nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-022-01397-5

Public use and public funding of science

Yian Yin[1,2,3], Yuxiao Dong[4,5], Kuansan Wang[4], Dashun Wang[1,2,3,6] ✉ and Benjamin F. Jones[1,2,6,7] ✉

Knowledge of how science is consumed in public domains is essential for understanding the role of science in human society. Here we examine public use and public funding of science by linking tens of millions of scientific publications from all scientific fields to their upstream funding support and downstream public uses across three public domains—government documents, news media and marketplace invention. We find that different public domains draw from various scientific fields in specialized ways, showing diverse patterns of use. Yet, amidst these differences, we find two important forms of alignment. First, we find universal alignment between what the public consumes and what is highly impactful within science. Second, a field's public funding is strikingly aligned with the field's collective public use. Overall, public uses of science present a rich landscape of specialized consumption, yet, collectively, science and society interface with remarkable alignment between scientific use, public use and funding.

Science is often seen to provide substantial impacts beyond the community of scientists themselves—for technological progress, government function, basic human curiosity and more[1–9]. Given the potential benefits, many nations have built institutional architectures to support science through public investment, following the logic of public goods[10–12]. Like a public park, which is funded by the government and can be visited for free, scientific research is substantially funded by governments, with its results placed in the public domain. This institutional design seeks to enable broad use of scientific ideas and avoid under-investment by private actors (for further background on the nature of public goods, see Supplementary Note 3). Yet, in turning to public funding, this approach in part relies on the idea that public investment in science can match the public interest in science.

Although public investment in science is a central feature of the scientific ecosystem[11–13], empirically examining the varied public uses of science and testing whether there is alignment between public funding and public use has remained elusive, mainly owing to the difficulty in collecting systematic data. Moreover, the lack of measurement has invited substantial scepticism. Indeed, many observers view scientific research as a cloistered or 'ivory tower' activity that rarely corresponds to the public interest[14–18]. For example, the 'two communities' and 'two cultures' theories highlight substantial knowledge and interest gaps between scientists and policymakers, disconnecting scientific research from policy insights[19–22] and suggesting little relationship between the quality of research and its public usage[20,23,24]. Meanwhile, scientists may have peculiar interests, with little exposure to real-world problems or incentives to tackle them[7,25]. These potential gaps further animate root concerns over the public funding of science and its proper allocation[26–29]. For example, policymakers have long criticized the National Science Foundation for funding frivolous research and have called for greater transparency around the relevance of science[26,27]. Some prominent academics and commentators, including Nobel-Prize winner Milton Friedman, have taken the position that the government should not fund science, favouring purely private sector research instead[28,29].

In this Article, we advance a measurement framework to study public uses of science, the public funding of science and how public use and public funding relate. Building on prior research that considers the use of science within a given public domain[30–35], here we integrate five large-scale datasets that link scientific publications from all scientific fields to their upstream funding support and downstream public uses across three public domains. Our first dataset (D_1) is scientific publications, using Microsoft Academic Graph (MAG)[36], which is one of the largest bibliometric databases of scientific research in the world (Methods and Supplementary Note 1.1). Our second dataset (D_2) leverages the Microsoft Bing search engine to collect about 6 million government documents available online across all branches of the US government[37]. Using a machine reading technology, we systematically identify academic publications that are referenced in these government documents and match these references to MAG. This pipeline allows us to collect a high-scale dataset on how government documents consume scientific knowledge (Methods and Supplementary Note 1.2). In total, we identify 389,896 unique academic publications cited by 43,014 government documents. We further leverage a secondary policy documents database, Overton, to help validate results obtained from D_2 (Supplementary Note 2.1). Our third dataset (D_3) uses the Altmetric data[31,32] to track academic publications covered by mainstream media reports. Matching these publications to the MAG data yields 724,849 unique papers covered by 2,701 media outlets (Methods and Supplementary Note 1.3). Building on prior work[33–35,38], our fourth dataset (D_4) links all patents granted by the US Patent and Trademark Office (USPTO) to the academic papers they reference, yielding 4,276,940 papers cited by 1,932,642 patents (Methods and Supplementary Note 1.4). Our main results focus on papers published between 2005 and 2014, a common period covered by all three datasets, resulting in 128,465, 275,536 and 1,296,922 papers cited in government, news and patent documents, respectively. Finally, we integrate funding records, using the Dimensions[39] dataset (D_5), which includes 5 million projects funded by over 400 funding agencies worldwide and links each funded project with its resulting publications (Methods and Supplementary Note 1.5). The Methods section and Supplementary Notes 1 and 2 further detail the construction of each dataset and additional validations.

[1]Center for Science of Science and Innovation, Northwestern University, Evanston, IL, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA. [3]McCormick School of Engineering, Northwestern University, Evanston, IL, USA. [4]Microsoft Research, Redmond, WA, USA. [5]Department of Computer Science, Tsinghua University, Beijing, China. [6]Kellogg School of Management, Northwestern University, Evanston, IL, USA. [7]National Bureau of Economic Research, Cambridge, MA, USA. ✉e-mail: dashun.wang@northwestern.edu; bjones@kellogg.northwestern.edu

