Hildegard Matthies · Dagmar Simon (Hrsg.)

Wissenschaft unter Beobachtung

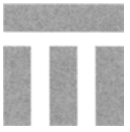
LEVIATHAN
Zeitschrift für Sozialwissenschaft

Sonderheft 24/2007

Hildegard Matthies · Dagmar Simon (Hrsg.)

Wissenschaft unter Beobachtung

Effekte und Defekte von Evaluationen



VS VERLAG FÜR SOZIALWISSENSCHAFTEN

Bibliografische Information Der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

1. Auflage 2008

Alle Rechte vorbehalten
© VS Verlag für Sozialwissenschaften | GWV Fachverlage GmbH, Wiesbaden 2008

Lektorat: Anne Vonderstein, Textetage Berlin

Der VS Verlag für Sozialwissenschaften ist ein Unternehmen von Springer Science+Business Media.
www.vs-verlag.de

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Umschlaggestaltung: Horst Dieter Bürkle, Darmstadt
Druck und buchbinderische Verarbeitung: Wilhelm & Adam, Heußenstamm
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier
Printed in Germany

ISBN 978-3-531-15457-2

Inhalt

Hildegard Matthies und Dagmar Simon
Einführung 9

I. Evaluation und Gesellschaft

Michael Power
Research Evaluation in the Audit Society 15

Marcel Weber
Wissenschaftstheorie der Evaluation 25

II. Qualitätsverständnis und Praktiken des Prüfens in historischer Perspektive

Rüdiger vom Bruch
Qualitätsmaßstäbe bei der Forschungsmittelvergabe in historischer Perspektive 47

Stefan Hornbostel
Neue Evaluationsregime? Von der Inquisition zur Evaluation 59

Jürgen Enders
Professor werden ist sehr schwer, Professor sein dann gar nicht mehr? Ein Beitrag zur Personalstrukturreform an den Hochschulen 83

III. Latente Funktionen und unintendierte Effekte

Dietmar Braun
Evaluation und unintendierte Effekte – eine theoretische Reflexion 103

Bruno S. Frey
Evaluitis – eine neue Krankheit 125

Sabine Maasen und Peter Weingart
Unternehmerische Universität und neue Wissenschaftskultur 141

Martina Röbbecke
Evaluation als neue Form der „Disziplinierung" – ein nicht intendierter Effekt? 161

Dagmar Simon
Als Konsequenz mehr Kohärenz? Intendierte und nicht intendierte Wirkungen von institutionellen Evaluationen 178

IV. Adäquatheit von Methoden und Gegenstand

Robert Salais
Evaluation und Politik: Auf der Suche nach guten Indikatoren für die Forschung 193

Wim Blockmans
Zählen die Geistes- und Sozialwissenschaften mit? 213

Matthias Bergmann und Thomas Jahn
Intendierte Lerneffekte: Formative Evaluation inter- und transdisziplinärer Forschung 222

Eva Barlösius
Urteilsgewissheit und wissenschaftliches Kapital 248

V. Von anderen lernen?

Thomas Widmer
Evaluationsansätze und ihre Effekte: Erfahrungen aus verschiedenen Politikfeldern 267

Torsten Strulik
Evaluationen in der Wirtschaft: Ratingagenturen und das Management des Beobachtetwerdens 288

Helmut Wiesenthal
Evaluation als Organisationslernen 315

Epilog I

Hildegard Matthies, Dagmar Simon und Andreas Knie
„Gefühlte" Exzellenz: Implizite Kriterien der Bewertung von Wissenschaft als Dilemma der Wissenschaftspolitik 331

Epilog II

Manfred Kohnke
Von Sternen und Kochmützen: Evaluation in der Haute Cuisine 347

Verzeichnis der Autorinnen und Autoren **355**

Hildegard Matthies und Dagmar Simon

Einführung

Die Prüfung und Bewertung wissenschaftlicher Leistungen spielt eine Rolle, solange es organisierte Wissenschaft gibt. Das gilt insbesondere für den epistemischen Gehalt von Wissen und die Erkenntnisansprüche der Wissenschaft. Ebenso war Wissenschaft immer schon durch Konformitäten geprägt. Die Einhaltung von impliziten Normen und Verhaltenserwartungen ist seit jeher ein wichtiges identitätsstiftendes Merkmal von Wissenschaft, wie etwa Robert Merton (1942) in seinen Analysen zum Ethos der Wissenschaft konstatiert.

Wiewohl Mertons Vorstellungen vielfach als zu idealistisch kritisiert worden sind, besteht kein Zweifel an der Feststellung, dass die Wissenschaft sich mittels einer Vielzahl von impliziten Regeln ein System der Selbststeuerung geschaffen hat, das auch Aspekte der Qualitätssicherung einschließt. Zu nennen sind etwa die Verfahren der Personalauswahl bei der Besetzung von Lehrstühlen, die Praktiken der Fördermittelvergabe, der Einführung neuer Lehrstühle oder Forschungseinrichtungen, und nicht zuletzt auch die Verfahren der Promotion und Habilitation des wissenschaftlichen Nachwuchses.

Dennoch haben sich sowohl die Diskussionen um die Leistungen der Wissenschaft als auch die Verfahren zu ihrer Kontrolle verändert. Ein Ausdruck dieses Wandels ist die flächendeckende Einführung von Evaluationen. Sie kann historisch betrachtet als Versuch der modernen Gesellschaft gelesen werden, bisher eher implizite Verfahren der Leistungsbewertung und Qualitätssicherung durch systematisierte Indikatoren und Bewertungsmodi expliziter und damit transparenter zu machen und somit auch die Bewertung von Wissen zu rationalisieren.

Allerdings zielt die Einführung von Evaluationen keineswegs nur auf die Sicherung der Qualität von Forschung und Lehre. Vielmehr erhoffen sich die wissenschaftspolitischen Akteure damit auch eine Stärkung der finanziellen und administrativen Kontrolle der von ihnen alimentierten wissenschaftlichen Einrichtungen. So ist es gerade die Verknüpfung von Buchführung und Evaluation, die Evaluationen in der Wissenschaft zu einer neuen „Form der Selbstregierung" (Strathern 2000: 287) gedeihen lässt und dafür sorgt, dass sie sich zu einem wichtigen Instrument der externen Steuerung von Wissenschaft entwickeln.

Die Suche nach der „golden goose of effectiveness" (Stanley Cohen, zit. nach Power 1997: 141) ist jedoch keineswegs ein Phänomen allein in der Wissenschaft. So beschreibt Power (1997) einen allgemeinen Trend zur Audit Society und meint damit einen Zustand, bei dem das Auditing als Teil einer Struktur der „Kontrolle der Kon-

trolle" fungiert. Tatsächlich hat der Anspruch, durch systematisches Prüfen mehr Gewissheit über die Legitimität öffentlicher Ausgaben herzustellen, stark zugenommen und lässt sich in zahlreichen anderen gesellschaftlichen Organisationen oder politischen Handlungsfeldern beobachten.

Derlei an Evaluationen geknüpfte Rationalisierungserwartungen könnten jedoch insbesondere für die Wissenschaft problematisch sein. Sie führen nicht nur zu überzogenen Erwartungen an das Instrument der Bewertung, sondern lassen in Evaluationsverfahren oftmals auch das Ziel der Bewertung unscharf werden: Geht es beispielsweise um die Bewertung von Erkenntnissen im Hinblick auf ihre epistemische Signifikanz, geht es um ihre gesellschaftliche Relevanz, oder geht es schließlich um den effizienten Einsatz von Ressourcen und um eine Optimierung ihrer Erträge, also um Fragen des Managements von wissenschaftlichen Institutionen im Hinblick auf generelle Stärken- und Schwächenanalysen, an deren Ergebnis häufig auch Entscheidungen über Ressourcenverteilungen geknüpft sind?

