, B. (2025). Could a novelty indicator improve science? *Nature*, 642(8068), 543. <https://doi.org/10.1038/d41586-025-01882-7>

⁷ <https://www.ukri.org/opportunity/metascience-sandpit-scientometrics-for-research-assessment/>

The sandpit projects, totalling ~£5m of investment plus in-kind support from collaborators, are as follows:

- A team led by Chenghua Lin at University of Manchester are developing evidence-grounded indicators of scientific rigour.
- Michael Foster at University of Sheffield is leading a team developing indicators of whether a research output has produced new artefacts such as code or primary data.
- Angelo Salatino at the Open University and his team are developing new AI-driven indicators of novelty and disruption.
- Katherine Button at the University of Bath and her team are developing new population-level research environment hygiene indicators, drawing on measures of excess statistical significance to identify where positive results may be over-reported.
- Alma Rahat at Loughborough University and his team are developing a framework for evaluating the reliability of AI-augmented assessment so that humans can decide when, why, and how to trust them.
- Anil Fernando at University of Strathclyde and his team are developing indicators to detect multi-modal research fraud, looking at reported data and image manipulation.
- Enrico Vanino at University of Sheffield and his team are developing indicators which estimate both the likelihood and potential economic value of research projects before their outcomes are fully realised.

Joining these teams are relevant investments and activities from other Metascience funding schemes, including:

- Work continues on the aforementioned Novelty Indicator Challenge, led by Sarah Otner at University of Sussex, and with Challenge winners, led by Jann Weinan at Forschungszentrum Julich GmbH.
- Kathryn Oliver at Kings College London and team are developing indicators of research policy relevance.
- Ramana Nanda at Imperial College London is developing indicators of research commercial potential.
- Mike Thelwall is evaluating the use of LLMs to undertake holistic REF-style assessments of research outputs.
- Other projects funded through our Grants Round 2, which either conceive or use new scientometric indicators.

Alongside these more technical scientometric investments, the Unit are also supporting more socio-technical dialogues: with philosophers of science, global policymakers, and system actors. This includes...

- Ongoing discussions in-house, with other funders, bibliometric data providers and journals about the practical implementation of indicators for different use cases.
- A deep and regular dialogue led by the Research on Research Institute with many national governments and research funders on automation of national research assessment.⁸
- A project led by RAND Europe looking at the political and sociotechnical considerations of the use of indicators in settings including UKRI portfolio evaluation and national research assessment.
- A working group of leading UK Philosophers of Science convened to think about more normative questions facing this agenda, such as: *Who ought to decide the weightings in a basket of metrics? Which aspects of research are intrinsically valuable, and thus should be*

⁸ <https://researchonresearch.org/project/agorra/>

incentivised with metrics, and which are not? And is automated assessment less justified for the arts and humanities than for STEM?

It is hoped that over the coming years, we can move towards better aligned, responsible, robust, and efficient approaches to research evaluation, which arm journals for sifting through poor quality papers, help to guide billions of pounds of research funding, while also rewarding those automated loops which truly accelerate scientific progress. But the opposite may turn out to be true: the evidence may lead us to reject indicators and accept the alternatives – be that human review, citations, or using LLMs in a more holistic fashion. Whatever the answers are, it is a vital terrain to explore.

To be clear, our raison d’être as the Metascience Unit is to build evidence – not to set policy. This is about new exploratory research to test whether AI-driven indicators can be developed, validated, and shown to be reliable, not a commitment to deploy them within UKRI or elsewhere.

This scientometrics programme complements a wider set of Metascience Unit investments seeking to understand how AI is changing science, including:

- The National AI in Research Survey⁹
- Early career fellowships in AI Metascience¹⁰
- Grant funding in AI Metascience¹¹
- Reforming UKRI’s grant processes¹²

⁹ <https://www.rand.org/randeurope/research/projects/2026/ai-in-research-mapping-adoption-attitudes-and-impact-in-the-uk.html>

¹⁰ <https://www.ukri.org/news/international-fellowships-to-explore-ais-impact-on-science/>

¹¹ <https://www.ukri.org/opportunity/metascience-research-grants-round-2/>

¹² <https://www.nature.com/articles/d41586-025-02090-z>