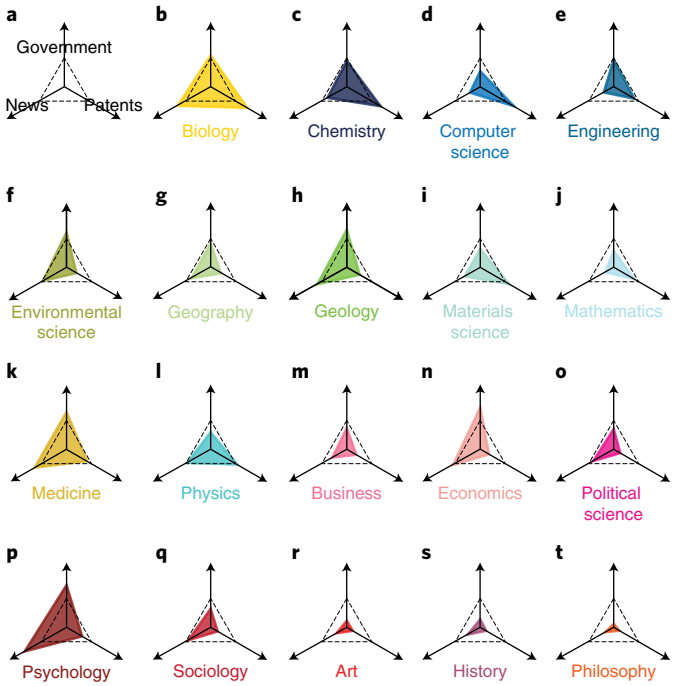

Fig. 1 | Diversity in public use. a–t, Different scientific fields experience distinct and typically specialized public uses. The usage metric RCI for the three public domains are presented for each field (**b–t**). The dashed triangles represent a null model where each paper has the same chance to be used (**a**). The colour scheme highlights four high-level areas of research—the physical sciences, life sciences, social sciences, and ecology and earth sciences—following the four major clusters of science detected by ref. [62] and suggesting commonalities in patterns of public use within these four areas.

Results

Diversity in public use. Our first analyses measure the usage of scientific research in the three public domains. To conduct this analysis, we first leverage the MAG's classification of papers across 19 top-level fields. To account for cross-field differences in publication volume, we define a relative consumption index (RCI). For a given public domain (d) and scientific field (f), RCI measures the fraction of papers in the field consumed by that public domain, normalized by the same fraction calculated on all fields for that domain. That is,

$$\mathrm{RCI}_d^f = \frac{\#\text{ papers in field } f \text{ consumed by domain } d / \#\text{ papers in } f}{\text{Total }\#\text{ papers consumed by domain } d / \text{Total }\#\text{ papers}}.$$

We find that the public uses of science are diverse, with many fields showing substantially specialized usage in public domains (Fig. 1). Computer science, materials science, mathematics and engineering (Fig. 1d,i–j) present substantially larger RCI values for patents than for government or news. By contrast, environmental science and geology (Fig. 1f,h) contribute relatively strongly in government and media documents compared with patents. Finally, physics, chemistry, medicine and biology present a broader range of use (Fig. 1b,c,k,l). Among all fields, biology is the only one over-represented across all three channels, demonstrating a uniquely general relevance to these broad domains beyond science.

Social sciences, by contrast, exhibit a visibly different pattern of public use. The social sciences are strongly consumed in government and media domains while showing systematically low usage in patents (Fig. 1m–q). Economics sees especially strong government use, while psychology, sociology and political science see relatively strong media use. Arts and humanities (philosophy, art and history; Fig. 1r–t) are relatively under-represented in all three domains.

Specialization in public use further appears at subdomain levels (Supplementary Fig. 6). For government, different agencies consume very different scientific research. For example, the US Department of Treasury draws especially on economics and business research, the US Department of Energy draws especially on geology and engineering and the US Department of Defense draws unusually on history. Different patenting fields further exhibit highly specialized relationships with specific scientific fields. By contrast, in media, while *The Washington Post* draws unusually heavily on political science research, mainstream media sources in general are more consistent in the fields they report, with especially strong and widespread interest in medicine and psychology.

The specialization in public use is further accompanied by substantial differences in time lags in the use of science by the different public domains. Whereas the news media places a particular focus on very recent work, the government and inventive domains have wider reach into prior discovery (Supplementary Note 4.5). For example, in the news media, 63% of citations to scientific articles cover research papers published within the year. By contrast, government documents and patent inventions draw more widely over past work, with a median citation lag of 10 years between scientific publication and use (Supplementary Fig. 7). Importantly, while the public domains differ considerably in time lags, we find that the RCI comparisons are extremely similar when considering either the recent decade of scientific publications (Fig. 1) or the stock of scientific publications over a substantially longer history (Supplementary Fig. 8), indicating that the results in Fig. 1 are robust controlling for time lags.

Overall, these results highlight a large set of specialized relationships between specific domains of public use and specific fields of scientific research. From a public goods perspective, if we think of scientific fields as akin to a series of national parks, we see that each park is embedded in particular communities of public use. Collectively, these parks spread across diverse regions of knowledge and are accessed by diverse segments of the public. A few fields, especially biology, receive visitors at relatively intense rates from a broad range of public domains—a 'Yellowstone Park' of science.