Solche multiplen Zielsetzungen gehen nicht per se ineinander auf, sie stehen vielmehr auch in Spannung zueinander. Ganz abgesehen davon, dass man durchaus darüber streiten kann, ob sich wissenschaftliche Effizienz überhaupt messen lässt: Effizienz und Qualität sind insbesondere in der Wissenschaft eben nicht zwei Seiten einer Medaille. Genauso wenig wie etwa eine bahnbrechende wissenschaftliche Erkenntnis an effiziente Produktionsbedingungen geknüpft ist, steht eine effizient geführte wissenschaftliche Einrichtung mit einem hohen Output automatisch auch für wissenschaftliche Spitzenleistungen. Wohl aber kann das Wissen um den multifunktionalen Einsatz von Evaluation unter den Akteurinnen und Akteuren in der Wissenschaft ungewollte Anpassungsleistungen dergestalt hervorrufen, dass nicht mehr die Suche nach neuem Wissen handlungsleitend ist, sondern die Frage, mit welchen Aktivitäten, Themen und Produkten am ehesten eine positive Evaluierung zu erwarten ist. Kritiker sprechen deshalb auch von einer „moderne[n] Form der Selbstgeißelung" (Weingart 2005), und die Zahl derer, die Evaluationen lieber heute als morgen abschaffen wollen, wächst. Auch wenn derlei Fundamentalkritiken oft ein Strukturkonservatismus anhaftet und sie einen nicht unproblematischen Status quo verteidigen – die Sorge, dass der Einsatz von Evaluationen als Multifunktionswaffe kontraproduktive Effekte für die Wissenschaft hervorruft, weil sich etwa kritische Forschung nicht mehr lohnt oder Organisationen durch Strukturempfehlungen in Sackgassen geführt werden, ist durchaus berechtigt.

Die unterschiedlichen Ziele – und vor allem die nicht intendierten Effekte und latenten Funktionen – von Evaluationen in der Wissenschaft in den Blick zu nehmen, ist das zentrale Anliegen dieses Buches. Es versammelt Beiträge von Autoren und Autorinnen, die sich vom 1. bis 3. Juni 2006 auf einer am Wissenschaftszentrum Berlin für Sozialforschung zum Wandel der Qualitätssicherung von Wissenschaft durchgeführten Tagung zusammengefunden haben, sowie von weiteren prominenten Autoren der Wissenschafts- und Evaluationsforschung.

Allen Beiträgen ist gemeinsam, dass sie einen reflexiven Zugang zum Thema wählen – sei es aus dem Blickwinkel der Wissenschaftstheorie, der Evaluationsforschung oder der Evaluationspraxis. Dabei lassen sich fünf Diskussionsblöcke voneinander unterscheiden, die im Folgenden kurz skizziert werden sollen.

Evaluation und Gesellschaft

In diesem einführenden Teil wird zum einen das offensichtlich gewachsene Bedürfnis des Prüfens und Bewertens in den gesellschaftspolitischen Kontext der Audit Society gestellt, zum anderen wird die Evaluation von Wissenschaft philosophisch beleuchtet und im Sinne des Seinsollens eine Weiterung des Evaluationsbegriffs im Hinblick auf gesellschaftliche Anforderungen an die Wissenschaft unternommen.

Qualitätsverständnis und Praktiken des Prüfens in historischer Perspektive

Im zweiten Block wird das Verständnis von Qualität sowie die Praktik des Prüfens in der Wissenschaft aus einer historischen Perspektive beleuchtet. Dabei geht es um Fragen wie: Worin unterscheiden sich die neuen Verfahren der Evaluation im Vergleich zu früheren traditionellen Methoden und Praktiken der Bewertung von Wissenschaft? Nach welchen Kriterien etwa wurden Lehrstühle eingerichtet und besetzt, Forschungsgelder verteilt oder der wissenschaftliche Nachwuchs ausgewählt? Welche der alten, mehr oder weniger impliziten, Bewertungsmodi bestehen in den neuen Verfahren – beabsichtigt oder unbeabsichtigt – fort? Oder anders gefragt: Was ist das eigentlich Neue an den modernen Verfahren der Evaluation von Wissenschaft?

Latente Funktionen und unintendierte Effekte von Evaluationen

Im dritten Block stehen die latenten Funktionen sowie die unintendierten Effekte von Evaluationen im Zentrum der Diskussion. In diesem Zusammenhang aufgeworfenen Fragen sind beispielsweise: Haben wir durch Evaluationen Disziplinierungseffekte zu befürchten, indem etwa Wissenschaftler und Wissenschaftlerinnen kritische Themen meiden und es vorziehen, ihre Forschungsthemen am Mainstream wissenschaftlicher Erkenntnis auszurichten? Bestimmen im Zuge des Evaluationsgeschehens zunehmend Moden das Geschäft in der Wissenschaft? Droht der Wissenschaft möglicherweise sogar der Verlust ihrer Autonomie, wenn das Evaluieren zunehmend zum Geschäft professioneller Evaluationsagenturen außerhalb der Wissenschaft wird?

Adäquatheit von Methoden und Gegenstand

Im vierten Diskussionsblock werden vor allem Fragen zur Adäquatheit von Methoden und Gegenstand aufgeworfen, etwa: Lässt sich Angemessenheit überhaupt bestimmen? Oder: Welche Probleme implizieren die gängigen Verfahren im Hinblick auf besonders komplexe Evaluationsgegenstände, etwa die interdisziplinäre Forschung

oder Forschungseinrichtungen, die sich nicht mehr in das klassische Raster von Grundlagen- oder anwendungsorientierter Forschung einordnen lassen?

Von anderen lernen?

Die Beiträge im fünften Block verfolgen schließlich das Ziel, mit dem Buch auch Antworten auf die Frage nach den Möglichkeiten der Vermeidung negativer Effekte von Evaluationen zu geben, z. B. welche Indikatoren und Methoden sich als geeignet identifizieren lassen, um einen von seiner Zielsetzung her so unbestimmten Gegenstand wie Wissen angemessen zu bewerten, oder welchen Nutzen die Wissenschaft aus Bewertungsverfahren in anderen Politikfeldern oder aus der Wirtschaft ziehen kann. Der Abschnitt schließt mit der Frage, welche Bedingungen erfüllt sein müssen, damit Evaluationen ein Impulsgeber für das Lernen von wissenschaftlichen Einrichtungen sein können.

Epilog

Das Buch endet mit zwei verschiedenen Formen der Nachrede. Im ersten Epilog geht es um das Kontrolldilemma, dem der Staat im Kontext von externen Evaluationen ausgesetzt ist, weil er sich sein Urteil über Exzellenz und Effizienz der Wissenschaft nur mit Hilfe der Wissenschaft bilden kann. Er ist somit auf eine vertrauensvolle Zusammenarbeit mit der Wissenschaft angewiesen, was den Möglichkeiten der externen Kontrolle Grenzen setzt. Im zweiten Epilog wird das Feld der Wissenschaft verlassen und ein Ausflug in die gehobene Küche unternommen. Was einem beim Speisen vermutlich kaum je durch den Kopf gehen dürfte: in Evaluationsangelegenheiten liegen beide Bereiche gar nicht so weit auseinander. Die Frage jedenfalls, was jemanden dafür qualifiziert, über gute oder schlechte Ergebnisse zu entscheiden, treibt auch die Restaurantkritik um.

Literatur

Power, Michael (1997): The Audit Society: Rituals of Verification. Oxford: Oxford University Press.

Merton, Robert (1942): The normative structure of science. In: Robert K. Merton (1973): The Sociology of Science. Theoretical and Empirical Investigations. Chicago/London: The University of Chicago Press.

Strathern, Marilyn (2000): Afterword: Accountability ... and ethnography. In: Dies. (Hg.): Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy. London: Routledge.

Weingart, Peter (2005): Die Wissenschaft der Öffentlichkeit. Weilerswist: Velbrück.

I. Evaluation und Gesellschaft

Michael Power

Research Evaluation in the Audit Society[1]

Introduction

A challenging analysis of the impact of a new Total Quality Management (TQM) system on the basic research practices of a commercial laboratory demonstrates how the perceived 'value incongruence' between the tasks of the scientists and TQM can create distrust, anger and resistance (Sitkin/Stickel 1996). Indeed, the introduction of formal controls with the manifest intention of enhancing trust and transparency within the organisation seemed to have the opposite effect. The study reports how the performance measures embodied in the TQM system were perceived as 'inappropriately precise and deterministic' in relation to a basic research task, which was seen to be inherently ambiguous and highly uncertain. In one reported example, a laboratory manager refused to inform staff about the patent goals of the organisation, arguing, in a manner consistent with Goodhart's law, that such targets were not only distracting but might affect the propensity of scientists to take intellectual risks (204). The manager had effectively established himself as a 'buffer' between the creative scientists and the management system.

The conflict was bilateral: management was critical of the scientists' refusal to share in the larger values of the organisation and to import a TQM system which had demonstrated its success in other areas. In turn, scientists were critical of the enforced uniformity of TQM, believing their position to be in some sense unique and beyond the reach of control systems.

The key features of this story have become a somewhat familiar starting point for scholars interested in the evaluation of science and its consequences. TQM is emblematic of the broader 'audit society' thesis (Power 1997), which draws attention to the counterproductive invasion of spheres of professional autonomy through apparently crude managerial instruments, such as accounting, performance measurement and audit. For the private sector scientists in the above circumstances, we might substitute teachers, doctors and social workers with only a slight change of emphasis. The observed value incongruence between the tasks of scientists and quality management is part of a broader perceived incongruity between the professional actors and formal systems of evaluation and control.

---

1 An earlier version of this paper was presented at a meeting of the Network of Institutes for Advanced Study (NetIAS), April 20, 2007, Bologna.

In the United Kingdom, an idealised model of the financial auditing process gained organisational and social significance as a template for evaluative practice during the 1980s. This normative idea of audit – and the implied possibility of making visible private clubs of practice – became highly attractive to reformers of the public sector and played a central role in the conception of a new public management as manifested in the United Kingdom. From this point of view, audit is far from being a neutral practice of monitoring and is a vehicle for the dissemination of distinctively 'managerial' values and ideas. TQM and similar initiatives can be understood as practices of self-observation and control which can also be externally observe and validated. Audit seems to be about innocent checking but is in fact part of a larger programme for radical change. The idea of the 'audit society' has provided a focus for counterdiscourses of resistance for different professional groups, particularly in the field of education.

Yet, as the contributions to this volume show, beneath this almost conventional account of professional autonomy, under threat is a more complex picture, particularly in the case of the evaluation of science. This short essay focuses on three substantive themes which deserve further attention in progressing the audit society debate and which are developed in the following chapters. First, we deal with the progressive transformation of evaluative practice by audit. Second, we consider the irony that auditing and evaluation, far from being 'value incongruent' with science, in fact draw on science for their own legitimation. Third, we reflect on the complex relationship between peers and auditors. In conclusion, we consider some further aspects of the way in which science is becoming a managerial and regulatory object.

1 *The 'Auditisation' of Evaluation?*

An enduring puzzle about the 'audit society' hypothesis concerns the relation between audit and evaluation, and the significance of this relationship for science. The distinction is made in a variety of ways, although there is general acceptance that that there are two practice clusters which have had very different histories, techniques and purposes. For example, programme evaluation can be associated with expansionism in welfare programmes and the application of the social sciences – including forms of cost-benefit analysis – in the assessment of whether such programmes achieve their intended effects (Mosher 1979; Power 1997: 115-119). In contrast, the history of financial auditing reveals a relatively narrow focus on the quality of financial statements produced by private companies. Whereas audit has developed hand in hand with the creation of prescribed standards for accounting, evaluation operates according to the standards of enquiry customised for specific applications. Audit is more like a check of conformity to pre-existing standards of 'best practice'; evaluation is more like an investigation of what has happened, an analysis of effects. Audit has been typically, but not exclusively, the province of accountants; evaluation is more the

province of social scientists and professionals who understand the practice domain which they evaluate.

Yet, while the distinction between audit and evaluation can be, and often is, made in these or very similar terms, 'audit' and 'evaluation' should not be regarded as discrete categories which refer to specific forms of practice. They have a familial resemblance as mechanisms of monitoring which are continually evolving and, critically, coming into a relation with each other in the creation of new forms of practice and bodies of expertise. In particular, the phenomenon by which evaluation practices come to be 'auditised' deserves investigation. What is meant by 'auditisation' is the processes, explicit or implicit, by which practices of evaluation come to constitute themselves in the shadow of the financial audit model – specifically through the emergence of best-practice standards of performance which can be checked. It is striking that, even though the efficacy of the financial audit model is a subject of constant debate between financial auditing practitioners themselves, it has come to exert a normative force over adjacent fields of monitoring, which have changed and become more managerial in form.

The reasons why evaluation has become more like audit in its style of operation are undoubtedly complex, but we can point to a number of suggestive issues. First, we can point to the decline of science as a cultural model informing public policy even though, as noted below, echoes of science are inherent in the model of audit. This decline is not to deny that science is widely regarded as strategically significant for economies. Indeed, it is precisely because it is such a significant 'policy object' that its fortunes can no longer be entrusted to scientists alone. Macroeconomic strategy in many countries demands that science be made governable (e.g., OST 1993; Robson 1995). While efforts to manage public science in the name of competitive economic advantage have a long history, they have taken on a new form as part of a cluster of programmatic changes in public management which took place in the 1980s and 1990s. The effectiveness of publicly funded science is to be governed and managed at the organisational level by practices such as TQM discussed above, together with auditing and accounting.

The second reason is a mirror-image of the first, namely the emergence of audit as a powerful regulatory model exerting 'isomorphic' pressures on other related practices. Audit is attractive as a neoliberal policy instrument because of its normalising impact on organisations, specifically via its demands for the formal articulation of control systems, such as TQM, and of auditable performance metrics. Whereas complex social and economic impacts may be evaluated using social scientific instruments, effects are often long term and disputed by evaluators themselves. The development of auditable performance measures in the form of visible outputs within short reporting timeframes is widely admitted to be a simplification, but it is attractive for that very reason as a useable method to discipline practices, including scientific practices, in the name of policy ideals. The suggestion that evaluation has become increasingly 'auditised' focuses attention on the pressures for evaluative practices to adopt more

elements of the audit model. In this way a science of checking against best-practice standards has the potential to displace the science of enquiry and investigation (Gray and Jenkins 2007). Furthermore, the audit model is performative and has come to be perceived as increasingly legitimate as organisational structures are themselves transformed to be compatible with it. For example, the rise of 'audit' committees, rather than evaluation committees, may seem only verbal in nature but is symptomatic of the power of the audit model even as its spectacular failures have become apparent early on in the twenty-first century. Big science has also had its own share of scandals and fraud leading to heightened interest in 'data audit'. In this setting, the audit model is imported to shore up the legitimacy of science (e.g., Francis 1989).

2 *Auditing as Science*

In their analysis, Sitkin and Stickel (1996) note in passing an important puzzle about TQM, namely that, in resisting it, scientists were arguing in part that they had already solved the problem that TQM is trying to solve. In short, far from being an entirely alien body of knowledge, TQM shares many of the characteristics of science itself. These characteristics are easy to sketch, although their precise history is complex. One line of development shows how concerns about engineering practice and product quality led to the design of testing or feedback loops within production systems. Faults and errors became data for investigation and process reform. In some cases, systems could be stressed by the deliberate introduction of faults in a process analogous to falsification.

The quality movement projected this cybernetic or machine-like conception of control onto the management process itself, and as it moved further from its specific origins, the model became more abstract, formalised and capable of being written in the standardised form now familiar to us in ISO 9000 and its variants. The same can be said for the managerial ideas of internal control and risk management which have developed in parallel with quality assurance (Power 2007). So, an important element in the analysis of the corporate laboratory above is the reintroduction into science of ideas and processes which have been abstracted from, and externalised by, science itself. In TQM, science paradoxically confronts a hyper-real version of itself. TQM represents a success for science as a cultural force which simultaneously undermines that force; processes of 'scientisation' may be self-undermining.

Extending this argument to the case of financial auditing, it is not difficult to argue that auditing itself can be reconstructed as a scientific process, namely as a process of seeking to refute assertions about financial statement items. Indeed, early conceptualisations of auditing in the United States placed the evidence process at its very heart –reconstructing it as a hypothetic-deductive practice (Mautz/Sharaf 1961). The normative task here was to bring auditing closer to the cultural values of science and, in so doing, emphasise the quality of the evidence process. Not surprisingly perhaps,

auditing thought and practice in the United States began to import statistical sampling techniques from engineering after this period. In Europe, auditing was constituted within a more legalistic institutional framework.

Thus, an enduring puzzle about 'audit society' pressures in relation to science is that the very idea of audit itself can be said to be scientised. Audit practice has sought legitimacy by importing the values and routines associated with a cultural model of science, namely rationalisation, evidence and formalisation. In being audited, science confronts something which is not entirely alien, namely a 'scientised' version of itself. This 'science of accounts' has extended its reach to create new sensibilities about the financial accountability of science (Power 1994). In the process, a powerful managerial model of objectivity has been created with far-reaching consequences for how policy-makers think (Porter 1992).