Scientific impact and public use. Our second set of results examine whether the public domains tend to consume ideas that scientists themselves consider impactful. Longstanding arguments suggest that the public is not well equipped to evaluate science and may draw on poorly established scientific ideas, which would undermine the public good benefits of science[20,23,24]. Continuing the national parks metaphor, scientists may be primarily focused in a hard-to-reach backcountry, whereas the typical visitor may not have the tools to access this terrain nor gravitate to the same areas the scientists themselves consider attractive. To further examine public use, we therefore consider, at the article level, the alignment between public use and scientific use. Specifically, we calculate the probability of being a hit paper within science, defined as those papers in the top 1% of citations within the same field and year, and examine the relationship to usage in the public domains (Methods and Fig. 2b). We find that papers referenced in public domains have a remarkably high likelihood of being hit papers within science. Papers cited by government documents, news or patents exhibit hit rates of 14.1%, 18.0% and 9.1%, respectively, all large multiples of the baseline rate of 1%. Further, papers referenced in the intersection of different domains tend to be exceptionally impactful in science. For papers referenced in two public domains, approximately half are hit papers. Papers referenced by both government documents and news media have a hit rate of 45.1%. The results are broadly similar if we examine the intersection between government documents and patents

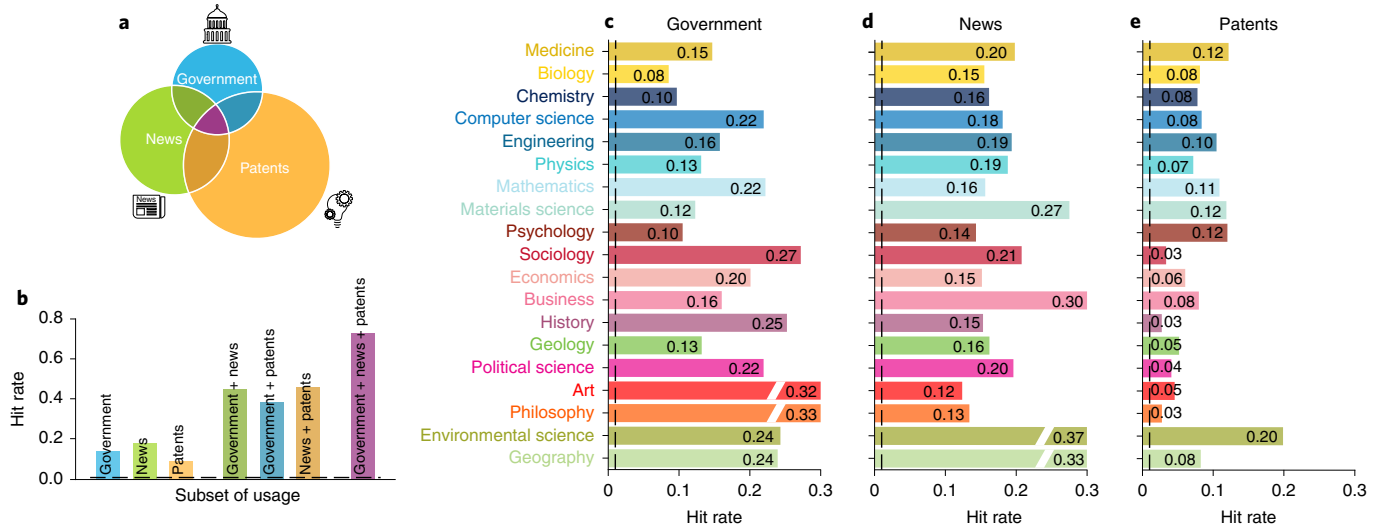

Fig. 2 | Public use and scientific use. The public tends to consume exceptionally high-impact science from all fields and in all three public domains, indicating alignment between public use and scientific use. **a**, Usage by domain for papers published from 2005 to 2014. The area of each subset is proportional to the square root of the paper count in the corresponding public domain. **b**, Hit rates for papers cited in at least one, two or three public domains. Hit papers are defined as those receiving citation counts, within science, in the top 1% within the field and year. **c**–**e**, Hit rates for each of the 19 fields consumed by government documents (**c**), news media (**d**) and patents (**e**). In all fields, and in all three domains, the consumed papers tend to have hit rates within science many times larger than the baseline rate of 1% (dashed line).

(38.7%) or news and patents (46.1%). A paper consumed in all three domains is a hit paper in science at a staggering 72.8 times the baseline rate. Reversing the exercise, we also see that, as the citation percentile of a paper rises, the probability for public use increases steeply, with extremely sharp increases at the very top of the citation distribution (Supplementary Figs. 9 and 10).

The use of high-impact papers is not only common across different public domains, it also appears universal across research areas. Papers covered by public domains tend to be highly cited in all scientific fields (Fig. 2c–e). These findings remain similar when varying the threshold for hit papers to the top 5% or 10% citations (Supplementary Note 4.4 and Supplementary Figs. 11 and 12). We also repeat our analyses for papers produced by United States-based researchers, arriving at the same conclusions (Supplementary Fig. 13). While government, media and patenting documents may cite science for a variety of reasons and our reference-based measures are proxies for uses of science[22,40,41], we see that the science referenced in public domains is not in conflict with what scientists themselves consider important; rather, impactful papers defined by these communities show substantial overlap. This finding stands in contrast to concerns over knowledge gaps, where the government and media in particular may be poorly positioned to assess high-impact scientific work or distinguish it from low-impact scientific work[20,23,24,31,42]. Considering the findings, one may note that, in each of these public domains, the initial step beyond science involves an intermediary—via the journalist in media, the inventor or other domain expert in patenting, the potential policy expert in government—all of whom may bring specialized capacities to bear in selecting what science they bring forth into their domain. The broader public use—among those who read a news article, use an invented product or experience a policy—will then depend upon these intermediaries, who may help bridge the knowledge gap. Overall, the public use of science, while marked by substantial specialization in use across research areas, presents a striking universality, where diverse public domains all draw on the highest-impact scientific papers within each field.

We further fine-grain the 19 broad research fields of papers into 294 subfields as indexed by MAG, and calculate the RCI score for each subfield in a given public domain. We visualize each field's RCI values, locating each field within a common triangle to compare each field's tendency toward usage in specific public domains (Fig. 3a). Fields in social science as well as arts and humanities are mostly used in media and government, whereas fields in science and engineering spread out widely within the triangle, again highlighting the field-level specialization yet collective diversity in the public uses of science.