3 Peers and Auditors

The third significant theme in the progression of the debate about evaluation and the audit society concerns the role of scientists themselves in the process of evaluation and, specifically, the conception of 'peer review'. In Sitkin and Stickel's analysis, 'value congruence' between scientists and TQM is the key theme. However, they focus on an early phase in the process of introducing the TQM system, and there is some suggestion that the resistance and anger exhibited by subjects waned over time. Further empirical research may have revealed that the scientists became resigned and fatalistic. Maybe they began to see the benefits of TQM. Or maybe they figured out ways to carry on as before, unaffected. Elements of all these strategies are likely to be true. Of particular interest is the possibility that scientists themselves would begin to own and run the TQM system, that is, that the evaluation of effectiveness would be conducted by 'peers'.

The idea of 'peer review' is a significant component of evaluation thinking for very obvious reasons. Peers are, by definition, experts in the field to be evaluated and already part of the self-checking process of science. There is, therefore, likely to be little or no value incongruence between evaluator and evaluated. Peers also play a potential role as 'buffers' between the integrity and autonomy of scientific endeavour and the demands of an audit society fixated on measurable outcomes. Peers are guardians of the scientific content of evaluation itself, being the human embodiment of a function which is constitutive of science itself – conjecture and refutation. On this view, evaluation by peers and practices of scientific replication can be regarded as closely related, and peer review can be seen as an antidote to the 'auditisation' of evaluation.

Yet, how much do we know about what happens to peers and peer review processes as they become more formalised and public? The history of clinical audit provides an instructive example. Beginning as the private analysis by medical practitioners

of data about their procedures – a local group peer process – clinical audit came to be co-opted, transformed and externalised as part of something called clinical governance (Scally/Donaldson 1998). Scientific peers themselves undergo complex changes in this process. Tightly defined 'epistemic communities' give way to new hybrid roles which confer power and prestige on the incumbents (Kurunmaki 2004), and the evaluation process begins to professionalise as an expert system in its own right (Lindeberg 2007).

In the United Kingdom, it is being suggested that the Research Evaluation Exercise (RAE) to be conducted in 2008 involving peer review might be the last of its kind and that a more 'metric-based' system might be used in future. Yet the contrast between peer review and metrics, such as citation analysis, is often overdrawn. Formally constituted panels of peers – not least because of workload reasons – operate with rules-of-thumb guides as to quality and draw on, for example, journal rankings and impact analysis. UK university administrations have also internalised metrics to varying degrees as a basis for organising and preparing for the RAE. Risk analyses of research productivity, down to the level of individual research staff, necessarily operate with quantitative benchmarks – all this in anticipation of the 'real' evaluation. So the current UK system for the evaluation of research operates as a hybrid because even the most well-intending peer needs the practical support of an auditable benchmark even if there are myths of 'reading everything'.

Accordingly, as the contribution by Eva Barloesius suggests, we need to understand more about the social dynamics of scientists when they enter formal peer review processes and the distinctive forms of symbolic capital which the evaluation process confers on them. The challenging hypothesis is that peers, far from being a buffer against the audit society, may in fact operate as its ambassadors. And because activities of evaluation are not themselves valued by an exercise like the United Kingdom's RAE – except indirectly via indicators of esteem – peer review may itself not attract the most competent reviewers in a field. Evaluation is not generally something that the best and most active scientists wish to do – they prefer the activity of science itself. All of this means that the category of 'peer' is more problematic than it often appears and should not be regarded as necessarily antithetical to the audit model.

4 *Science as a Managerial Object*

As the different contributions to this volume show, we are beginning to know more about the complexities of science as an object of management. We are also beginning to know more about the consequences of evaluative regimes for science. Science has always in some sense been evaluated and controlled (e.g., Crosland 1992). In recent times, journal reviewing processes have moved to the centre of the evaluation stage. Scientific replication can be understood as a check on the quality of science, although as Collins (1985) reminds us, pure replication never happens – in part because there is

no great reward for doing this without also advancing knowledge in the process. Yet, pure checking has also come to be institutionalised in light of concerns about data quality and the falsification of scientific results, particularly in the context of drug trials. New hybrid bodies of expertise have come into being and new journals, such as *Accountability in Research* created in 1989, are symptomatic of the growing mood of enquiry about the quality and effectiveness of science (Shapiro 1992). This is a managerial and regulatory mood which points to the significance of quality control systems for science, such as the one described by Sitkin and Stickel (1996).

The governance of science has emerged, albeit unevenly, in different countries as part of the rise of governance more generally (Drori 2006) within new neoliberal conceptions of public management, regulation and economic policy. Research organisations must account in new and more elaborate ways, and funding bodies in some countries have become more tightly coupled to macroeconomic policy objectives. Demands exist for research organisations to produce grand narratives of integration, interdisciplinarity and collective performance – constructed out of the activities of groups and individuals with a high degree of autonomy, and which also compete with one another. As in the case of public financial reports produced by large private companies, we can assume that very few of these reports will be read by the funding bodies and trusts which demand them – they are simply components of institutionalised chains of accountability linking science to politics. It is not fanciful to suggest that *production* of reports about the effectiveness of science and research activity has more institutional significance in conferring legitimacy than whether they are read or lead to action. At the limit, research units and universities become self-auditable as management attention shifts from individual scientists as experts to the capacity of scientific organisations to self-regulate. For example, the 2008 RAE in the United Kingdom will seek to evaluate the quality and esteem of the research environment in which individual scientists work. And as a result of a UK white paper in the early 1990s (OST 1993), there was a growth of project management courses and similar training packages being offered to scientists and research managers to enable them to cope with a new world of organisational accountability.

At the heart of the governance of science is the problem of performance measurement and the role of metrics, such as citation and impact analysis. Much is known about the limitations of metrics and the games of 'creative compliance' that take place around them. The chain of social production behind these indices becomes quickly lost from view, as Robert Salais' contribution in this book reminds us. While there are efforts to conceptualise indicators of research effectiveness as part of a broader, more intelligent conversation about science, we also know that indicators have a life of their own which defines identities and shapes motivations. Targets give rise to new forms of 'careerism' and strategic choice which, arguably, create a low appetite for intellectual risk-taking. Organisations and individuals discover that reputation is constructed by external rating and ranking bodies and adapt their behaviour accordingly.

Formalised indicators always contain the seeds of their own demise – as Goodhart's law suggests – and Bruno Frey's work on motivational crowding (in this book) defines an analytical and empirical space within which perverse side-effects may be analysed. At the extreme, auditised ex post evaluation mechanisms may be a 'fatal' remedy (Sieber 1981), in so far as productive capacities and appetites for risk are fundamentally damaged, not only in science, but also in teaching, medicine and much else besides. Citation-measures normalise the middle ground of the 'moderately new', and radical innovators – who potentially have no peers who might cite them – become invisible. Within medical research it can be observed that demonstrations of success are cited more often in comparison with demonstrations that a drug or procedure does not work – because of fashion and the politics in research funding.

In the face of these difficulties and side-effects, which are to a greater or lesser extent acknowledged and known, systems of evaluation for science exist in a near constant process of reform. There is always a search for 'better' indicators which address unintended consequences. For example, Sitkin and Stickel (1996: 211) suggest that the pathology of performance measures and value in congruence might be addressed by framing formal specifications 'in terms of broad objectives rather than *precise* measures,' (emphasis added). This is a remedy which is currently much discussed within financial regulation in general and accounting in particular. There is a growing policy belief that detailed rules create problems which could be solved by the use of principles or, in other words, broadly framed rules. Yet this strategy is more of a symptom rather than a cure. Studies in accounting and law show clearly a dynamic by which broad rule systems become defined more precisely over time, not least because organisational agents demand certainty and clarity. It can also be argued that the dynamic which produces precise rules is one which is also driven by deeply engrained values of 'auditability' (Power 2007). Precise measures of performance may have all the disadvantages listed above, but they also have one telling advantage: they make things auditable and checkable by remote centres of control.

The policy imperative for the production of auditable facts is most problematic in settings characterised by high uncertainty, such as basic research. However, it is important to be mindful of cross-disciplinary variations in the value incongruence problem. While audit culture may be highly threatening to anthropologists (Strathern 2000a; 2000b), there are signs that economics and the natural sciences operate with a reasonably well-defined evaluative architecture. There may be some self-referential citation 'clubs' and some evidence that North American scientists cite only North American work, but the availability of a very large number of citations may be the best low-cost proxy for quality.

Many contributions in this volume point to paradoxes and pitfalls of formal evaluation practices for science. Yet somehow the activities of scientists are increasingly being made auditable and manageable. Formal remedies which may be resisted, at least initially, nevertheless begin to have a policy and managerial life because they create a certain kind of policy knowledge which is demanded by the 'new' mood of

public management for science in different countries. These instruments are attractive because they create a window on the private club of science, albeit a window which obscures as much as it reveals. Science must be made 'readable' (Scott 1998) and auditable. It is only when we fully acknowledge the policy power of audit and the values of precision with which it operates, and which it diffuses into practices of evaluation, that we can begin to understand how the growing recognition of side-effects – not only in the pages of this volume, but also in policy circles – has had little effect to date on practice.

Finally, a speculation. The management guru Tom Peters is reported to have said that we should only measure the things we want to see more of. Following this line of reasoning, would it be possible to invent a better set of lead indicators of research quality which could indeed change behaviour to give us more of what we want from science? If we value interdisciplinary work, should we perhaps give citations more weight when they are from authors outside of the home discipline/field of the study? Or if we value a quality control process which weeds out poor research, could we count the number of times that papers are presented to groups of peers for serious challenge? And could we even develop indicators for what Kuhn has described as the 'serendipitous' nature of scientific creativity? Could the creation of institutional capacities 'to be lucky' ever be measurable and auditable? These are playful suggestions and no doubt any new measures would have their own perverse effects, but they might also create new conversations around the management of science and place citation metrics in their place as just one component within a rich evaluative methodology. One of Sitkin and Stickel's interviewees is reported to have said that 'finding measures is easy . . . you need courage to find real indicators.' Yet perhaps the real challenge is to allow different indicators to flourish in an open market for evaluation, to create a market democracy of different measures and to demonopolise the citation industry.

References

Collins, Harry (1965): Changing Order: Replication and Induction in Scientific Practice. London: Sage.

Crosland, Maurice (1992): Science under Control: the French Academy of Science 1795-1914. Cambridge: Cambridge University Press.

Drori, Gili S. (2006): Governed by Governance: The New Prism for Organizational Change. In: Gili S. Drori, John W. Meyer and Hokyu Hwang (Eds.): Globalization and Organization: World Society and Organizational Change. Oxford: Oxford University Press: 91-118.

Francis, Jere R. (1989): The Credibility and Legitimation of Science: A Loss of Faith in the Scientific Narrative Accountability in Research: Policies and Quality Assurance 1(1): 5-22.

Gray, Andrew/Bill Jenkins (2007): Checking Out? Accountability and Evaluation in the British Regulatory State. In: Marie-Louise Bemelmans-Vidsec, Jeremy Lonsdale and Burt Perrin (Eds): Making Accountability Work: Dilemmas for Evaluation and Audit. (London: Transaction): 143-162.

Kurunmaki, Lisa (2004): A hybrid profession: The acquisition of management accounting expertise by medical professionals. Accounting, Organisations and Society 29(3-4): 327-348.

Lindeberg, Tobias (2007): Evaluative Technologies: Quality and the Multiplicity of Performance. Copenhagen: Copenhagen Business School.

Mautz Robert. K./Hussain. A. Sharaf (1961): Philosophy and Principles of Auditing. Sarasota, FL.: American Accounting Association.

Mosher, Frederick C. (1979): The GAO: The Quest for Accountability in American Government Boulder, Co.: Westview Press.

OST (1993): Realising our Potential: A Strategy for Science, Engineering and Technology,

Porter, Theodore M. (1992): Quantification and the Accounting Ideal in Science. Social Studies of Science 22 (4): 633-651.

Power, Michael (1994): From the Science of Accounts to the Financial Accountability of Science, Science in Context 7(3): 355-387.

Power, Michael (1997): The Audit Society: Rituals of Verification. Oxford: Oxford University Press.

Power, Michael (2007): Organized Uncertainty: Designing a World of Risk Management. Oxford: Oxford University Press.

Robson, Keith (1995): Connecting Science to the Economic: Accounting Calculation and the Visibility of Research and Development. In: Michael Power (Ed.) Accounting and Science. Cambridge: Cambridge University Press: 151-169.

Scally, Gabriel/Liam J. Donaldson (1998): Clinical Governance and the Drive for Quality Improvement in the New NHS in England,' British Medical Journal 317: 61-65.

Scott, James (1998): Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. New Haven: Yale University Press.

Shapiro, Martin F. (1992): Data Audit by a Regulatory Agency: Its effect and Implication for Others. Accountability in Research: Policies and Quality Assurance 2(3): 219-229.

Sieber, Sam D. (1981): Fatal Remedies: The Ironies of Social Intervention. New York: Plenum.

Sitkin, Sim B./Darryl Stickel (1996): The Road to Hell: The Dynamics of Distrust in an Era of Quality. In: Roderick M. Kramer and Tom R. Tyler (Eds.): Trust in Organizations: Frontiers of Theory and Research Thousand Oaks. CA: Sage Publications: 196-215.

Strathern, Marilyn (2000a) (Ed.): Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy. London: Routledge.

Strathern, Marilyn (2000b): The Tyranny of Transparency. British Educational Research Journal 26(3): 309-321.

Marcel Weber[1]

Wissenschaftstheorie der Evaluation

Einleitung

Die Wissenschaftstheorie beginnt in der Regel damit, dass sie ein Ziel für die Wissenschaft setzt und dann philosophische Überlegungen anstellt, wie dieses Ziel erreicht werden kann. Als Ziele der Wissenschaft gelten traditionell Wissen oder Erkenntnis, die als System wahrer und begründeter Sätze verstanden werden. An Mitteln, Erkenntnis zu erreichen, stehen als die üblichen Kandidaten zur Verfügung: deduktives und induktives Schließen, Aufstellen und kritische Überprüfung von Hypothesen, Experimente usw. Hingegen kommt in der klassischen Wissenschaftstheorie die Frage, worin der Wert von wissenschaftlicher Erkenntnis eigentlich gründet, seit jeher zu kurz. Wissenschaftstheoretiker scheinen es für evident zu halten, dass Erkenntnis etwas Wertvolles ist, denn sie setzen diese als Ziel der Wissenschaft einfach voraus.

Wenn wir nun beginnen wollen, eine Wissenschaftstheorie der Evaluation zu entwickeln, scheint sich eine analoge Vorgehensweise aufzudrängen: Wir fragen zunächst, worin das Ziel der Evaluation besteht, und stellen dann Überlegungen an, welches die geeigneten Mittel zur Erreichung dieses Ziels sind. Das Ziel selbst ist schnell benannt: die Sicherung oder Verbesserung der wissenschaftlichen Qualität, wissenschaftliche Exzellenz sowie die Effizienz und „accountability" in Bezug auf die Vergabe von öffentlichen Geldern. (Voraussetzung dafür ist natürlich, dass die Evaluation an geeignete Steuerungsmittel gekoppelt wird.) Wesentlich umstrittener ist die Frage, mit welchen Methoden diese Ziele am besten zu erreichen sind. Zur Diskussion stehen die klassische Methode des Peer Review, empirische Indikatoren (z. B. die Bibliometrie), konsultative Verfahren mit Beteiligung von Experten bis hin zur demokratischen Partizipation. Ich möchte in diesem Beitrag zeigen, dass die Angemessenheit dieser Evaluationsmethoden nur auf der Grundlage einer Axiologie (d. h. einer Werttheorie) wissenschaftlichen Wissens fundiert beurteilt werden kann. Ich möchte außerdem verschiedene Möglichkeiten aufzeigen, wie eine solche Axiologie sich entwickeln ließe und welche Implikationen verschiedene Typen von Axiologien für die Wissenschaftsevaluation haben.

Im folgenden Abschnitt werde ich eine Unterscheidung einführen, die ich für die Diskussion über die Wissenschaftsevaluation für fundamental halte. Ich möchte gleich

1 Ich danke Paul Hoyningen-Huene, Martin Reinhart und Daniel Sirtes für wertvolle Anregungen und die kritische Durchsicht des Manuskripts.

zu Beginn klarstellen, dass „Evaluation" hier in einem weiteren Sinn verstanden wird als beispielsweise im Beitrag von Bruno Frey (in diesem Band). Ich verstehe darunter die Gesamtheit aller Kontexte, in denen wissenschaftliche Arbeiten bewertet werden, sei es durch Fachkolleginnen und -kollegen (Peers), durch Fachpersonen, die nicht selbst an der Forschung mitwirken (z. B. „science administrators") oder auch durch wissenschaftliche Laien, zum Beispiel Politkerinnen und Politiker. Im zweiten Teil möchte ich aufzeigen, welche Grundtypen einer Axiologie wissenschaftlichen Wissens es gibt und welche Implikationen diese für die Evaluation von Wissenschaft haben. Drittens möchte ich eine ganz bestimmte Konzeption kritisch beleuchten, nämlich Philip Kitchers Konzeption von „wohlgeordneter Wissenschaft". Im vierten und letzten Kapitel möchte ich auf Grundlage des bisher Erreichten einige programmatische Vorschläge zur Weiterentwicklung der Wissenschaftstheorie der Evaluation geben.