Public use and public funding. Together, these results raise a central question: To what degree does the funding input for science relate to the field's public use? The majority of scientific research is supported by public investment, which aims to advance not only science itself but also broader public interest[41]. The National Science Foundation, for example, formally introduced broader impacts as a key criterion for evaluating grant proposals in 1997. Here we focus on US-funded projects and use D_5 to calculate the average funding per paper in a given subfield as a proxy for public investment costs per unit of output.

We find that the public investment per paper differs dramatically across fields, spanning over five orders of magnitude. Yet comparing average funding per paper with RCI in each domain reveals substantial correlations between funding and the use of science across all three public domains, with $R^2 = 0.159$ for government, 0.272 for news and 0.376 for patents (Fig. 3b–d, Methods and Supplementary Table 1). To further test if the uncovered correlation is due to the heterogeneity in field size or parent field, we add the number of papers in the subfield as well as parent field fixed effects (for the 19 higher-level fields) into the regression, finding the strong correlation with RCI persists ($P < 0.001$ in all three cases). Notably, across the three domains, the representation of subfields in government documents has the lowest predictive power for funding, suggesting that public investments in science better reflect the overall public interest captured by media or patents. We further include funding from non-governmental sources or focus on papers by US researchers only, finding our conclusions remain the same (Supplementary Notes 4.1 and 4.2, Supplementary Figs. 19 and 20 and Supplementary Tables 4–7).

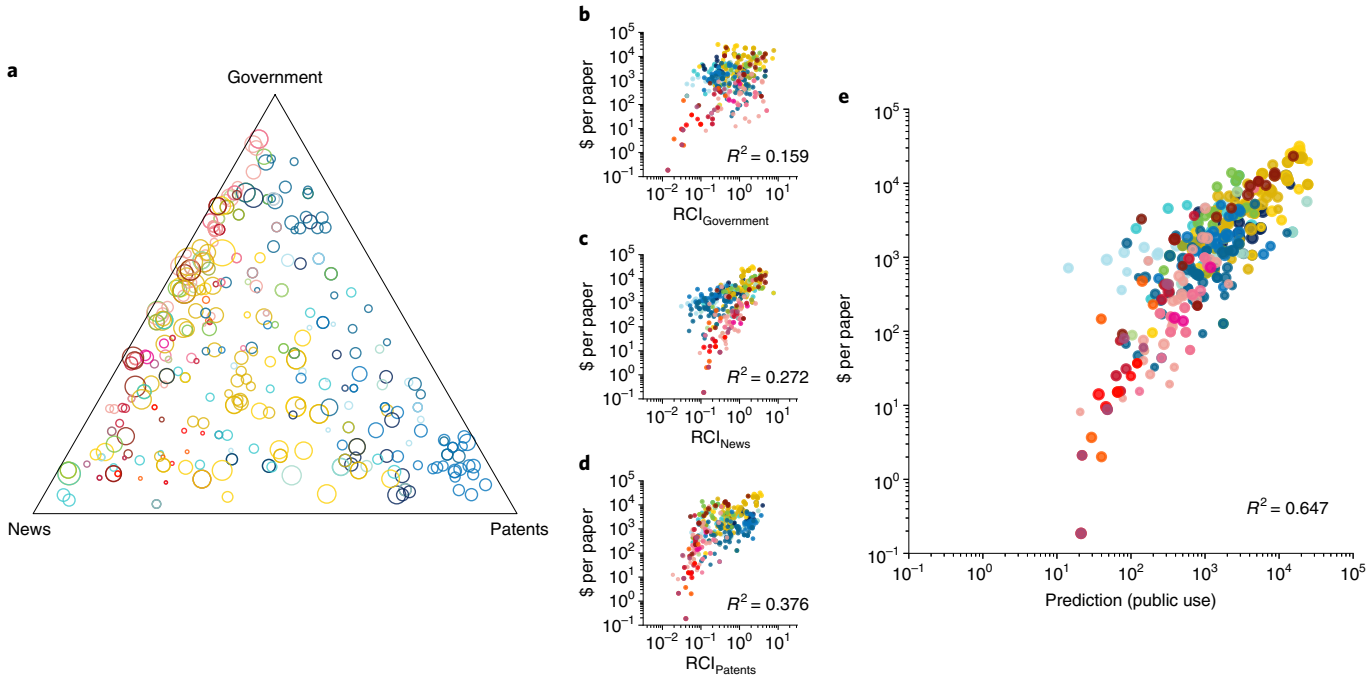

Fig. 3 | Public use and public funding. Amidst enormous diversity in public use across fields and domains, scientific funding for a given field is closely aligned with the totality of its public use. **a**, Ternary plot of RCI for 294 level-1 fields together, with the location of each field indicating its relative usage among the public domains. Circles are colour-coded according to their parent field in Fig. 1, and circle sizes reflect overall usage. **b**–**d**, Average funding (US$) per paper across fields is positively correlated with a field's RCI index in government (**b**), news (**c**) and patenting (**d**). The relationship remains statistically significant when combined with control variables ($P < 0.001$ in ordinary least squares regressions controlling for the number of papers and parent field fixed effects; see Supplementary Table 2 for details). **e**, Collectively, public uses beyond science strongly predict field level funding per paper.

Most strikingly, a simple linear regression model combining the three RCI values together yields a surprisingly high degree of agreement with funding, with an R^2 of 0.647 (Fig. 3e, Methods and Supplementary Table 2), providing at minimum a 72% increase in predictive power compared with using any of the three public domains alone. These results suggest that each public domain provides independent predictive power for understanding the allocation of public investment in science. The uncovered high predictive power of this analysis is especially striking given many complex factors and processes at work in appropriations, budget setting and grant review[43–48]. Although each research field differs substantially in its relative role and contribution in science and beyond, the combination of their impacts beyond science powerfully predicts funding, suggesting that, ultimately, what the public uses, what scientists use and what is funded are remarkably consistent.

Discussion

One source of this alignment could be that science follows the public interest. For example, scientists may prioritize or innately share areas of interest, such as coronavirus disease 2019 (COVID-19), where there is enormous public demand for solutions and where scientific attention has surged[30,49,50]. Another source could be that some scientists or science institutions are especially good at promoting their interests to the public, influencing what the public sees and funds. For example, one may wonder if high-prestige journals, eminent authors or funding for a paper drive attention to specific research. To test this, we further consider fine-grained, paper-level regressions that include journal fixed effects, author fixed effects and paper-level funding indicators. We find that the results are very similar, regardless of these controls (Supplementary Note 4.6, Supplementary Table 8 and Supplementary Figs. 21–23). Indeed, the relative attention to different fields (Fig. 1), the alignment between public use and high-impact science (Fig. 2) and the alignment with public funding (Fig. 3) all appear robust after accounting for journal placement, the scientists who produced the work or the funding status of the specific paper. Thus, while some scientists, journals or funders may have advantages in reaching the public, the forms of alignment we see appear primarily as features of a research area, rather than the specific promotion opportunities from a journal, scientist or funding. More generally, numerous mechanisms, institutional factors and policies may be at work in producing, increasing or reducing use and alignment, and unpacking these mechanisms is an exciting area for future work.

Altogether, the analysis probes quantitatively key features of the public use and funding of science. Measuring the usage of scientific research outside science itself, we uncover enormous diversity and specialization in how different fields of scientific enquiry are linked to different public domains. Yet, despite these differences, the different public domains (and subdomains) universally draw on highly cited papers within science, indicating that public use is strongly aligned with what scientists themselves consider impactful. And, critically, the public usage of scientific fields across the diverse domains provides simple yet powerful predictors for the level of public investment in each field.

Note that, although the three domains each represent an important dimension of the public space, they do not cover all domains that science may impact. Even within each of the three domains we studied, there may be consumption of science through channels that go beyond our datasets. For example, scientists and their ideas can appear through television, in congressional testimony and in private sector consulting. Scientific ideas may also enter industry and government through social networks, through the hiring of scientists, and through influencing managerial practices (Supplementary Fig. 24), which may augment and alter perspectives on the public

use of specific research fields. While there is much still to explore, this paper introduces a quantitative framework to examine public uses of science at the individual paper level, both across all scientific fields and diverse public domains, revealing individually specialized and collectively diverse uses, universality in impact and a remarkable alignment between the funding of science and its public use.

As society's support of science depends on a public goods model[11,13], and as legislators have called for more transparency in the usage and value of scientific funding[51], the framework developed in this paper provides an empirical tool, offering quantitative evidence to inform discussions around public interest features of science. The allocation of science funding involves chains of decisions by individuals and groups with different perspectives and priorities. These considerations range from legislative committees and the goals of individual political representatives, to funding agency leaders, to within-agency mechanisms that often incorporate insights from scientists, interacting in a complex process that must bridge across distinct communities. As such, one might expect a substantial disconnect between what is eventually funded and forms of public interest; metaphorically, funding of public parks in ways weakly related to public use. Yet, despite the massive diversity in the public uses of science and a complex funding process, there is remarkable alignment in the end result. What the public uses and what scientists themselves use are closely consistent. And the funding of science closely tracks quantifiable public use. These results suggest the connections between the ivory tower and the real world appear more aligned than is commonly imagined.

Methods

Microsoft Academic Graph (D_1). The publication and citation data are primarily obtained from Microsoft Academic Graph (MAG)[36,52]. MAG is among the largest open-source citation databases thus far and contains records of 209 million documents. We inter-linked different data tables to obtain the author, affiliation, year, publication venue and field information for each paper. Data pre-processing and summary statistics are further documented in Supplementary Note 1.1.

US government documents (D_2). To quantify references to scientific articles in the government domain, one needs to construct a large-scale dataset of government documents that can be linked to the scientific papers. The task has been difficult in part because government documents are spread across many sources. Furthermore, although a substantial fraction of such documents may cite scientific literature, such citations do not follow a common structure.

Our data collection starts with a list of 6 million URLs under the.gov domain, which is the domain name for government agencies and contains the vast majority of US government entities. We downloaded these pages using an automatic crawler and focused on all PDF files in this set, extracting the references cited in these files using Science Parse[53], an open-source tool for reference string extraction. We then matched this list to the MAG with a search engine-like system using title, journal, author and publication year information. Supplementary Note 1.2 documents technical details of this data pipeline. We also perform additional validation analysis using Overton, an independent dataset of policy documents (Supplementary Note 2.1).

Altmetric dataset (D_3). To study references to scientific publications in the news media, we use a dataset offered by Altmetric[31,32,54]. This dataset records approximately 26.2 million papers with at least one news media or social media mention. We then merge paper information with MAG. A vast majority (22.1 million) of such publications in the Altmetric database have unique digital object identifiers (DOI). We find that 17.2 million (78%) of the DOIs can be matched to records in MAG.