1 Enge versus weite Evaluation

Ich möchte unter *enger* Evaluation die Bestimmung verstehen, wie weit eine Forschungsarbeit den spezifischen Qualitätsstandards einer wissenschaftlichen Disziplin entspricht. Dagegen soll im Folgenden *weite* Evaluation die Bestimmung bezeichnen, wie gut eine Forschungsarbeit die von der Gesellschaft an sie gestellten Anforderungen erfüllt. Ich möchte betonen, dass diese Unterscheidung nicht notwendigerweise mit der zwischen interner und externer Evaluation zusammenfällt. Die letztere Unterscheidung ist strukturell; sie bezieht sich darauf, in welchem institutionellen Verhältnis die evaluierenden Personen zu der zu evaluierenden Forschung stehen, insbesondere ob sie Teil der relevanten Scientific Community sind oder nicht. Hier greift eine einfache Dichotomie wahrscheinlich zu kurz. Collins und Evans (2002) weisen darauf hin, dass es rund um das Unternehmen Wissenschaft nicht einfach bloß Experten und Laien gibt, wie häufig angenommen wird. Es muss mindestens zwischen partizipierenden oder beitragenden Experten in der gleichen Disziplin, partizipierenden Experten in einer Nachbardisziplin, interaktionalen Experten und Laien unterschieden werden. Zeitschriften-Peer-Review ist zum Beispiel ausschließlich Sache beitragender Experten im gleichen Fach, während die Begutachtung von Forschungsprojekten durch forschungsfördernde Institutionen oft auch partizipierenden Experten in benachbarten Disziplinen obliegt. Doch auch interaktionale Experten wie Hochschul- und Wissenschaftsadministratoren sowie Laien, zum Beispiel Politiker und Politikerinnen, wirken an der Evaluation und an der Steuerung der Forschung mit.

Meine Unterscheidung zwischen enger und weiter Evaluation bezieht sich nicht auf die Frage „*Wer* evaluiert?", sondern auf die Frage „Auf welcher *normativen Basis* wird evaluiert?". Sie ist daher keine strukturelle, sondern eine funktionale Unterscheidung. Sie bezieht sich auf die relevante Wertgrundlage, auf der die Evaluation aufbaut (man sollte nicht vergessen, dass Evaluation immer etwas mit Werten zu tun hat; wie

schon am Wort offensichtlich ist). Handelt es sich um die konstitutiven Normen und Werte einer spezifischen wissenschaftlichen Disziplin, spreche ich von enger, andernfalls von weiter Evaluation. Dabei soll gerade *nicht* angenommen werden, dass es sich bei der normativen Basis der engen Evaluation ausschließlich um epistemische und bei der weiten Evaluation nur um nicht epistemische Werte handelt. Das bedeutet unter anderem, weder von der Voraussetzung auszugehen, dass Nichtwissenschaftler die Ergebnisse der Forschung ausschließlich anhand ihres Anwendungsnutzens bewerten, noch davon, dass Wissenschaftler reine, interesselose Erkenntnissubjekte sind, denen allein an der Wahrheit gelegen ist. Was ich mit der Unterscheidung von enger und weiter Evaluation einfangen möchte, ist die Tatsache, dass wissenschaftliche Disziplinen mit einem spezifischen Satz von Normen und Werten verknüpft sind (Hoyningen-Huene 1993: 147-154), die gewisse Handlungen und gewisse Ergebnisse auszeichnen und andere sanktionieren, und dass diese von den Normen und Werten anderer Sektoren der Gesellschaft verschieden sind. Dieses Wertegefüge ist natürlich von den Werten und Interessen des größeren sozialen Kontextes, in dem die Wissenschaft praktiziert wird, nicht unbeeinflusst[2], dennoch ist es von diesem *unterscheidbar*. Dies genügt, um die Unterscheidung zwischen weiter und enger Evaluation signifikant zu machen. Ich werde nun als Nächstes einige Anmerkungen zur engen Evaluation machen. Der Rest dieses Aufsatzes wird dann aber den wissenschaftstheoretischen Grundlagen der weiten Evaluation gewidmet sein.

Die enge Evaluation wird traditionellerweise beitragenden Experten zu der Disziplin, aus dem die zu evaluierende Arbeit stammt, überlassen. In manchen Fällen werden aber auch beitragende Experten zu benachbarten Disziplinen herbeigezogen, vor allem bei der Vergabe von Forschungsmitteln. Diese kollektiv als Peer Review bezeichneten Verfahren sind in letzter Zeit wiederholt kritisiert worden. David Shatz (1996) hat etwa darauf hingewiesen, dass die Legitimation von Zeitschriften-Peer-Review auf der Grundlage von John Stuart Mills (1859) Argumentation für die Redefreiheit kritisiert werden kann. Dieser Argumentation zufolge steigt mit wachsender Zahl der veröffentlichten Ideen auch die Wahrscheinlichkeit, dass die jeweils wahren Ideen darunter sind. Ein Vorfilter durch Experten wie beim Peer Review kann, weil auch Experten sich irren können, der Wahrheitsfindung nur hinderlich sein, und ist daher nicht nur überflüssig, sondern potenziell sogar dem Erreichen des Ziels der Wissenschaft entgegengesetzt. Welches die wirklich guten Ideen sind, wird sich in der Diskussion zeigen; aber es ist wichtig, dass möglichst alle Ideen überhaupt zur Diskussion kommen. Shatz hat aber meines Erachtens bei dieser Argumentation zu wenig berücksichtigt, dass heute praktisch jede wissenschaftliche Arbeit – selbst wenn sie von sehr fragwürdiger Qualität ist – in irgendeiner Form veröffentlicht wird, und sei es nur in weniger bekannten oder sogar obskuren Zeitschriften oder im Internet.

2 Wie groß dieser Einfluss ist, war immer wieder Gegenstand wissenschaftssoziologischer Kontroversen, zum Beispiel in der Folge des „strong program" (Bloor 1976; Shapin/Schaffer 1985) oder von Robert K. Mertons erstmals 1938 veröffentlichten, äußerst einflussreichen Dissertation (Merton 1970). Ich kann hier in Bezug auf diese Frage neutral bleiben.

Gerade im Zuge des Vormarschs des Internets zeigt sich aber, dass die Aufnahmekapazität des Menschen für Informationen beschränkt ist und es durchaus sinnvoll sein kann, die Fülle der verfügbaren Publikationen nach ihrer Relevanz für bestimmte Leserinnen und Leser vorzusortieren. Auf der Basis von Mills Argumenten für die Redefreiheit lässt sich also aufzeigen, dass die eigentliche Funktion des Zeitschriften-Peer-Review möglicherweise nicht etwa in der Qualitätssicherung der Wissenschaft, sondern vielmehr im Management von Lesezeit liegt (Harnad 1998), und damit vor allem der Lektüre von Arbeiten zugute kommt, die gerade *nicht* im engeren Forschungsfeld eines Wissenschaftlers liegen[3]. Dieses Ergebnis mag zwar überraschend sein, aber es impliziert nicht, dass Zeitschriften-Peer-Review unwichtig ist. Außerdem ist die auf Mills Argumenten basierende Argumentation auf das Grant Application Peer Review überhaupt nicht anwendbar.

Peer Review ist ohnehin nicht mehr die einzige Methode der engen Evaluation. In zunehmenden Maß werden zu diesem Zweck auch bibliometrische Methoden eingesetzt. Obwohl bibliometrische Evaluationen häufig nicht von beitragenden Experten durchgeführt werden, ist die Zielsetzung solcher Evaluationen in der Regel die Beurteilung des Erfolgs von Forschungsarbeiten nach den Maßstäben der jeweiligen wissenschaftlichen Disziplin. Deshalb muss sie zur engen Evaluation gezählt werden. Möglicherweise ist gerade die Tatsache, dass hier enge Evaluation durch nicht beitragende, interaktionale Experten (z. B. „science administrators") durchgeführt wird, eines der wissenschaftssoziologisch auffälligsten Merkmale der Bibliometrie. Wie gut empirische Indikatoren bei der Erfassung wissenschaftlicher Qualität wirklich sind, ist eine empirische Frage, die wohl zurzeit kaum abschließend beurteilt werden kann. Auf jeden Fall handelt es sich dabei aber nicht um ein wissenschaftstheoretisches Problem, das mit Mitteln der philosophischen Analyse gelöst werden kann. Aus diesem Grund werde ich es hier nicht weiter verfolgen.