USPTO patent database (D_4). To study references to scientific publications in patents, we build on prior work and use a high-scale mapping from USPTO patents to MAG papers, which includes approximately 31.7 million citation pairs between patents and papers[35,38], from both the front page and full text of the patents. To classify patents into technology classes, we use the Cooperative Patent Classification system, drawn from PatentsView, a data platform based on USPTO bulk data[55] (Supplementary Note 1.4). Combining the two files provides technology class information for 97.5% of patents that reference scientific articles. The small share of missing technology class cases corresponds to patents recently granted, which have not been updated in our data.

Dimensions scientific funding data (D_5). To understand how research funding from various sources is allocated into different scientific fields, we leverage research funding data from Dimensions[39,56], which includes approximately 5 million research projects supported by over 400 funding agencies worldwide. To be consistent with the rest of our analysis, we focus on projects funded during the same 10-year period (2005–2014). A unique opportunity provided by Dimensions is a linkage table between supporting grants and resulting publications, which allows us to categorize the field of each grant according to its resulting publications. Together we link 292,875 funded projects with at least one publication (for detailed descriptions of our linkage procedure, see Supplementary Note 1.5).

Citation percentiles and hit papers. While citations are widely used as a proxy for scientific impact[2,9,57–59], direct comparison of citation counts received by papers across time and field can be problematic without normalization[60]. We therefore calculate citation percentiles for papers within the same publication year and field. Here, following prior studies[33,47,61], we define 'hit papers' (also known as 'home runs') as papers ranking in the top 1% of citations received. We further test robustness of these results by tuning the threshold from 1% to 5% or 10% (Supplementary Note 5.3 and Supplementary Figs. 11 and 12).

Regression models. To understand the association between public use and funding for different scientific fields, we use linear regression models (ordinary least squares). We first note that all three RCI measures are highly skewed (Supplementary Fig. 5a–c), prompting us we take the natural logarithm, ln RCI, in our linear regressions (Supplementary Fig. 5d–f). The same transformation is taken on the average funding per paper. The variables are defined as follows:

Dependent variable. $\ln Y_i$, defined as the natural logarithm of average funding per paper for the level-1 field i.

Predictors of interest. We examine the extent to which different impact measures can predict funding, including $\ln \mathrm{RCI}_{ji}$ for the three public domains. To include all data points in the regression, for the rare cases when an impact measure is 0, we add 1 to avoid zeros in the logarithm. We further include the natural logarithm of the number of papers published in the 10-year period, $\ln p_i$, as a control variable.

Fixed effects. To control for the possibility that fields under different broad categories may have specific funding and public use norms, we introduce F_{fi}, fixed effect terms for each level-0 field. Specifically, $F_{fi} = 1$ if the level-1 field i is a child field of the level-0 field f according to MAG's classification structure. Note that some level-1 fields belong to two level-0 fields simultaneously (for example, mathematical physics is the child field of both mathematics and physics).

We start with bivariate regressions examining the relationship between each RCI (that is, for government, media or patenting) and average funding (Fig. 3b–d, Supplementary Table 1 and Models 1–3). That is,

$$\ln Y_i = \beta_j \ln \mathrm{RCI}_{ji} + \varepsilon_i.$$

In multivariate regressions, we further include controls for heterogeneity in field size or parent field fixed effects (Supplementary Table 2 and Models 4–6).

We further investigate the joint predictive power of the three RCIs (Fig. 3e, Supplementary Table 2 and Model 7):

$$\ln Y_i = \sum_j \beta_j \ln \mathrm{RCI}_{ji} + \varepsilon_i$$

which shows that each measure contributes independently and substantially to explaining the variation in funding.

Finally, we add further control variables into Model 8 (Supplementary Table 2 and Model 8):

$$\ln Y_i = \sum_j \beta_j \ln \mathrm{RCI}_{ji} + \beta_p \ln p_i + \sum_f \beta_f F_{fi} + \varepsilon_i$$

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

MAG raw data are publicly available at https://docs.microsoft.com/en-us/academic-services/graph/. MAG-USPTO linkage data are publicly available at https://doi.org/10.5281/zenodo.3575146. Those who are interested in raw data of Altmetric and Dimensions should contact Digital Science directly. Those who are interested in raw data of Overton should contact Open Policy Ltd directly. The de-identified data necessary to reproduce main plots and statistical analyses are freely available at https://kellogg-cssi.github.io/science_public.

Code availability

Government document data are collected with web crawler programs (customized bash code and Science Parse v1). Raw datasets are further linked using customized code in Python 3 and Elasticsearch 7.0. Data are analysed with customized code in Python 3 and Stata 14.0 using standard software packages within these programs. The code necessary to reproduce main plots and statistical analyses is freely available at https://kellogg-cssi.github.io/science_public.

Received: 21 April 2021; Accepted: 19 May 2022;
Published online: 07 July 2022

References

1. Disraeli, B. *Inaugural Address Delivered to the University of Glasgow Nov. 19, 1873* (Longmans, Green, and Co., 1873).
2. Wang, D. & Barabási, A.-L. *The Science of Science* (Cambridge Univ. Press, 2021).
3. Merton, R. K. *The Sociology of Science: Theoretical and Empirical Investigations* (Univ. of Chicago Press, 1973).
4. Gibbons, M. *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies* (Sage, 1994).
5. Mokyr, J. *The Gifts of Athena: Historical Origins of the Knowledge Economy* (Princeton Univ. Press, 2002).
6. Etzkowitz, H. & Leydesdorff, L. The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Res. Policy* **29**, 109–123 (2000).
7. Committee on Prospering in the Global Economy of the 21st Century, National Academy of Sciences & National Academy of Engineering Institute of Medicine. *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*. (National Academies Press, 2014).
8. Hjort, J., Moreira, D., Rao, G. & Santini, J. F. *How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities*. Report No. 0898-2937 (National Bureau of Economic Research, 2019).
9. Fortunato, S. et al. Science of science. *Science* **359**, eaao0185 (2018).
10. Jefferson, T. No patent on ideas. Letter to Isaac McPherson (1813).
11. Arrow, K. in *The Rate and Direction of Inventive Activity: Economic and Social Factors* (National Bureau of Economic Research) 609–626 (Princeton Univ. Press, 1962).
12. Stiglitz, J. E. Knowledge as a global public good. *Global Public Goods* **1**, 308–326 (1999).
13. Stephan, P. E. The economics of science. *J. Econ. Lit.* **34**, 1199–1235 (1996).
14. Jewkes, J. *The Sources of Invention* (Springer, 1969).
15. Gibbons, M. & Johnston, R. The roles of science in technological innovation. *Res. Policy* **3**, 220–242 (1974).
16. Landau, R., Rosenberg, N. & National Academy of Engineering. *The Positive Sum Strategy: Harnessing Technology for Economic Growth* (National Academies Press, 1986).
17. Mansfield, E. Academic research and industrial innovation. *Res Policy* **20**, 1–12 (1991).
18. Klevorick, A. K., Levin, R. C., Nelson, R. R. & Winter, S. G. On the sources and significance of interindustry differences in technological opportunities. *Res Policy* **24**, 185–205 (1995).
19. Caplan, N. The two-communities theory and knowledge utilization. *Am. Behav. Sci.* **22**, 459–470 (1979).
20. Dunn, W. N. The two-communities metaphor and models of knowledge use: an exploratory case survey. *Knowledge* **1**, 515–536 (1980).
21. National Research Council. *Knowledge and Policy: The Uncertain Connection* (The National Academies Press, 1978).
22. National Research Council. *Using Science as Evidence in Public Policy* (National Academies Press, 2012).
23. Landry, R., Lamari, M. & Amara, N. The extent and determinants of the utilization of university research in government agencies. *Public Adm. Rev.* **63**, 192–205 (2003).
24. Snow, C. P. *Science and Government* (Harvard University Press, 2013).
25. Langrish, J., Gibbons, M., Evans, W. G. & Jevons, F. R. *Wealth from Knowledge: Studies of Innovation in Industry* (Springer, 1972).
26. Hatfield, E. Proxmire's golden fleece award. *Relationship Research News* **4**, 5–9 (2006).
27. Coburn, T. *The National Science Foundation: Under the Microscope*. (Senator Tom Coburn, 2011).
28. Ridley, M. *The Evolution of Everything: How New Ideas Emerge*. (HarperCollins, 2015).
29. Kealey, T. The case against public science. *Cato Unbound* **5** https://www.cato-unbound.org/2013/08/05/terence-kealey/case-against-public-science/ (2013).
30. Yin, Y., Gao, J., Jones, B. F. & Wang, D. Coevolution of policy and science during the pandemic. *Science* **371**, 128–130 (2021).
31. Thelwall, M., Haustein, S., Larivière, V. & Sugimoto, C. R. Do altmetrics work? Twitter and ten other social web services. *PLoS ONE* **8**, e64841 (2013).
32. Costas, R., Zahedi, Z. & Wouters, P. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *J. Assoc. Inf. Sci. Technol.* **66**, 2003–2019 (2015).
33. Ahmadpoor, M. & Jones, B. F. The dual frontier: patented inventions and prior scientific advance. *Science* **357**, 583–587 (2017).
34. Fleming, L., Greene, H., Li, G., Marx, M. & Yao, D. Government-funded research increasingly fuels innovation. *Science* **364**, 1139–1141 (2019).
35. Marx, M. & Fuegi, A. Reliance on science: worldwide front-page patent citations to scientific articles. *Strategic Management Journal* **41**, 1572–1594 (2020).
36. Wang, K. et al. Microsoft Academic Graph: when experts are not enough. *Quant. Sci. Stud.* **1**, 396–413 (2020).
37. Kosack, S. et al. Functional structures of US state governments. *Proc. Natl Acad. Sci. USA* **115**, 11748–11753 (2018).
38. Marx, M. & Fuegi, A. *Reliance on Science by Inventors: Hybrid Extraction of In-Text Patent-to-Article Citations* (National Bureau of Economic Research, 2020).
39. Herzog, C., Hook, D. & Konkiel, S. Dimensions: bringing down barriers between scientometricians and data. *Quant. Sci. Stud.* **1**, 387–395 (2020).
40. Weiss, C. H. The many meanings of research utilization. *Public Adm. Rev.* **39**, 426–431 (1979).
41. Bornmann, L. What is societal impact of research and how can it be assessed? A literature survey. *J. Am. Soc. Inf. Sci. Technol.* **64**, 217–233 (2013).
42. Selvaraj, S., Borkar, D. S. & Prasad, V. Media coverage of medical journals: do the best articles make the news?. *PLoS ONE* **9**, e85355 (2014).
43. Boudreau, K. J., Guinan, E. C., Lakhani, K. R. & Riedl, C. Looking across and looking beyond the knowledge frontier: intellectual distance, novelty, and resource allocation in science. *Manag. Sci.* **62**, 2765–2783 (2016).
44. Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* **534**, 684–687 (2016).
45. Ginther, D. K. et al. Race, ethnicity, and NIH research awards. *Science* **333**, 1015–1019 (2011).
46. Ma, A., Mondragón, R. J. & Latora, V. Anatomy of funded research in science. *Proc. Natl Acad. Sci. USA* **112**, 14760–14765 (2015).
47. Wang, Y., Jones, B. F. & Wang, D. Early-career setback and future career impact. *Nat. Commun.* **10**, 1–10 (2019).
48. Yin, Y., Wang, Y., Evans, J. A. & Wang, D. Quantifying the dynamics of failure across science, startups and security. *Nature* **575**, 190–194 (2019).
49. Hill, R., Yin, Y., Stein, C., Wang, D. & Jones, B. F. Adaptability and the pivot penalty in science. *Available at SSRN 3886142* (2021).
50. Else, H. How a torrent of COVID science changed research publishing–in seven charts. *Nature* **588**, 553–554 (2020).
51. Hearing: The science of science and innovation policy. *Committee on Science and Technology* (2010).
52. Wang, K. et al. A review of Microsoft Academic Services for science of science studies. *Front. Big Data* **2**, 45 (2019).
53. AI2. *Science Parse* https://github.com/allenai/science-parse (2019).
54. Adie, E. & Roe, W. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learn. Publ.* **26**, 11–17 (2013).
55. *PatentsView* http://www.patentsview.org (2019).
56. Hook, D. W., Porter, S. J. & Herzog, C. Dimensions: building context for search and evaluation. *Front. Res. Metrics Anal.* **3**, 23 (2018).
57. Garfield, E. & Merton, R. K. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Vol. 8 (Wiley New York, 1979).
58. Price, D. J. D. S. Networks of scientific papers: the pattern of bibliographic references indicates the nature of the scientific research front. *Science* **149**, 510–515 (1965).
59. Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342**, 127–132 (2013).
60. Radicchi, F., Fortunato, S. & Castellano, C. Universality of citation distributions: toward an objective measure of scientific impact. *Proc. Natl Acad. Sci. USA* **105**, 17268–17272 (2008).
61. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
62. Rosvall, M. & Bergstrom, C. T. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS ONE* **6**, e18209 (2011).