Abgesehen vielleicht von der vorhin kurz andiskutierten Frage nach der Legitimität von Zeitschriften-Peer-Reviews, wirft die enge Evaluation eigentlich überhaupt keine neuartigen wissenschaftstheoretischen Fragen auf. Denn die Kriterien guter Wissenschaft waren schon immer Gegenstand der Wissenschaftstheorie. In besonderem Maß gilt dies natürlich für Fragen nach der Begründung und der Erklärungskraft wissenschaftlicher Theorien, die zusammen mit der Struktur von Theorien und der Dynamik des wissenschaftlichen Wandels im Zentrum klassischer wissenschaftstheoretischer Studien stehen (Stegmüller 1969-1986, Band 1-2). Die wissenschaftstheoretische Diskussion dieser Aspekte geht natürlich weiter; sie zu verfolgen ist hier aber nicht der geeignete Ort. Ich möchte mich vielmehr der Wissenschaftstheorie der weiten Evaluation zuwenden.

Im Gegensatz zur engen Evaluation sind die wissenschaftstheoretischen Grundlagen der weiten Evaluation bisher noch wenig untersucht worden, und es besteht keine Klarheit darüber, welches die dafür geeigneten Methoden sind. Es steht lediglich fest, dass Bibliometrie eine gänzlich ungeeignete Methode für die weite Evaluation ist,

3 Letztere Einsicht verdanke ich Daniel Sirtes.

denn der Zweck der Forschung besteht nicht darin, möglichst hohe Zitationszahlen zu erreichen. Moderne Staaten, private Stiftungen und Unternehmen unterstützen die Forschung, weil sie sich davon einen Gewinn erhoffen, sei es in Form von technologischen und anderen Innovationen, hochqualifizierten jungen Wissenschaftlerinnen und Wissenschaftlern sowie wissenschaftlichem Know-how, sei es in Form von Erkenntnis. Wissenschaftliche Exzellenz ist dabei zwar sicherlich etwas von den Geldgebern Angestrebtes, allerdings wohl nicht als Selbstzweck. Das lässt sich allein schon daran ablesen, dass die Allokation von Forschungsmitteln im riesigen Spektrum wissenschaftlicher Disziplinen oft schwierigen Verhandlungen unterliegt. Klar ist, dass solche Entscheidungen – die nach meiner Definition zur weiten Evaluation gehören – nicht beitragenden Experten überlassen werden können und ihnen auch nicht überlassen werden. Hier sind politische Entscheidungen unerlässlich. Doch welches ist die wissenschaftstheoretische Grundlage solcher Entscheidungen? Gibt es eine solche Grundlage überhaupt, und wenn ja: kann die Philosophie etwas dazu beitragen? Ich möchte in den folgenden beiden Abschnitten zeigen, dass es durchaus Aspekte dieses Problems gibt, die durch philosophische Überlegungen erhellt werden können. Ein solcher Aspekt ist die Axiologie (Werttheorie) wissenschaftlichen Wissens, die ich im folgenden Abschnitt behandeln werde.

2 *Zur Axiologie wissenschaftlichen Wissens*

Ich schlage zunächst zwei Adäquatheitsbedingungen für eine Axiologie der Wissenschaften vor. Die erste Bedingung sei die der *Inklusivität*. Damit meine ich, dass eine solche Axiologie auf alle Wissenschaften anwendbar sein soll, also auf Natur- ebenso wie auf Geistes-, Sozial- und Kulturwissenschaften. Dies bedeutet nicht, dass die Axiologie allen Wissensformen letztlich denselben Wert zuweisen muss, aber sie darf nicht von Beginn an bestimmte Wissenschaften bevorzugen, etwa die Naturwissenschaften. Die zweite Adäquatheitsbedingung ist der *Wertepluralismus*, also die hohe Variabilität der Werte, Interessen und Präferenzen in einer modernen Gesellschaft. Eine Axiologie wissenschaftlichen Wissens muss dem Rechnung tragen, weil sie sonst droht, nur eine elitäre oder elitezentrierte Legitimation für die Existenz von öffentlichen Universitäten, Forschungsinstituten und forschungsfördernden Institutionen bereitstellen zu können.

Die Frage ist nun, wie gut verschiedene Axiologien diese zwei Adäquatheitsbedingungen erfüllen. Es gibt zwei verschiedene Grundtypen: subjektive und objektive Werttheorien. Da die beiden Begriffe „subjektiv" und „objektiv" notorisch vieldeutig sind, muss kurz erläutert werden, wie sie hier zu verstehen sind. Subjektive und objektive Theorien unterscheiden sich vor allem darin, welchen Status sie Werturteilen zuweisen. Ein Werturteil hat die Form „x ist wertvoll" (qualitativ), oder „x hat Wert u" (quantitativ) oder „x ist wertvoller als y" (komparativ). Nach einer objektiven Werttheorie ist die Wahrheit oder Falschheit von Werturteilen unabhängig davon, ob je-

mand diese Wahrheit oder Falschheit faktisch anerkennt. Dagegen ist nach einer subjektiven Werttheorie die Anerkennung der Wahrheit oder Falschheit solcher Werturteile gerade für deren Wahrheit oder Falschheit konstitutiv. Mit anderen Worten, nach einer objektiven Werttheorie kann etwas einen Wert haben, selbst wenn niemand ein positives Werturteil darüber als wahr anerkennt. Nach einer subjektiven Werttheorie ist dies nicht der Fall; es muss nach einer solchen Theorie mindestens eine Person geben, die ein positives Werturteil als wahr anerkennt.

Die Werturteile, um die es hier geht, sind von der Art „*zu wissen, dass p* ist wertvoll" oder „*zu wissen, dass p* hat Wert u" oder „*zu wissen, dass p* ist wertvoller als *zu wissen, dass q*" (die kursiv gesetzten Ausdrücke zeigen die Träger des Werts an um Missverständnisse zu vermeiden), wobei p und q beliebige Aussagen sein können. In unserem Zusammenhang werden p und q wissenschaftliche Aussagen sein, also die Ergebnisse wissenschaftlicher Forschung, wie zum Beispiel „das Produkt des Tumorsuppressor-Gens BRCA1 ist an der Reparatur der DNA beteiligt" oder „alle menschlichen Sprachen verfügen über eine universale generative Grammatik", um zwei Beispiele aus ganz unterschiedlichen Wissenschaftszweigen zu nennen. Wir sprechen von Wissen, wenn jemand solche Aussagen kennt und gute Gründe für die Annahme hat, dass sie wahr seien[4]. *Warum* jemand es wertvoll finden mag, diese Aussagen zu kennen und über Gründe zu verfügen, dass sie wahr seien, bleibt hier völlig offen. Es geht vorerst nur um die Frage, worin die Geltung von Werturteilen über den Besitz von Wissen gegründet ist. Für den Wertsubjektivisten sind sie darin gegründet, dass jemand diese Urteile faktisch anerkennt, und zwar unabhängig davon, ob und wie *dieses* Urteil begründet ist. Dagegen glaubt der Objektivist, dass die Geltung solcher Werturteile nicht davon abhängt, ob jemand sie faktisch anerkennt oder nicht.

Einige potenzielle Missverständnisse im Zusammenhang mit objektiven Werttheorien müssen hier gleich ausgeräumt werden: Erstens, der Objektivismus muss Wissen nicht als *intrinsisch* wertvoll betrachten, das heißt den Besitz von Wissen per se für wertvoll halten. Wissen kann auch aus objektivistischer Sicht als Mittel zu anderen Zwecken verstanden werden. Solange die Werturteile über diese anderen Zwecke objektiv sind, können es auch Werturteile über Wissen sein, das nicht in sich von Wert sein muss. Zweitens, der Objektivist muss nicht auf eine universale menschliche Natur rekurrieren, wie zum Beispiel eine angeborene Neugier. Es kann sein, dass der Wert des Wissens nicht von bestimmten natürlichen Neigungen abhängt, sondern von kulturell vermittelten Bedürfnissen und Zielen. Dies tut der Objektivität der Werturteile keinen Abbruch, kann aber bedeuten, dass diese einem historischen Wandel unterliegen.

4 Ich verwende hier die philosophische Standarddefinition von Wissen, nach der Wissen mehr ist als bloße Information. Wissen muss begründet sein, und wenn wir von „wissenschaftlichem Wissen" sprechen, so verbinden wir damit einen besonders hohen Begründungsstandard. Die Frage, welche Aussagen diesen Begründungsstandard erreichen, gehört in den Bereich der engen Evaluation. Hier befassen wir uns ausschließlich mit der Frage, worin der Wert besteht, über bestimmte Arten von Wissen zu verfügen, und zwar nicht aus der Sicht des Wertekanons einer wissenschaftlichen Disziplin, sondern aus der Sicht der Allgemeinheit.

Wie könnte eine objektive Werttheorie des Wissens aussehen? Ich sehe grundsätzlich zwei verschiedene Varianten. In der ersten Variante wird der Wert des Wissens darin begründet, dass es Mittel zu bestimmten Zwecken ist, zum Beispiel Gesundheit, Nachhaltigkeit, Entwicklung, aber auch Gerechtigkeit oder politische Stabilität usw.[5] Kann eine solche Konzeption die eingangs aufgestellten Adäquatheitsbedingungen erfüllen? Es gibt folgendes Problem damit: Um die zweite Adäquatheitsbedingung, Verträglichkeit mit dem Wertepluralismus, zu erfüllen, werden nur solche Zwecke als wertkonstitutiv gelten können, bei denen man davon ausgehen kann, dass ihre Verfolgung im Interesse *aller* liegt, und zwar unabhängig vom kulturellen Hintergrund. Das heißt, die Forschung muss Bedürfnisse bedienen, die alle Menschen haben. Dies mag bei manchen Zwecken der Fall sein, etwa bei der Gesundheit. Fast alle Menschen wollen gesund sein. Ein potenzielles Problem dabei ist, dass es in einer pluralistischen Gesellschaft verschiedene Vorstellungen davon geben kann, was es heißt, gesund zu sein. Ein hartgesottener Wertobjektivist wird dagegen argumentieren, dass es keine Rolle spielt, was verschiedene Menschen über die Gesundheit denken; Gesundheit im Sinne einer evidenzbasierten Medizin ist gut für alle. Auf ähnliche Weise könnte man versuchen, andere universale Zwecke zu verteidigen, wie etwa Gerechtigkeit.

Im Zusammenhang mit unserer Thematik ergibt sich jedoch dabei ein Problem. Selbst wenn es gelingen sollte (was durchaus fraglich ist), Dinge wie Gesundheit oder Gerechtigkeit als Ziele zu verteidigen, die im objektiven Interesse *aller* Menschen liegen, wird eine darauf aufbauende Axiologie des Wissens Probleme hinsichtlich der Inklusivität bekommen, also mit der Forderung, dass eine Werttheorie potenziell für alle Wissenschaften Geltung haben soll. Selbst wenn anerkannt wird, dass Biologie und Psychologie zur Gesundheit oder dass Philologie, Soziologie, Geschichte und Gender Studies zu einer gerechteren Gesellschaft, zu interkultureller Verständigung (und damit zur politischen Stabilität) beitragen – welchen *universalen* Bedürfnissen dienen etwa die Evolutionstheorie, die Urknall-Kosmologie oder die Paläoanthropologie? Neuen Umfragen zufolge wird zum Beispiel die Evolutionstheorie in vielen Ländern von großen Teilen der Bevölkerung abgelehnt (siehe Miller/Scott/Okamoto 2006). Bei diesem Typ von Axiologie scheint also das Problem zu bestehen, dass der Versuch, die zweite Adäquatheitsbedingung zu erfüllen, indem nur universale Zwecke als wertkonstitutiv anerkannt werden, sofort zu Schwierigkeiten mit der ersten Bedingung führt, der Inklusivität.[6]

Die zweite Variante einer objektiven Werttheorie hält Wissen für eine Voraussetzung dafür, dass wir *überhaupt* Dinge wertschätzen können. Die Bewertung von irgendwelchen Gütern, seien diese materieller oder immaterieller Natur, setzt nach dieser Theorie immer voraus, dass wir über das erforderliche Wissen verfügen, um die

5 Die von der UNESCO/ICSU World Conference on Science 1999 verabschiedete *Declaration on Science* bekräftigt etwa den Nutzen wissenschaftlicher Forschung zu diesen und weiteren Zwecken, darunter auch intellektuelle Bereicherung. Siehe hierzu auch Hoyningen-Huene/Oberheim/Weber (1998), Hoyningen-Huene/Weber/Oberheim (1999).

6 Zu beachten ist auch, dass die Anrufung eines Zwecks wie Bildung die Frage lediglich verlagert, denn man muss dann angeben können, worin der Wert von Bildung gegründet ist.

Interrelationen zwischen den Dingen und anderen Dingen, etwa solchen, die als Mittel zum Erlangen dieser ersten Sorte von Dingen in Betracht kommen, verstehen zu können. Ohne Wissen kann es daher gar nichts geben, das irgendeinen Wert hätte; Wissen ist eine Voraussetzung für die Möglichkeit von Werten überhaupt. Dieses Argument ist zwar philosophisch sehr elegant und schneidet bei der zweiten Adäquatheitsbedingung sehr gut ab (was immer man wertvoll findet, es ist besser, Wissen zu haben), jedoch stößt es ebenfalls auf Schwierigkeiten mit der ersten Bedingung, der Inklusivität. Denn wir brauchen wohl kaum die Urknall-Kosmologie, um Dinge wertschätzen zu können. Diese Argumentation wurde denn auch in einem etwas anderen Kontext entwickelt, nämlich, um die Quelle epistemischer Normativität herauszuarbeiten (Kornblith 2002). Sie gibt eine gute Antwort auf die Frage, woher epistemische Normen, wie zum Beispiel die Norm der Widerspruchsfreiheit, ihre präskriptive Kraft beziehen. Diese epistemischen Normen sind aber sehr allgemein und an keine spezifischen Wissensformen geknüpft. Eine solche Theorie kann höchstens die Frage beantworten, worin der Wert eines funktionierenden kognitiven Systems und der Befolgung logischer Regeln besteht und nicht, worin der Wert spezifischer Wissens*inhalte* liegt.

Angesichts dieser Schwierigkeiten ist es mehr als fraglich, ob eine objektive Werttheorie des Wissens angemessen sein kann. Ich gehe deshalb zur Diskussion der subjektiven Theorien über.

Für einen Wertsubjektivisten erhält ein Ding nur dadurch einen Wert, dass jemand es als wertvoll anerkennt. Dabei spielt es keine Rolle, aus welchen Gründen diese Anerkennung erfolgt; die bloße Tatsache, *dass* jemand etwas anerkennt, reicht aus. Während es für Objektivisten Ziele gibt, die es wert sind, verfolgt zu werden, selbst wenn niemand dies tatsächlich tun will, ist eine subjektive Werttheorie an die subjektiven Präferenzen tatsächlicher Personen gebunden. Dabei muss es für diese Präferenzen selbst keinerlei Rechfertigung geben; es kann sich auch um völlig willkürliche Vorlieben handeln. Für eine Axiologie des Wissens hat dies den Vorzug, dass jede Wissenschaft, so exotisch sie auch sein mag, einen Wert haben kann, solange es Individuen gibt, deren Präferenzen sie bedient. Dazu kommt noch, dass auch Präferenzen wie die bloße Neugier oder Lust am Wissen wertkonstitutiv sind. Aus diesem Grund hat eine subjektive Axiologie des Wissens meines Erachtens weniger Schwierigkeiten, die erste Adäquatheitsbedingung (Inklusivität) zu erfüllen. Tendenziell hat sie natürlich ein Problem mit der zweiten Bedingung, der Verträglichkeit mit dem Wertepluralismus. Denn Wertepluralismus bedeutet ja, dass höchstens ein Teil der Mitglieder einer Gesellschaft über die entsprechenden Präferenzen verfügt, die wissenschaftliches Wissen wertvoll machen können. Nicht alle haben schließlich die motivationale Verfassung eines Alexander von Humboldt. Gleichzeitig muss anerkannt werden, dass die Berücksichtigung der Präferenzen *aller* bei der Verfolgung gemeinsamer Großprojekte und bei der Allokation staatlicher Ressourcen ein Problem ist, das jede pluralistische Gesellschaft lösen *muss*. Hier sitzt die Wissenschaft im gleichen Boot wie etwa die staatlich geförderte Kunst oder eine Stadtbildkommission.