Acknowledgements

We thank I. Shen, D. Eide and all members of Microsoft Academic group for their invaluable help and J. Trefethen at Open Philanthropy for his comments. This work uses data sourced from Altmetric.com and Dimensions.ai through researcher access plans. D.W. is supported by the Air Force Office of Scientific Research under award numbers FA9550-17-1-0089 and FA9550-19-1-0354, National Science Foundation grant SBE 1829344, the Alfred P. Sloan Foundation G-2019-12485 and Peter G. Peterson Foundation 21048. Y.Y. is supported in part through the computational resources and staff contributions provided for the Quest high-performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost,

the Office for Research and Northwestern University Information Technology. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

D.W., B.F.J. and K.W. conceived the project and designed the experiments; Y.Y. and Y.D. collected data; Y.Y. performed empirical analyses with help from D.W., B.F.J., Y.D. and K.W.; all authors discussed and interpreted results; Y.Y., B.F.J and D.W. wrote the manuscript; all authors edited the manuscript.

Competing interests

K.W. and Y.D. were employees of Microsoft Corporation when the study was conducted. Y.Y., B.F.J. and D.W. declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01397-5.

Correspondence and requests for materials should be addressed to Dashun Wang or Benjamin F. Jones.

Peer review information *Nature Human Behaviour* thanks Carolin Haeussler, Paula E. Stephan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Benjamin F. Jones and Dashun Wang

Last updated by author(s): May 19, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Government document data is collected with web crawler programs (customized bash code and Science Parse v1). Raw datasets are further linked using customized code in Python 3 and Elasticsearch 7.0.
Data analysis	Data is analyzed with customized code in Python 3 and Stata 14.0 using standard software packages within these programs. The code necessary to reproduce main plots and statistical analyses is freely available at https://kellogg-cssi.github.io/science_public.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

MAG raw data is publicly available at https://docs.microsoft.com/en-us/academic-services/graph/. MAG-USPTO linkage data is publicly available at https://doi.org/10.5281/zenodo.3575146. Those who are interested in raw data of Altmetric and Dimensions should contact Digital Science directly. Those who are interested in raw data of Overton should contact Open Policy Ltd directly. The deidentified data necessary to reproduce main plots and statistical analyses is freely available at https://kellogg-cssi.github.io/science_public.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A study to quantify public uses and funding of science. We collect and integrate five large-scale datasets that link scientific publications from all scientific fields to their upstream funding support and downstream public uses across three public domains -- government documents, the news media, and marketplace invention. We then develop quantitative measurements to study field-level patterns of public use, as well as the relationship between public use, public funding, and scientific use of science.
Research sample	Scientific publications from Microsoft Academic Graph, together with their linkages to US government documents (.gov documents ranked as tier-0 documents by Microsoft Bing search engine), mainstream media news (collected by Altmetric), patents (recorded by USPTO), and funded research projects (collected by Dimensions).
Sampling strategy	No statistical methods were used to predetermine sample size. Each of the datasets represents the state-of-the-art large-scale corpus of its kind, and no random/snowball/stratified sampling is used. We used the full sample of scientific publications published in a 10-year period (2005-2014), a common period where the public uses of papers are covered by all three datasets. We also repeat our analysis on a larger sample across 30-year period (1985-2014) as robustness checks, finding our main conclusions remain the same.
Data collection	This study is based on pre-existing datasets and web crawling.
Timing	The MAG data was collected in 2018. The patent linkage data was collected in 2021. The other datasets were collected in 2019.
Data exclusions	The analysis has no data exclusions. Selection criteria within a dataset are described in Methods.
Non-participation	There are no participants in this study.
Randomization	This is a data driven study, not a randomized experiment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging