fteval JOURNAL

for Research and Technology Policy Evaluation

Open Issue 2024

#55

FEDERAL MINISTRY OF EDUCATION, SCIENCE AND RESEARCH
Minoritenplatz 5, 1014 Vienna

Dr. Patrick Svensson-Jajko, E: Patrick.Svensson-Jajko@bmbwf.gv.at
Mag.ª Simone Mesner, E: simone.mesner@bmbwf.gv.at

FEDERAL MINISTRY FOR LABOUR AND ECONOMY
Stubenring 1, 1014 Vienna

Mag. Bernd Zimmer, bernd.zimmer@bmaw.gv.at

FEDERAL MINISTRY FOR CLIMATE ACTION, ENVIRONMENT, ENERGY, MOBILITY, INNOVATION AND TECHNOLOGY
Radetzkystraße 2, 1030 Vienna

Dr. Rupert Pichler, E: rupert.pichler@bmk.gv.at
MSc (WU) Philipp Witibschlager, BA, E: philipp.witibschlager@bmk.gv.at
Lisa Koch, BBA MSc, Kmsr., E: lisa.koch@bmk.gv.at

ACR – AUSTRIAN COOPERATIVE RESEARCH
Sensengasse 3, 1010 Vienna

Dr.in Sonja Sheikh, E: sheikh@acr.ac.at

AIT – AUSTRIAN INSTITUTE OF TECHNOLOGY
Giefinggasse 4, 1210 Vienna

Mag. Michael Dinges, E: michael.dinges@ait.ac.at
Anahi Montalvo-Rojo, E: anahi.montalvo-rojo@ait.ac.at

AQ AUSTRIA – AGENCY FOR QUALITY ASSURANCE AND ACCREDITATION AUSTRIA
Franz-Klein-Gasse 5, 1190 Vienna

Dr.in Elisabeth Froschauer-Neuhauser , E: elisabeth.froschauer-neuhauser@aq.ac.at
Dr.in Eva Maria Freiberger, E: eva.maria.freiberger@aq.ac.at

AWS – AUSTRIA WIRTSCHAFTSSERVICE GMBH
Walcherstraße 11A, 1020 Vienna

Mag. Norbert Knoll, E: n.knoll@awsg.at

CDG – CHRISTIAN DOPPLER RESEARCH ASSOCIATION
Boltzmanngasse 20, 1090 Vienna

DIin Magª. Brigitte Müller, E: brigitte.mueller@cdg.ac.at

CONVELOP – COOPERATIVE KNOWLEDGE DESIGN GMBH
Kaiserfeldgasse 7/II, 8010 Graz

DIin Dr.in Karin Grasenick, E: karin.grasenick@convelop.at
Konstantin Melidis, E: konstantin.melidis@convelop.at
Erdbergstraße 82/4, 1030 Wien

FFG – AUSTRIAN RESEARCH PROMOTION AGENCY
Haus der Forschung, Sensengasse 1, 1090 Vienna

DIin Dr.in Sabine Mayer, E: sabine.mayer@ffg.at
Mag. Leonhard Jörg, E: leonhard.joerg@ffg.at
Dr. Jakob Eder, MA, E: jakob.eder@ffg.at

FH CAMPUS WIEN
Favoritenstraße 226, 1100 Wien

Dr. Mario Steyer, E: mario.steyer@fh-campuswien.ac.at

FORWIT - AUSTRIAN COUNCIL FOR SCIENCES, TECHNOLOGY, AND INNOVATION
Pestalozzigasse 4/DG 1, 1010 Vienna

Priv.-Doz.in DIin Mag.ª Dr.in techn. Alexandra Mazak-Huemer, E: amh@rfte.at

FWF – AUSTRIAN SCIENCE FUND
Georg-Coch-Platz 2, 1010 Wien

Dr.in Tina Olteanu, E: tina.olteanu@fwf.ac.at
Dr. Falk Reckling, E: falk.reckling@fwf.ac.at

IHS – INSTITUTE FOR ADVANCED STUDIES
Josefstädter Straße 39, 1080 Vienna
Dr.in Angela Wroblewski, E: wroblews@ihs.ac.at

INDUSTRIEWISSENSCHAFTLICHES INSTITUT – IWI
Mittersteig 10, 1050 Wien
FH-Hon.Prof. Dr. Dr. Herwig W. Schneider, E: schneider@iwi.ac.at
Mag. Philipp Brunner, E: brunner@iwi.ac.at

JOANNEUM RESEARCH FORSCHUNGSGESELLSCHAFT MBH
Sensengasse 3/3/1, 1090 Wien
Michael Ploder, E: Michael.Ploder@joanneum.at
Mag. Jürgen Streicher, E: juergen.streicher@joanneum.at

KMU FORSCHUNG AUSTRIA –AUSTRIAN INSTITUTE FOR SME RESEARCH
Gusshausstraße 8, 1040 Vienna
Mag. Peter Kaufmann, E: p.kaufmann@kmuforschung.ac.at
Harald Wieser, PhD, E: h.wieser@kmuforschung.ac.at
Jakob Kofler, MSc MPA, E: j.kofler@kmuforschung.ac.at

LUDWIG BOLTZMANN GESELLSCHAFT
Nußdorfer Straße 64, 1090 Vienna
Elvira Welzig, E: elvira.welzig@lbg.ac.at
Sherin Alias, E: Sherin.Alias@lbg.ac.at
Georg Russegger, E: georg.russegger@lbg.ac.at

OEAW – AUSTRIAN ACADEMY OF SCIENCE
Dr. Ignaz Seipel-Platz 2, 1010 Vienna
Nikolaus Göth, MSc, E: nikolaus.goeth@oeaw.ac.at

TECHNOPOLIS GROUP AUSTRIA
Rudolfsplatz 12/11, 1010 Vienna
Mag.a Katharina Warta, E: warta@technopolis-group.com
Tobias Dudenbostel, MA, E: tobias.dudenbostel@technopolis-group.com

VIENNA BUSINESS AGENCY.
A SERVICE OFFERED BY THE CITY OF VIENNA.
Mariahilfer Straße 20, 1070 Vienna
DI Mag. Alfried Braumann, E: braumann@wirtschaftsagentur.at

WIFO – AUSTRIAN INSTITUTE OF ECONOMIC RESEARCH
Arsenal, Objekt 20, Postfach 91, 1103 Vienna
Dr. Jürgen Janger, E: juergen.janger@wifo.ac.at
Mag.a Dr.in Agnes Kügler, E: agnes.kuegler@wifo.ac.at

WPZ RESEARCH GMBH
Mariahilfer Straße 115/16, 1060 Vienna
Dr.in Brigitte Ecker, E: brigitte.ecker@wpz-research.com
Prof.(FH) Mag.a Dr.in Verena Régent, E: verena.regent@wpz-research.com

WWTF – VIENNA SCIENCE AND TECHNOLOGY FUND
Schlickgasse 3/12, 1090 Vienna
Dr. Michael Stampfer, E: michael.stampfer@wwtf.at
Dr. Michael Strassnig, E: michael.strassnig@wwtf.at

ZSI – CENTRE FOR SOCIAL INNOVATION
Linke Wienzeile 246, 1150 Vienna
Dr. Klaus Schuch, E: schuch@zsi.at
Stefanie Schuerz, MA, E: schuerz@zsi.at

fteval JOURNAL
for Research and
Technology Policy Evaluation

EDITORIAL

DEAR READERS,

If you're wondering why you cannot hold this issue in your hands anymore, the answer is simple: we have gone fully digital. And honestly, what issue could be more fitting to mark this shift than one treating challenges in evaluation and R&I, artificial intelligence and transformation?

Against this background we are happy to share our small digital revolution with you: Going digital allows us to better harness the advantages of virtual publishing. As you may already have noticed that articles are now released on a rolling basis, meaning each contribution is published as soon as it is ready. That is why the first articles in this issue appeared as early as 2023, while the final piece was only added in 2025. We aim to further improve publication speed and continue exploring the potential of digital tools: not just in how we publish, but also in how we review, share, and access evaluation research.

As mentioned, this issue brings together a wide variety of thought-provoking articles, which we briefly introduce below:

We open this issue with a contribution that addresses the evolving demands placed on evaluation in the context of grand societal challenges.

In ***Realising Societal Challenges: Towards Challenge-Led Monitoring and Evaluation***, Vincent Baarslag, Amber Geurts and Frans van der Zee explore how evaluation practices must evolve to effectively support innovation policies aimed at tackling major societal challenges. Based on a literature review, the authors identify key bottlenecks that hinder meaningful monitoring and evaluation in this context and argue that these cannot be addressed without rethinking governance structures and ingrained institutional practices. They propose a "challenge-led" approach to evaluation that moves beyond technical adjustments and calls for systemic change in how we steer research and innovation toward societal goals.

The following two articles turn to gender equality as a central dimension of evaluation quality—highlighting both strategic blind spots and opportunities for more inclusive practices.

In ***Monitoring as a Steering Tool for Gender Equality Policy in Higher Education Institutions: Ideal and Reality***, Angela Wroblewski examines how monitoring practices in gender equality policy often fall short of their intended steering function. While policy design ideally builds on evidence-based gender analysis and sets measurable objectives, Wroblewski shows that in practice, the development of indicators is frequently driven by data availability rather than policy goals. Using the Austrian higher education sector as a case, she reflects on the challenges of aligning monitoring tools with strategic objectives and highlights the structural limitations that hinder effective policy steering.

In ***Enhancing Evaluation Tender Quality: Integrating Gender for Robust Results***, Karin Grasenick highlights a critical blind spot in evaluation practice: namely, the lack of systematic gender and diversity considerations in tendering processes. While evaluation standards typically focus on the work of evaluators, the article argues that quality must also begin upstream, in the design of calls for proposals. Drawing on a guideline developed by the German Evaluation Society's Working Group on Gender Mainstreaming, Grasenick proposes practical ways to embed gender and intersectionality from the outset. The goal: more inclusive, legally sound, and socially responsible evaluations particularly in the context of research, technology, and innovation policy.

The next three articles are the result of dedicated working groups initiated by the fteval Platform, addressing two pressing themes in current evaluation discourse: artificial intelligence and indicators for sustainability. They reflect collective efforts to better understand and shape the role of evaluation in times of technological disruption and transformative societal change.

In ***The Intersection of AI and RTI (Policy) Evaluation: Principles and Considerations***, Stefanie Schuerz, Charlotte D'Elloy and Michael Dinges present key reflections and practical guidance on the use of artificial intelligence in evaluation processes. As part of a working group, the authors explored the applicability of AI tools across different phases of the evaluation cycle. Their contribution outlines ethical principles, practical use cases, and critical limitations of AI in this context, emphasizing the need for transparency, accountability, and human oversight. The paper offers a valuable starting point for evaluators navigating the rapidly evolving intersection between digital technologies and evaluation practices.

In ***How Does Artificial Intelligence Affect the Evaluation System? Discussion Starters for Shaping the Evaluation System of Tomorrow***, Thomas Palfinger, Felix Gaisbauer, Isabella Wagner and Susanne Beck explore the systemic

impact of generative AI on evaluation practices – using research and innovation policy evaluation as a case in point. Moving beyond familiar concerns such as data protection, ethics, and methodological reliability, the authors shift the focus to the relational dynamics within evaluation systems. Based on the work of a dedicated fteval working group, they present a conceptual model that helps reflect on how generative AI might disrupt established actor constellations and interaction patterns. While rooted in the Austrian context, the model is designed to stimulate reflection in similarly structured systems elsewhere.

In ***More Sustainable, Faster, Deeper, Broader: Capturing RTI Contributions to Sustainable Transformation Processes***, Jakob Kofler, Isabella Wagner and Harald Wieser present the findings of a working group that explored how monitoring and evaluation can better capture the role of research, technology and innovation (RTI) in driving sustainability transitions. Reviewing both Austrian practices and literature, the authors identify conceptual and methodological challenges in tracing transformative impact. They propose avenues for expanding existing indicator frameworks and stress the importance of embedding transformation logic more explicitly in monitoring design. Their contribution lays the groundwork for more robust and future-oriented evaluation approaches in the sustainability context.

Building on the theme of transformative societal change, the article that follows turns to the often-overlooked role of social innovation. It explores how intentional shifts in social practices can complement technological approaches in driving systemic transformation.

In ***Theoretical Reflections on the Role of Social Innovation in Challenge- and Goal-Oriented R&I Agendas***, Klaus Schuch makes a case for integrating social innovation more systematically into transformative research and innovation programmes. Arguing that technological solutions alone are insufficient to address today's grand societal challenges, he calls for a conceptual expansion of current innovation policy frameworks. Drawing on practice theory, the paper defines social innovation as intentional changes in social practices and explores how such innovations can contribute to systemic transformation. Schuch positions social innovation within the multi-level perspective on socio-technical change and offers guidance on how policy and programme design must evolve to realise its transformative potential.

Fittingly, the final article in this issue returns to the topic of artificial intelligence—this time from a comparative governance perspective. By examining how Austria and Australia approach the regulation and strategic

use of AI, it offers a broader view on how states navigate the balance between innovation, ethics, and control in the face of rapid technological change.

In ***AI Governance in Austria and Australia: Learning from the Antipodeans***, Peter Biegelbauer explores how two high-income democracies approach the complex task of governing artificial intelligence. Comparing Austria and Australia, the article examines how each country balances innovation, ethics, and regulation - both in terms of national strategy and in the state's own use of AI. Despite shared ambitions, the two cases reveal notable differences in governance structures, policy instruments, and implementation approaches. Biegelbauer's comparative lens offers valuable insights into the diverse pathways governments can take when navigating the promises and risks of AI.

We hope these contributions will not only inform but also inspire—by highlighting the many ways in which evaluation can engage with the major shifts shaping research and innovation policy today. Whether it's artificial intelligence, gender equity, sustainability, or governance challenges across continents, the articles in this issue share a common concern: how to design evaluation systems that are both forward-looking and fit for purpose.

Our next open issue is scheduled for 2026, and submissions are already welcome.

Until then, enjoy reading—and rethinking—what evaluation can be.

Yours,
Isabella Wagner

fteval JOURNAL
for Research and
Technology Policy Evaluation

January 2024, Vol. 55, pp. e1, 1-22
DOI: 10.22163/fteval.2024.632

REALISING SOCIETAL CHALLENGES: TOWARDS CHALLENGE-LED MONITORING AND EVALUATION

VINCENT BAARSLAG, AMBER GEURTS & FRANS VAN DER ZEE
DOI: 10.22163/FTEVAL.2024.632

ABSTRACT

Challenge-led innovation policies place societal challenges and transitions at the focal centre and goal of innovation policy. This new genre of innovation policy not only requires new approaches in agenda-setting, programming, implementation and management, but also requires a renewed view and practice of **monitoring and evaluation** in order to be able to steer innovation policies towards societal goals. In this paper, we focus on the necessity and usefulness of a different view and way of monitoring and evaluating challenge-led R&I policy and its implementation – which we term **challenge-led monitoring and evaluation**. To define this new approach, we conducted a literature review. Our analysis identifies the bottlenecks as well as potential routes to arrive at an appropriate monitoring and evaluation framework for challenge-led innovation policies. Next, our analysis highlights how governance and institutionalised evaluation culture and practice are presented as (part of) solutions to all four identified bottlenecks but usually considered an afterthought requiring 'experimentation'. However, overcoming other identified bottlenecks in challenge-led monitoring and evaluation is contingent upon altering (a) governance and organisational structures as well as (b) institutionalised assumptions and practices. Therefore, we argue that in order to further develop challenge-led monitoring and evaluation, the roles of governance and organisational structures as well as institutionalised assumptions and practices should be problematised and prioritised as bottlenecks.

Keywords: challenge-led innovation policy, transformative innovation policy, mission-oriented innovation policy, monitoring, evaluation, transitions, literature review, accountability, learning, operationalisation, attribution.

1. INTRODUCTION

Innovation policy in the Netherlands and Europe has been developing rapidly in recent years. Rather than only targeting economic growth and productivity, we see an increasing recognition – both in policy practice as well as in science - for a new genre of innovation policy that aims to focus research and innovation more strongly on addressing persistent societal challenges (Kuhlman & Rip, 2018; Schot & Steinmueller, 2018). Consequently, previous innovation policy frameworks, which focused on fixing market failures by repairing private firms' underinvestment in research and development (frame 1) or on fixing system failures in national innovation systems and strengthen innovation networks (frame 2), are being complemented with a new generation of innovation policies to address and find solutions for societal challenges (frame 3, Schot & Steinmueller, 2018). Based on different emerging streams of literatures, these frame 3 innovation policies are framed as 'mission-oriented innovation policy (Mazzucato, 2018; Hekkert et al., 2020) or 'transformative innovation policy' (Diercks et al., 2019; Haddad et al., 2022). In this paper, we place these different perspectives under a common umbrella that we call 'challenge-led innovation policy (CIP)' (Rathenau Instituut, 2020; 2021).

This new genre of challenge-led innovation policy not only requires new approaches in agenda-setting, programming, implementation and management (Janssen et al., 2021). In light of this new type of policy, we argue that **monitoring and evaluation** also requires an innovative approach, which is different on a number of dimensions from the usual, 'traditional' monitoring and evaluation of research and innovation policies. **Challenge-led monitoring and evaluation** is meant to effectively use the potential and increase the effectiveness of utilising research and innovation in solving societal challenges and realising system transitions.

Monitoring and evaluation is important in addressing societal challenges and transitions (see e.g. Janssen, 2019; Luederitz et al., 2017; Turnheim et al., 2015; Weber et al., 2014; Wittmann et al., 2022). Monitoring and evaluation is a key element in the realisation of a balanced and reliable evidence base that enables politics and public administrations to conduct an informed debate and to make targeted assessments about the deployment, progress, and outcomes of poli-

cies for societal challenges and transitions. Monitoring and evaluation provide the evidence to enable political and administrative accountability and hence fulfil a vital democratic function. At the same time, monitoring and evaluation also enable politics and public administration to maintain or, where necessary, shift policy direction by learning and offering room for improvement and adjustment. Learning and adjustment can mean intensifying, accelerating, but also phasing out or even stopping particular research and innovation (R&I) efforts, thus being able to even better steer mission and transition processes towards the envisioned societal goal. Monitoring and evaluation can thus provide knowledge bases and policy intelligence to inform system-level transformative policies for R&I now and in the future.

It also becomes increasingly clear that monitoring and evaluation of challenge-led R&I policies should differ in important respects from the usual, more traditional evaluation of R&I policies as it sets different design requirements in terms of the evaluation perspective, framework and approaches to be used. Furthermore, as the limitations of traditional monitoring and evaluation frameworks are becoming increasingly evident, calls for reflection, reconceptualization and experimentation with altogether different monitoring and evaluation practices increases (Arnold et al., 2018; Haddad et al., 2022; Molas-Gallart et al., 2021; Wittmann et al., 2022). However, the complexity and wickedness of societal challenges impose various conceptual challenges for such a challenge-led monitoring and evaluation, including '**the multidimensionality and interaction of effects, the different analytical levels, the long time horizon associated with mission goals and the empirical diversity of missions**' (Wittmann et al., 2022, p.31).

While there have been numerous attempts to initiate and substantiate the discussion of what it takes to develop a monitoring and evaluation framework that can be considered 'challenge-led', there has been no convincing synthesis, yet, of what this innovative view and role means for monitoring and evaluation, let alone how it could be further shaped and implemented. With the increasing importance of challenge-led R&I in policy and society, whereby the bottlenecks, potential pitfalls and possible solutions of challenge-led monitoring and evaluation have until now been underexposed, such a synthesis would fulfil a societal need and serve a public interest. Renewal of monitoring and evaluation practices, but also a government that learns and adjusts more consistently and transparently on the basis of (continuous) monitoring and evaluation is necessary in view of the pressing societal urgency of solving various societal challenges. Expectations regarding the role of R&I in solving societal challenges,

and the associated use of public funds accentuate this necessity. In this article, we therefore focus on the emergence of a different perspective on the monitoring and evaluation of challenge-led R&I policy. The research question central to this article thus reads: **"How does the emergence of challenge-led research and innovation policies impact requirements for monitoring and evaluation, and how can challenge-led monitoring and evaluation strengthen policy and governance of research and innovation aimed at societal challenges?"**

We address our main research question using a systematic literature review. In the next sections, we describe our methodology for the literature review and present our main findings. We conclude our analysis with a discussion of the subsequent implications and steps to be taken as from here.

2. METHODOLOGY

In order to address our central research question we conducted an in-depth literature review (Elsbach & van Knippenberg, 2020; Torraco, 2016) focusing on generating insights as well as identifying active debates regarding why and how challenge-led monitoring and evaluation is different from 'traditional' monitoring and evaluation and what is needed to implement challenge-led monitoring and evaluation practices. This review process consists of several steps, including the search for relevant articles, their selection , and their analysis. To complement our literature review, we have also conducted semi-structured interviews with civil servants and evaluation experts at different policy settings and levels, involved with different societal challenges. These interviews helped us to identify the main bottlenecks and crystallise our findings further.

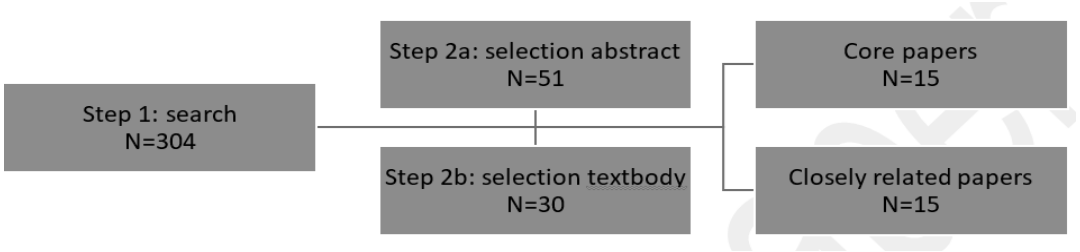
SEARCH AND SELECTION

To arrive at a set of relevant articles, we conducted a Web of Science article search from 2004 -2022. Our search included search terms in which 'innovation policy' or 'innovation programme' was combined with 'evaluation', 'assessment' or 'monitoring' concepts. This approach means that any governance or government capacities notions around monitoring and evaluation were excluded from the search process. This decision was made to focus on what challenge-led monitoring means, rather than whether government has the capacity to do so. In addition, we included search terms regarding 'sustainability', 'transition', 'transformation', 'missions', 'SDG' or 'societal challenge' in combination with 'evaluation', 'assessment' or 'monitoring' concepts. We used several

variations of these keywords to increase the number of potentially relevant search results. Finally, to augment the articles identified during our literature review of academic peer reviewed papers and to correct possible omission bias we also searched Google for grey literature and other articles that might be useful. This resulted in a total of 304 articles and book chapters.

The main selection process involved two rounds. In the first round, we primarily judged articles based on the title, the abstract and keywords. This resulted in a key literature list of 51 articles. In the second round, we inspected the full texts of articles in order to check whether and how the main constructs we are interested in have been mentioned in the body of the text. In case we were unsure about an article, two of our authors would debate the relevance of the article. As a result of this selection process, several articles were excluded for the final analysis step. After the second step, we ended up with 30 relevant articles, out of which 15 articles discussed core issues to our research questions, while 15 other articles discussed closely related issues. A summary of our search and selection process is shown in Figure 1.

Figure 1: Flowchart article selection



3. ANALYSIS

The analysis of our selected sample of core and closely related papers was conducted in two rounds. First, we analysed the core texts to identify bottlenecks and solutions that are discussed, described or highlighted in the papers. In this step, we largely relied on the text and wording used in the papers and solely identified a difference between 1) the discussion of a bottleneck or 2) the discussion of a solution. In the second step, we analysed the identified bottlenecks to define aggregate dimensions. In this step we identified four dimensions of bottlenecks, namely 'operationalisation', 'attribution', 'accountability' and 'learning'. Next, we also analysed the identified solutions and found six dimensions of proposed solutions, namely 'operationalisation', 'attribution', 'ac-

countability', 'learning', 'governance and organisation', and 'institutional rigidity'. We refined our analysis by revisiting the literature and our analyses frequently, and corroborating our findings using semi-structured interviews with experts. In the case we were unsure about a dimension, the authors would debate the dimensions and its characteristics.

4. FINDINGS

Our analysis shows that with the emergence of the new genre of challenge-led innovation policy, the anticipated scope of monitoring and evaluation has also changed. As such, challenge-led monitoring and evaluation focuses not only on the result (the innovation outcome), but also on the question of how that outcome relates to the desired impact (i.e. addressing the societal challenge) and whether ongoing (intermediate) innovation results are still in line with and steering towards that desired impact. That is, societal challenges and the context in which they are shaped are often complex, with goals that cover a longer term, whereby the solution directions can sometimes also shift and change over time. The complexity and systemic character of societal challenges, stemming from their wicked nature, have implications for monitoring and evaluation. These implications are discussed next.

4.1. IDENTIFYING BOTTLENECKS OF CHALLENGE-LED EVALUATION

By analysing the relevant literature, four important bottlenecks for challenge-led monitoring evaluation can be demarcated. These bottlenecks concern 1) accountability, 2) learning, 3) operationalisation, and 4) attribution. Though we acknowledge that these bottlenecks can also be observed in some of the 'traditional' R&I policy evaluations, our analysis shows that the wicked character of societal challenges that challenge-led innovation policy aims at, exacerbates these bottlenecks (Duckett et al., 2016; Wanzenböck et al., 2020). That is to say, the four demarcated bottlenecks are not unique to challenge-led monitoring and evaluation, but manifest themselves differently and more extensively when compared to a 'traditional' innovation policy monitoring and evaluation context.

4.2. CATEGORISING BOTTLENECKS: UNDERSTANDING THE CHANGE DIMENSION

When further analysing these bottlenecks it was found that they can be placed into two different categories: (**1**) bottlenecks related to a fundamental shift in the role of R&I policy evaluation within a challenge-led innovation policy (CIP) context, and (**2**) bottlenecks related to the operationalisation and attribution of transformative impact in a CIP context. The two bottlenecks related to **accountability** and **learning** originate from a fundamental shift in the role of monitoring and evaluation within a CIP context: the socio-economic and environmental legitimations of CIPs, as well as the increased emphasis on reflexivity and adaptivity alter the evaluation perspective, imposing additional requirements on monitoring and evaluation of CIPs in terms of providing accountability and facilitating learning. In contrast, identified bottlenecks related to **operationalisation** and **attribution** in large part preceded the emergence of CIPs and are merely magnified by the wicked context that characterises CIPs. Figure 2 illustrates the identified bottlenecks and their categorization. Next, the aforementioned bottlenecks and their categorisation are further explained.

Figure 2: Four bottlenecks of challenge-led monitoring and evaluation

Cause:	Bottleneck:	Why?:
Fundamental shift in the role of evaluation within a CIP context	Accountability	IP now (partially) accountable for adequately addressing societal challenges through supported transition pathways
	Learning	Increased need for learning due to reflexivity and adaptivity being emphasised in CIP frameworks
Wicked nature of CIP context	Operationalization	Measuring additionality increasingly difficult due to wickedness of challenges, long timespan
	Attribution	Attributing results increasingly difficult due to complexity of policy mixes, need for outcome-based evaluation

Note: IP is short for Innovation Policy, CIP is short for Challenge-led Innovation Policy

4.2.1 A FUNDAMENTAL SHIFT IN THE ROLE OF EVALUATION WITHIN A CIP CONTEXT

The change in policy goals and legitimization towards addressing societal challenges constitutes a fundamental shift in the role of innovation policy (Schot & Steinmueller, 2018). In contrast to innovation policy frameworks for economic growth and competitiveness, CIPs no longer primarily aim to enhance economic growth and increase productivity through stimulating technological progress and enhancing national systems of innovation (Bush, 1945; Lundvall, 1992). With

the shift towards CIP frameworks, innovation policies are directly assigned a role in addressing and finding solutions for societal challenges. Accordingly, we find that the intervention logic of CIPs is oftentimes inverted with respect to 'traditional' innovation policy frameworks. Using Theories of Change (ToCs), CIPs often take desired impacts rather than specific interventions as their point of reference (Weiss, 1995). This inverted intervention logic has consequences when considering monitoring and evaluation for accountability as well as for learning. Based on our analysis, we argue that this reframing comprises a fundamental shift in the role of monitoring and evaluation within a challenge-oriented context, calling for broader scoping when providing accountability and stimulating learning for CIPs.

First**, accountability**. Establishing democratic and political accountability is one of the primary functions of policy monitoring and evaluation as it contributes to citizens' trust in government and provides evidence regarding policy performance that can be used to justify implemented policy as well as inform future policy making decisions (OECD, 2020). With the shift towards CIPs, justification of R&I policy now also comes from ameliorating societal challenges (Schot & Steinmueller, 2018). This means that CIPs contribute to and are (at least partially) to be held accountable for generating solutions to societal challenges, despite the complex character and the long timescales at which (solutions to) societal challenges unfold (Janssen, 2019; Robinson & Mazzucato, 2019). As such, the concept of accountability is fundamentally different with respect to 'traditional' accountability for innovation policy: in addition to accountability for the expenditure of public resources on innovation, accountability now also encompasses adequately addressing societal challenges through supported transition-pathways. This includes taking into consideration the socio-economic and ecological implications of supported transition pathways and a review of the uptake of solutions (Wanzenböck, 2020). As framings of both problems and fitting solutions are oftentimes highly contested, accountability is meant to support stakeholders' need for trust in government and deliver evidence regarding policy performance longitudinally – answering the question whether government is still doing the right thing given the societal goal (rather than the question did we do things right). Finally, because societal challenges often lack clear ownership structures, their governance tends to be more distributed, bottom-up - and less unidirectional and top-down - thus inhibiting or at least making accountability more challenging (Arnold et al., 2018; Hertting & Vedung, 2012).

Second, **learning**. Monitoring and evaluation can assist in the assessment of value in regard to the aims, goals or objectives of a programme or policy. Aside from gaining insight into achievements, such an evaluation also enables reflection and the possibility to assist in the identification of (future) changes needed (Kuhlmann & Rip, 2018; Mazzucato, 2018; Haddad, 2022). That is, the wickedness and long term horizon of societal challenges raise the need for reflexivity to look at the progress made and how. However, it also raises the need for adaptivity to be built into the monitoring and evaluation framework in order to mitigate transformational failures (Amanatidou et al., 2014; Weber & Rohracher, 2012). Various CIP-frameworks emphasise this need for reflexivity and adaptivity in some shape or form by highlighting the need to integrate learning throughout the policy cycle and/or the multiple organisational/governance levels (Kattel & Mazzucato, 2018; Kuhlmann et al., 2018; Loorbach, 2010; Lindner et al., 2016; Molas-Gallart, 2021). This need for reflexivity and adaptivity stems from the uncertainty in developments of the societal challenge on the one hand and possible solutions on the other hand. Accordingly, effectively facilitating diverse learning needs of various stakeholders across governance levels of CIP is an active field of study (Amanatidou et al., 2014; Aranguren et al., 2017; Borras, 2011; Haddad et al., 2022; Janssen, 2022; Luederitz; 2017, Magro & Wilson, 2019). Aranguren et al. (2017), for instance, noted the importance of demand for evaluation when it comes to transformative capacity. In a similar vein, Magro & Wilson (2019) highlighted legitimacy of the governance of evaluation as a prerequisite for dealing with contestation. Facilitating policy learning in practice remains, however, difficult as it is highly complex, can conflict with providing accountability, and often does not enjoy priority from all involved stakeholders (Amantidou et al., 2014; Magro & Wilson, 2019).

4.2.2. OPERATIONALISATION AND ATTRIBUTION OF TRANSFORMATIVE IMPACT IN A CIP CONTEXT

Next to the fundamental shift in the role of evaluation within a CIP context, our analysis shows that the wicked context of societal challenges and CIPs also accentuates the limitations of current monitoring and evaluation practices, giving prominence to bottlenecks related to the **operationalisation** of transformative impact and the **attributing** of results to specific policy interventions and policy targets (Arnold et al., 2018; Duckett et al., 2016; Grillitsch et al., 2018; Magro & Wilson, 2019; Janssen, 2019; Luederitz, 2017; Rittel & Webber, 1973). That is, within CIP the (a) growing number of policy targets and instruments, that are (b) interacting with one another and (c) adapted in an uncontrolled way makes building an evidence-base for policy making and policy learning increasingly

difficult (Adam, 2018). As such, CIPs exacerbate pre-existing bottlenecks related to operationalisation and attribution (Amanatidou et al., 2014; Aranguren et al., 2016; Haddad, 2022).

First, **operationalisation**. Operationalisation focuses on collecting and interpreting data that gives insight in performance, including expected and realised results. The shift towards CIPs makes operationalisation increasingly difficult, given the wickedness of the societal challenges, the potential of change over time, and the increasing blurring of policy domains (Kivimaa, 2022; Lindner et al., 2021; McLaren & Kattel, 2022). Measuring the additionality of innovation policies applied to complex systems is already a difficult matter (Arnold, 2004; Janssen, 2019; Magro & Wilson, 2013). However, measuring additionality in terms of transformative impact is even more complex due to the difficulty of 1) attributing transformative impact to policy measures and 2) the long timescales at which impact tends to manifest (Amanatidou et al., 2014; Haddad, 2022; Janssen, 2019; Molas-Gallart, 2021). In their evaluation framework centred around transformative outcomes, Molas-Gallart et al. (2021), for instance, emphasize the difficulty of establishing causal links between innovation policy and transformative outcomes. In a similar vein, Janssen (2019) explains that "**as transformative policy involves adaptation of elements of the specific environment firms are active in, it leads to systemic change which opens up possibilities also for firms not directly involved. The result is the lack of a good counterfactual**" (p. 82). Hence, the need for outcome/impact-based monitoring and evaluation, combined with elevated contextual complexity associated with CIPs complicates the operationalisation of challenge-led monitoring and evaluation.

Second**, attribution**. The need for outcome-based evaluations, stemming from the inverted intervention logic associated with CIPs, makes attributing results to individual policy elements increasingly difficult. Additionally, because (the solutions to) societal challenges span across various policy domains, governance levels and regional borders, CIPs are often part of a complex policy mix consisting of policy instruments, strategies, characteristics and processes spanning across multiple dimensions (Flanagan, 2011; Rogge & Reichardt, 2016). As a consequence, CIPs can rarely be evaluated outside of the policy context it is part of, magnifying the so-called attribution problem (Adam, 2018; Bovaird, 2014; Kern, 2019; Schuch, 2017). Belcher & Hughes (2021) explained that "**if the treatment itself is multi-pronged, evolving, and/or under-specified, there will be uncertainty as to which variation of the emergent research-informed innovation is responsible and for which specific effects**" (p. 160). Accordingly,

Haddad et al. (2022) identified "**attributing the effects of policy and performing ex-ante evaluation**" as one of the main challenges for policy practitioners working with CIPs. The inherent limitations with regards to attribution in a complex policy mix, in conjunction with the need for outcome-based evaluations, has led scholars and practitioners engaged in the evaluation of CIPs to increasingly forgo establishing the effects of a single policy instrument, focusing on establishing contribution instead (Janssen, 2019; Kivimaa & Kern, 2016; Molas-Gallart, 2021).

Our analysis of the literature thus indicates that the monitoring and evaluation of challenge-led innovation policies requires consideration of the bottlenecks from a challenge-led perspective. While we acknowledge that not all challenge-led policies require a completely novel approach to evaluation (i.e. there might not be the need to renew everything all at once, nor to renew everything every time), the identification of these bottlenecks ensures that policy-makers can identify issues that are likely to occur based on the bottleneck. Furthermore, our distinction of two categories of bottlenecks shows that the challenges towards making monitoring and evaluation challenge-led is different depending on the category of changes.

4.3 LITERATURE ON ADDRESSING BOTTLENECKS OF CHALLENGE-LED EVALUATION

Our analysis of relevant literature also indicates suggestions to address the bottlenecks that we discussed in the previous section. Figure 3 summarises the main issues per bottleneck, as well as the possible solutions that have been proposed or developed in the literature. Note, however, that the literature screened is in itself not elaborate or clear regarding what these solutions could look like in practice, or what would ensure that these solutions will make a difference in the monitoring and evaluation practice. Next, we will discuss the potential routes to arrive at an appropriate monitoring and evaluation framework for challenge-led innovation policies as identified in the literature.

Figure 3: Issues and solutions of addressing bottlenecks of challenge-led evaluation

Bottleneck:	Issue:	Possible solution:
Accountability	Complexity, uncertainty & contestation	Theory-based evaluation, multi-level governance and evaluation system, separate formative & summative evaluation
Learning	Insufficient facilitation of reflexivity and adativity	Formative evaluation across governance levels and domains
Operationalisation	Ambiguity	Qualitative and normative approach, Flexible ToC
Attribution	Interaction effects, Ambiguity	Contribution to outcome, Contribution to impact
Governance & organisation	Lacking reflexive & adaptive governance, politicisation & continuity of the approach	Experimentation in governance and organizational structures
Institutional rigidity	Institutionalised evaluation practice not designed for CIP	Experimentation in evaluation approaches

Note: ToC denotes theory of change.

4.3.1 POTENTIAL ROUTES TO ADDRESS BOTTLENECKS RELATED TO A FUNDAMENTAL SHIFT IN THE ROLE OF R&I POLICY EVALUATION WITHIN A CIP CONTEXT

The **accountability** bottleneck concerns the need to establish accountability for adequately addressing societal challenges through supported transition pathways. CIPs are often nested within a complex challenge-led policy mix and are characterised by a high degree of uncertainty and contestation both in terms of challenges and solutions (Arnold et al., 2018; Belcher & Hughes, 2021; Wanzenböck et al., 2020). Taking into consideration these characteristics, scholars and evaluation practitioners have suggested providing accountability using: theory-of-change-based evaluation, a transparent and reflexive multi-level governance and evaluation system with a variety of stakeholders, and/or inclusive or participative formative evaluation combined with separate external summative evaluation (Amanatidou et al., 2014; Arnold, 2018; Belcher & Hughes, 2021; Magro & Wilson, 2019; Miyaguchi, 2022; Molas-Gallart et al., 2021). Thus, challenge-led monitoring and evaluation is defined by its transparency of what can be evaluated along the theory-of-change, and its inclusiveness of stakeholders in the monitoring and evaluation process over time.

The **learning** bottleneck deals with the elevated need for reflexivity and adaptivity that characterises CIP frameworks and imposes additional requirements

regarding the extent to which challenge-led monitoring and evaluation should facilitate learning. In response, numerous frameworks that promote learning through formative evaluation across governance levels and policy domains have been presented (Amanatidou et al., 2014; Aranguren et al.,, 2017; Janssen, 2019; Luederitz, 2017; Molas-Gallart, 2021; Räkköläinen and Saxén, 2022; Wittmann et al., 2022). Hence, challenge-led monitoring and evaluation's emphasis on learning and adaptivity emphasizes that addressing societal challenges is a process, which thus requires steering towards the societal goal.

Interestingly, the focus of and attention for accountability and learning is, in some sense, also paradoxical as it arises from an inherent tension that exists between accountability and learning. That is, as accountability is demanded but increasingly complex from a challenge-led perspective, actors might be less willing to share information that could inform learning. In addition, as evaluation with the goal of learning also concerns expenditure (time, extra staff/human capital), doing so should pay off but should also be accounted for. And, finally as accountability and learning from a challenge-led perspective demands an inclusive process, it will become increasingly difficult to actually ensure accountability (e.g. a butcher inspecting and testing its own meat) (Amanatidou et al., 2014; Arnold et al., 2018; Haddad et al., 2022; Magro & Wilson, 2019; van der Steen et al., 2018).

4.3.2 POTENTIAL ROUTES TO ADDRESS BOTTLENECKS RELATED TO THE OPERATIONALISATION AND ATTRIBUTION OF TRANSFORMATIVE IMPACT IN A CIP CONTEXT

Such paradoxical tensions appear less so considering the bottlenecks related to the operationalisation and attribution of transformative impact in a CIP context. The **operationalisation** bottleneck has to deal with ambiguity that stems from the wickedness of societal challenges, their longitudinal horizon and their blurring of policy domains. Consequently, it has been proposed that theory-based qualitative, normative and flexible approaches should be used to operationalise societal challenges for evaluation, in which complex, context-specific considerations can be taken into account more (Amanatidou et al., 2014; Aranguren, 2017; Arnold, 2018; Janssen, 2019; Luederitz, 2017; Molas-Gallart, 2021).

Finally, the main issue that the **attribution** bottleneck is confronted with are interaction effects that stem from the complexity of challenge-led policy mixes. Thus, as a solution, scholars and evaluation practitioners have suggested mov-

ing away from aiming to establish attribution, towards establishing contribution (Janssen, 2019; Kivimaa & Kern, 2016; Wittmann et al., 2022). As such, room is provided to consider the extent to which a solution contributes to addressing a societal goal (or not), and whether the policy is still doing the right thing(s).

4.3.3 ADDITIONAL POTENTIAL ROUTES TO ADDRESS BOTTLENECKS

Finally, our analysis also indicates two additional bottlenecks that appear insufficiently problematised and addressed in the literature and policy practice. These bottlenecks appear as they are presented as (part of) solutions to all four identified bottlenecks, but they are presented as an afterthought that needs 'experimentation'. These bottlenecks are: 1) 'governance and organisation' that is lacking in reflexivity and adaptivity and increasingly politicised at the expense of the continuity of the approach, and 2) 'institutional rigidity' sustaining an institutionalised evaluation culture and practice that is not designed to fulfil the evaluation needs within CIP frameworks. These bottlenecks highlight potentially significant issues and solutions concerning further development of challenge-led monitoring and evaluation. That is, ameliorating other identified bottlenecks in challenge-led monitoring and evaluation is contingent upon altering (a) governance and organisational structures as well as (b) institutionalised assumptions and practices. Therefore, we argue that in order to further develop challenge-led monitoring and evaluation, the roles of governance and organisational structures as well as institutionalised assumptions and practices should be problematised and prioritised as bottlenecks.

5. DISCUSSION – CONCLUSIONS AND RECOMMENDATIONS

Our literature review has shown that existing literature provides ample input on the bottlenecks, issues and possible solutions to develop challenge-led monitoring and evaluation (see Figure 2 and 3). By delineating the bottlenecks and their potential solutions, we aim to provide policy-makers and evaluators strategic guidance to understand what it takes to develop a challenge-led monitoring and evaluation 'fit-for-purpose', and which issues might hamper possible ways forward.

An important implication of our review is that governance and institutionalised monitoring and evaluation culture and practice are important bottlenecks

to consider. While existing literature does not considered these a bottleneck per se, the literature does mention that both governance and institutionalised culture and practice both require 'experimentation' (see Figure 3). To develop challenge-led monitoring and evaluation further, we therefore argue that we should de-emphasise the well-known bottlenecks and focus instead on the governance and organisation of monitoring and evaluation for societal challenges, and question the institutionalised monitoring and evaluation culture and practices. That is, we argue that altering governance and organisational structures as well as institutionalised assumptions and practices should be prioritised because ameliorating other identified bottlenecks in challenge-led evaluation is contingent upon doing so. In other words, while the main issues of the bottlenecks towards a challenge-led monitoring and evaluation can be indicated (see Figure 2), the potential routes forward are largely dependent on the institutionalisation of challenge-led monitoring and evaluation practices that are incorporated into the governance and organisation. After all, a challenge-led monitoring and evaluation is not automatically ingrained, and therefore a different design, and a change of mind-set and practices within the existing monitoring and evaluation practice is required. We thus argue that knowing that challenge-led monitoring and evaluation is necessary does not automatically lead to a different monitoring and evaluation practice. Current practices, with their habits and routines, standards, and socio-culturally accepted norms regarding monitoring and evaluation, whether imposed or not, play a role and are difficult to set aside. What is needed, therefore, is a culture change and the institutionalisation of a new, monitoring and evaluation norm or standard.

As a result, challenge-led monitoring and evaluation touches on a more institutional question of recalibrating how we deal with monitoring and evaluation in policy assessments and society at large, and what the role of policymakers, evaluators, and financiers and intermediaries is in this context. We thus argue that this not only requires further reflection and elaboration of the meaning and consequences of monitoring and evaluating bottlenecks for societal challenges, but above all a change in the governance and organisation of monitoring and evaluation as well as the culture regarding the where, who, when and how monitoring and evaluation should be carried out. That is, as the bottlenecks of challenge-led evaluation ensure the emergence of novel practices, actors, or owners that are not institutionalised, these practices, actors and owners compete with highly institutionalised ones. Consequently, these competitions are the institutional sources of contestations that policymakers, evaluators, financiers and intermediaries that want to develop a challenge-led evaluation are likely to be confronted with.

5.2 CONCLUSION

An emerging stream of research can be identified that sets out to identify and address the bottlenecks of developing monitoring and evaluation approaches for challenge-led innovation policies. Although these efforts are still fragmented, they are starting to contribute to the establishment of a common understanding of how the emergence of challenge-led innovation policies impacts the requirements for monitoring and evaluation, and how challenge-led monitoring and evaluation could strengthen policy and governance of research and innovation aimed at societal challenges. This understanding helps in the development of challenge-led monitoring and evaluation approaches 'fit-for-purpose'. However, to establish consensus, new norms among actors, and novel practices regarding the monitoring and evaluation of challenge-led innovation policies, dealing with identified governance and institutional bottlenecks requires more attention.

REFERENCES

Adam, C., Steinebach, Y., & Knill, C. (2018). Neglected challenges to evidence-based policy-making: The problem of policy accumulation. **Policy Sciences**, **51**(3), 269-290.

Amanatidou, E., Cunningham, P., Gök, A., & Garefi, I. (2014). Using evaluation research as a means for policy analysis in a 'new'mission-oriented policy context. **Minerva**, **52**(4), 419-438.

Aranguren, M. J., Magro, E., & Wilson, J. R. (2016). Regional competitiveness policy evaluation as a transformative process: From theory to practice. **Environment and Planning C: Politics and Space,** 35(4), 703-720.
Arnold, E. (2004). Evaluating research and innovation policy: a systems world needs systems evaluations. **Research evaluation**, 13(1), 3-17.

Arnold, E., Åström, T., Glass, C., & de Scalzi, M. (2018). **How should we evaluate complex programmes for innovation and socio-technical transitions?** Retrieved from technopolis |group| United Kingdom website: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-250284

Belcher, B.M. & Hughes, K. (2021). Understanding and evaluating the impact of integrated problem-oriented research programmes: Concepts and consider-

ations. **Research Evaluation**, 30(2), 154–168, https://doi.org/10.1093/reseval/rvaa024

Berger, P. L. & Luckmann, T. (1966). **The Social Construction of Reality: A Treatise in the Sociology of Knowledge**. Garden City, NY: Anchor Books.

Borrás, S. (2011). Policy learning and organizational capacities in innovation policies. **Science and Public Policy**, 38(9), 725-734.

Bovaird, T. (2014). Attributing outcomes to social policy interventions–'gold standard'or 'fool's gold'in public policy and management?. **Social Policy & Administration**, 48(1), 1-23.

Bush, V., (1980). Science--the endless frontier : a report to the President on a program for postwar scientific research. [Washington, D.C.] :**National Science Foundation.**

Diercks, G., Larsen, H. & Steward, F. (2019). Transformative innovation policy: Addressing variety in an emerging policy paradigm. **Research Policy** 48, 880–894.

Duckett, D. Feliciano, D. Martin-Ortega, J. Munoz-Rojas, J. (2016). Tackling wicked environmental problems: the discourse and its influence on praxis in Scotland. **Landsc. Urban Plan.**, 154, pp. 44-56

Elsbach, K.D. & van Knippenberg, D. (2020). Creating high impact literature reviews: An argument for integrative reviews. **Journal of Management Studies**.

Flanagan, K., Uyarra, E., & Laranja, M. (2011). Reconceptualising the 'policy mix'for innovation. **Research policy**, 40(5), 702-713.

Grillitsch, M., Hansen, T., Coenen, L., Miörner, J., & Moodysson, J. (2019). Innovation policy for system-wide transformation: The case of strategic innovation programmes (SIPs) in Sweden. **Research Policy**, 48(4), 1048-1061.
Haddad, C. R., Nakić, V., Bergek, A., & Hellsmark, H. (2022). Transformative innovation policy: A systematic review. **Environmental Innovation and Societal Transitions**, 43, 14-40.

Hajer, M. (1995). **The politics of environmental discourse: ecological modernization and the policy process.** Clarendon Press.

Hekkert, M. P., Janssen, M. J., Wesseling, J. & Negro, S. O. (2020). Mission-oriented innovation systems. **Environmental Innovation and Societal Transitions**, 34, 76-79.

Hertting, N., & Vedung, E. (2012). Purposes and criteria in network governance evaluation: How far does standard evaluation vocabulary takes us?. **Evaluation**, 18(1), 27-46.

Janssen, M. J. (2019). What bangs for your buck? Assessing the design and impact of Dutch transformative policy. **Technological Forecasting and Social Change**, 138, 78-94.

Janssen, M. J., Torrens, J., Wesseling, J. H. & Wanzenböck, I. (2021). The promises and premises of mission-oriented innovation policy—A reflection and ways forward. **Science and Public Policy**, 48(3), 438–444.

Kattel, R., & Mazzucato, M. (2018). Mission-oriented innovation policy and dynamic capabilities in the public sector. **Industrial and Corporate Change**, 27(5), 787–801. https://doi.org/10.1093/icc/dty032

Kern, F., Rogge, K. S., & Howlett, M. (2019). Policy mixes for sustainability transitions: New approaches and insights through bridging innovation and policy studies. **Research Policy**, 48(10), 103832.

Kivimaa, P. (2022). Transforming innovation policy in the context of global security. **Environmental Innovation and Societal Transitions**, 43, pp. 55-61

Kivimaa, P., Kangas, H. L., & Lazarevic, D. (2017). Client-oriented evaluation of 'creative destruction'in policy mixes: Finnish policies on building energy efficiency transition. **Energy Research & Social Science**, 33, 115-127.

Kivimaa, P., & Kern, F. (2016). Creative destruction or mere niche support? Innovation policy mixes for sustainability transitions. **Research policy**, 45(1), 205-217.

Kuhlmann, S. & Rip, A. 2018. Next-Generation Innovation Policy and Grand Challenges. **Science and Public Policy**, 45(4): 448–454.

Lindner, Ralf et al. (2021). Mission-oriented innovation policy: From ambition to successful implementation, Perspectives - Policy Brief, No. 02 / 2021, FraunhoferInstitut für System- und Innovationsforschung ISI, Karlsruhe.

Loorbach, D. (2010). Transition Management for Sustainable Development: A Prescriptive, Complexity-Based Governance Framework. **Governance**, 23(1), 161–183. https://doi.org/https://doi.org/10.1111/j.1468-0491.2009.01471.x

Luederitz, C., Schäpke, N., Wiek, A., Lang, D. J., Bergmann, M., Bos, J. J., Burch, S., Davies, A., Evans, J., König, A., Farrelly, M. A., Forrest, N., Frantzeskaki, N., Gibson, R. B., Kay, B., Loorbach, D., McCormick, K., Parodi, O., Rauschmayer, F., Schneidewind, U., Stauffacher, M., Stelzer, F., Trencher, G., Venjakob, J., Vergragt, P. J., von Wehrden, H. & Westley, F. R. (2017). Learning through evaluation – A tentative evaluative scheme for sustainability transition experiments. **Journal of Cleaner Production**, 169, 61- 76.

Lundvall, B. A. (1992). National systems of innovation: towards a theory of innovation and interactive learning.

Molas-Gallart, J., Boni, A., Giachi, S., & Schot, J. (2021). A formative approach to the evaluation of Transformative Innovation Policies. **Research Evaluation**, 30(4), 431-442.

Magro, E., & Wilson, J. R. (2013). Complex innovation policy systems: Towards an evaluation mix. **Research policy**, 42(9), 1647-1656.

Magro, E., & Wilson, J. R. (2019). Policy-mix evaluation: Governance challenges from new place-based innovation policies. **Research policy**, 48(10), 103612.

Mazzucato, M. (2018). Mission-oriented innovation policies: challenges and opportunities. **Industrial and Corporate Change,** 27, 803–815.

McLaren, J. and Kattel, R. (2022). Policy capacities for transformative innovation policy: A case study of UK Research and Innovation. UCL Institute for Innovation and Public Purpose, Working Paper Series (IIPP WP 2020-04). Available at: https://www.ucl.ac.uk/bartlett/public-purpose/wp2022-04.

Miyaguchi, T. (2022). Importance and Utilization of Theory-Based Evaluations in the Context of Sustainable Development and Social-Ecological Systems. In: Uitto, J.I., Batra, G. (eds) **Transformational Change for People and the Planet. Sustainable Development Goals Series.** Springer, Cham. https://doi.org/10.1007/978-3-030-78853-7_15

OECD (2020), **Improving Governance with Policy Evaluation: Lessons From Country Experiences**, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/89b1577d-en.

Räkköläinen, M., Saxén, A. (2022). Pathway to the Transformative Policy of Agenda 2030: Evaluation of Finland's Sustainable Development Policy. In: Uitto, J.I., Batra, G. (eds) **Transformational Change for People and the Planet. Sustainable Development Goals Series**. Springer, Cham. https://doi.org/10.1007/978-3-030-78853-7_16

Rathenau Institute (2020). **Maak werk van opgavegericht innovatiebeleid: Bericht aan het parlement.** Den Haag: Rathenau Institute.

Rathenau Institute (2021). **EU-missies voor maatschappelijke opgaven – Inspiratie uit Horizon Europe voor opgavegericht innovatiebeleid**. Den Haag: Rathenau Institute.

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. **Policy sciences**, 4(2), 155-169.

Robinson, D. K. R., & Mazzucato, M. (2019). The evolution of mission-oriented policies: Exploring changing market creating policies in the US and European space sector. **Research Policy**, 48(4), 936–948. https://doi.org/10.1016/j.respol.2018.10.005

Rogge, K. S., & Reichardt, K. (2016). Policy mixes for sustainability transitions: An extended concept and framework for analysis. **Research Policy**, 45(8), 1620-1635.

Schot, J. & Steinmueller, W. E. (2018). Three frames for innovation policy: R&D, systems of innovation and transformative change. **Research Policy**, 47(9), 1554-1567.

Schuch, K., Campbell, D., Carayannis, E. G. and Edler, J. (2017): Evaluation of Research, Development, and Innovation. In: **Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship**; p. 1-8; share and cite: doi:10.1007/978-1-4614-6616-1_200025-1

Scott, R. (1995). **Institutions and organizations**. London, Sage Publications.

van der Steen, M., Faber, A., Frankowski, A., & Norbruis, F. (2018). Opgavegericht evalueren. **Beleidsevaluatie voor systeemverandering**. Den Haag: NSOB.

Termeer, C. J., & Dewulf, A. (2019). A small wins framework to overcome the evaluation paradox of governing wicked problems. **Policy and Society**, 38(2), 298-314.

Torraco, R. J. (2016) Writing integrative literature reviews: Using the past and present to explore the future. **Human Resource Development Review**, 15(4).

Turnheim, B., Berkhout, F., Geels, F., Hof, A., McMeekin, A., Nykvist, B. & van Vuuren, D. (2015). Evaluating sustainability transitions pathways: Bridging analytical approaches to address governance challenges. **Global Environmental Change**, 35, 239-253.

Wanzenböck, I,. Wesseling, J.H., Frenken, K., Hekkert, M.P., Weber, K.M., A framework for mission-oriented innovation policy: Alternative pathways through the problem–solution space, **Science and Public Policy**, Volume 47, Issue 4, August 2020, Pages 474–489, https://doi.org/10.1093/scipol/scaa027

Weber, K. M., & Rohracher, H. (2012). Legitimizing research, technology and innovation policies for transformative change: Combining insights from innovation systems and multi-level perspective in a comprehensive 'failures' framework. **Research policy**, 41(6), 1037-1047.

Weber, M.; Polt, M. (2014): Assessing mission-orientated R&D programs: combining foresight and evaluation. **Fteval - Journal for Research and Technology Policy Evaluation**, (39), 5–10.

Weiss, C. H. (1995). Nothing as Practical as Good Theory : Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families. **New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts**, 1, 65--92.

Wittmann, F., Hufnagl, M., Roth, F., Lindner, R., & Kroll, H. (2022). Towards a framework for impact assessment for mission-oriented innovation policies. A formative Toolbox approach. **Fteval – Journal for Research and Technology Policy Evaluation**, 53, 31-42.

AUTHORS

CORRESPONDING AUTHOR:
VINCENT BAARSLAG
Rathenau Instituut, Anna van Saksenlaan 51, The Hague,
The Netherlands
Email: v.baarslag@rathenau.nl, vincent.baarslag@gmail.com
Orcid: https://orcid.org/0000-0003-2030-053X

AMBER GEURTS
Institution & Address: TNO, TNO Vector, Anna van Buerenplein 1, The Hague,
The Netherlands
Email: amber.geurts@tno.nl
Orcid: https://orcid.org/0000-0003-0478-0713

FRANS VAN DER ZEE
Institution & Address: Rathenau Instituut, Anna van Saksenlaan 51, The Hague,
The Netherlands
Email: f.vanderzee@rathenau.nl

fteval JOURNAL
for Research and
Technology Policy Evaluation

January 2024, Vol. 55, pp. e2, 1-20
DOI: 10.22163/fteval.2024.634

MONITORING AS A STEERING TOOL FOR GENDER EQUALITY POLICY IN HIGHER EDUCATION INSTITUTIONS: IDEAL AND REALITY

ANGELA WROBLEWSKI
DOI: 10.22163/FTEVAL.2024.634

ABSTRACT

In an ideal-typical process of gender equality policy development and implementation, empirical evidence plays a central role. The gender analysis at the beginning of the process describes the gender issue to be addressed. Based on this baseline, objectives and priorities are defined, policies are designed and implemented, and these are then monitored and, ideally, evaluated. The aim of monitoring is to provide information on the implementation of the gender equality measure so that countermeasures, if needed, can be taken at an early stage in the event of deviations from the planned implementation. However, this requires that the monitoring indicators are derived from the policy objectives and are measurable. Practice shows that the availability of data, rather than the objectives, determines the development of indicators. The procedure for developing indicators is described and reflected on using the example of monitoring the implementation of the recommendations of the Austrian Convention of Higher Education Institutions on strengthening gender competencies in higher education processes. The preconditions for the steering function of monitoring and its limitations are discussed.

Keywords: gender equality policy, monitoring of policy implementation, higher education institutions, policy steering

INTRODUCTION

When the monitoring of the implementation of policies or measures is set in a complete and ideal-typical process (May and Wildavsky 1978), the central role of empirical evidence in the design and implementation of policies becomes clear. In this setting, monitoring also has enormous potential for steering policy implementation (Markiewicz and Patrick 2016). This is especially relevant in the case of gender equality policy, an area often characterised by a lack of strong evidence. The article argues that monitoring can be a powerful tool for steering gender equality policy implementation and shows that its use remains limited, and for what reasons.

Monitoring is set against the background of reflexive gender equality policy (Wroblewski and Palmén 2022; Wroblewski and Leitner 2022). Evidence plays a central role in a complete and reflexive process of designing and implementing gender equality policy: first in the analysis of the gender equality problem (gender analysis), from which goals and priorities are defined in a further step, which in turn form the basis for designing measures. The implementation of measures is subject to monitoring. This involves the systematic collection and analysis of information on the relevant context and the implementation of the measures. On this basis, it is determined whether the measures are being implemented as planned or whether adjustments are necessary in the event of deviations. In this way, monitoring contributes to the efficient implementation of measures. To fulfil this function, monitoring indicators must be derived from the objectives of the action.

However, this model is based on a number of assumptions, not all of which are tested when developing a policy. Usually, they are only tested as part of an evaluation. The model of a complete policy cycle described above assumes that gender equality policies are designed and implemented based on the following preconditions:

- That management is not only committed to gender equality goals but also to changing the structures and processes which cause gender inequalities.
- That all actors involved have a common understanding of the gender equality problem to be addressed by the policy, as well as of the objectives and priorities to be addressed.
- That sufficient resources are allocated to the implementation of the measures and that they address the gender equality policy issue.

- That monitoring is meaningful and that there is a willingness to use it for steering purposes.

There are a number of examples that could be used to illustrate that these prerequisites are not always met in practice (see e. g. Wroblewski 2021, Wroblewski 2017). In the following, the limited steering power of monitoring will be discussed using the example of monitoring the implementation of the recommendations on strengthening gender competencies in higher education processes formulated by the Austrian Convention of Higher Education Institutions (*Hochschulkonferenz*, HSK; BMBWF 2018; Wroblewski and Englmaier 2023).

THE HSK RECOMMENDATIONS ON STRENGTHENING GENDER COMPETENCIES IN HIGHER EDUCATION PROCESSES

Austria has a long tradition of gender equality policies in science and research. The existing policy mix, which has been developed over the years, addresses the three central objectives that also characterise European gender equality policy in the field of science and research: (1) achieving gender balance in all areas and at all hierarchical levels in science and research (fixing the numbers), (2) integrating the gender dimension into higher education structures and processes (fixing the institution) and (3) anchoring the gender dimension in research content and teaching (fixing the knowledge).

With the existing bundle of gender equality policy measures, Austria is considered a country with a comprehensive and consistent policy mix in a European comparison (Wroblewski 2021). However, there is a need to improve the effectiveness of existing measures, as formulated as an objective in the Austrian ERA Roadmap 2016-2020 (BMWFW 2016). In order to achieve this goal, a working group was set up in 2016 to formulate recommendations for strengthening gender competencies in higher education processes. In the first phase, the working group developed and agreed on a definition of gender competence which is defined as a basic competence all members of a higher education institution should have and is thus distinguished from gender expertise.

"Gender competence comprises of the fundamental recognition of the relevance of gender attributions in one's own work and sphere of influence (knowledge). This recognition is connected to the willingness (desire) and ability to deal with these issues in day-to-day work and throughout study life - if necessary, supported by gender experts and with knowledge from gender theories - and to take action based on this knowledge (skills). Recognition, discussion and action are subject to a constant process of reflection (reflection)." (BMBWF 2018: 33, translated by author)

The working group consisted of representatives of universities, universities of applied sciences, university colleges of teacher education and private universities, as well as representatives of the Austrian National Union of Students and gender and equality experts. The task was to develop recommendations for members of higher education institutions to improve gender competence and raise awareness of gender diversity. The recommendations are intended to serve as a guideline for the persons and committees involved at the universities and to provide concrete, action-oriented suggestions (BMBWF 2018). The recommendations are addressed (1) to the universities, but also (2) to the Austrian National Union of Students, (3) to the Austrian Convention of Higher Education Institutions itself, (4) to the Austrian Conference of Universities (Uniko), the Austrian Conference of Universities of Applied Sciences (FHK), the Austrian Conference of Private Universities (PUK) and the Austrian Rectors' Conference of Universities of Teacher Education (RÖPH), and (5) to the Federal Ministry of Education, Science and Research (BMBWF).

The working group formulated a total of 36 recommendations that address four subject areas. For each recommendation, it is made explicit who is responsible for its implementation and who benefits from its realisation.

(1) Gender competent management: The 18 recommendations assigned to this area aim to achieve a clear commitment on the part of higher education institution (HEI) management to a cultural change in the sense of gender- and diversity-equitable organisation.

(2) Gender competent action: The implementation of the nine recommendations assigned to this area is intended to support the responsibility of all HEI members to build and develop their own gender competence. All members of the HEI are thus called upon to actively help shape their own field of work and activity in a gender-competent manner.

(3) Gender competent teaching: These seven recommendations are aimed at anchoring the gender dimension more firmly in the curricula as well as gender- and diversity-appropriate teaching.

(4) Gender competent research: These two recommendations aim to consider the gender dimension in research content as a cross-cutting issue.

OVERVIEW 1 RECOMMENDATIONS FOR GENDER COMPETENT MANAGEMENT

No.	What?	Who is responsible?	Who profits?
1	Setting targets to develop and strengthen gender competencies	All HEI members	All HEI members
2	Consideration of existing strategy papers and instruments	All HEI members	All HEI members
3	Consideration of knowledge platforms	All HEI members	All HEI members
4	Responsibility of a member of management to develop and strengthen gender competencies	All HEI members	All HEI members
5	Provision of resources to develop and strengthen gender competencies	Management of Universities of applied sciences, University colleges of teacher education and private universities	All members of Universities of applied sciences, University colleges of teacher education and private universities
6	Integration of gender competencies into the profile for commissions/committees	All HEI members	Members of decision making bodies, future applicants
7	Integration of gender competence into organisation development and quality management	All HEI members	All HEI members
8	Developing expertise to create an index for gender competencies	All HEI members	HEI management, BMBWF
9	Organising a cross-sectoral networking meeting	Austrian Convention of HEIs	All HEI members
10	Cooperation when developing gender competence trainings	All HEI members, human resources department	All HEI members, especially teachers and researchers

11	Establishment of platforms to bundle knowledge on gender competence	FHK, RÖPH, PUK	All members of Universities of applied sciences, University colleges of teacher education and Private universities
12	Bundling and developing human resources in the area of gender education/gender studies	Steering groups of the networks developing teacher education	All HEI involved in teacher education, all students of pedagogy and teacher education
13	Designing the infrastructure so it does justice to gender diversity	Owners of HEI buildings, HEI management	Non-binary HEI members
14	Integration of gender competence into the development and finance plan	BMBWF, FHK, RÖPH	All members of Universities of applied sciences and University colleges of teacher education
15	Legal anchoring of gender competencies	BMBWF	All HEI members
16	ÖH-department on gender at every HEI	BMBWF	Austrian National Union of Students (ÖH), students
17	Consideration of real gender diversity when collecting data	BMBWF, HEI management, Statistics Austria	Non-binary HEI members
18	Design forms based on gender diversity	BMBWF, HEI management	Non-binary HEI members

Source: BMBWF 2018: 28f.

OVERVIEW 2 RECOMMENDATIONS FOR GENDER COMPETENT ACTION

No.	What?	Who is responsible?	Who profits?
19	Provision of information on gender aspects of the HEI	HEI management	All HEI members
20	Recognition and processing of information on gender issues	All members of HEI	All HEI members
21	Discussion of gender competence among personnel managers	Human resources department	All HEI members, future applicants
22	Creation of individual and collective reflection opportunities, also with external support	HEI management	All HEI members, future applicants
23	Development of an self-assessment tool	HEI management in cooperation	Specific groups of HEI members
24	Development and/or provision of a guide on gender sensitive language	HEI management	All HEI members
25	Low-threshold access to language guide and promotion of its use	HEI management	All HEI members
26	Provision of training for ÖH representatives and tutors	Austrian National Union of Students (ÖH)	ÖH representatives, students
27	Collection of good practice examples by departments/disciplines	HEI management in cooperation	Teachers, researchers

Source: BMBWF 2018: 29.

OVERVIEW 3 RECOMMENDATIONS FOR GENDER COMPETENT TEACHING

No.	What?	Who is responsible?	Who profits?
28	Integration of Gender Studies in Curricula	Collegial bodies responsible for curriculum development	Teachers, students
29	Anchoring the development of gender competence in the content of AHS and BHS curricula	Curriculum commission	Students in secondary education
30	Anchoring the development of gender competence in STEOP	Collegial bodies responsible for curriculum development	Students
31	Provision of further training in gender-sensitive didactics	Human resources department	Teachers, students
32	Individual coaching for teachers to develop the gender dimension	Human resources department	Teachers, students
33	Award for theses addressing gender issues	HEI management	Students
34	Award for gender competent teaching	HEI management	Students

Notes: AHS – Academic secondars school, BHS – Vocational secondary school; STEOP –introductory phase of the study programme

Source: BMBWF 2018: 30.

OVERVIEW 4 RECOMMENDATIONS FOR GENDER COMPETENT RESEARCH

No.	What?	Who is responsible?	Who profits?
35	Targeted funding for research projects	BMBWF, HEI management	Researchers, students
36	Establishment of gender professorship(s)	HEI management	Researchers, students

Source: BMBWF 2018: 30.

AIM OF MONITORING AND ACCESS TO INDICATOR DEVELOPMENT

The monitoring which accompanies the implementation of the HSK recommendations was initiated and funded by the BMBWF. On the one hand, the monitoring provided a comparative presentation of the activities undertaken by the universities to enhance gender competence. On the other hand, the monitoring should provide input for the accompanying process of the BMBWF which supports the implementation of the recommendations through a collegial consultation process and the provision of a platform for exchange and discourse (see recommendations 9 and 10).

The development of the indicators for monitoring the implementation of the HSK recommendations was based on four principles (see also Wroblewski et al. 2017):

- Indicators must meet established quality criteria for social scientific measurement, i.e. they must be objective, valid and reliable.
- Valid criteria must make explicit the underlying understanding of gender and gender competence and relate to specific gender equality policy objectives.
- Relevant stakeholders must be involved in the process of developing indicators to ensure their acceptance and use.
- The development of indicators should be understood as a reflexive process, analogous to the development of strategies to enhance gender competence in higher education processes. The analysis of the indicators should be combined with a reflection on the meaningfulness of the indicators for the particular context of application and thus constitute the starting point for the further development of the indicators or the respective data base.

In light of these principles, the development of the monitoring programme was designed to be participatory and discursive. Drawing on the approaches of feminist institutionalism and practice theory (see Lipinsky, Wroblewski 2021), the project was based on the assumption that the development and expansion of gender competence among all HEI members is a prerequisite for achieving the central goal of gender equality policy in science and research - cultural change. The creation of structures for gender equality as well as the development and implementation of concrete measures should support the development of gender competence (Löther et al. 2021).

The specific aim of the monitoring was to contribute to an evidence-based discourse on gender equality, in particular on the status quo and changes in activities and measures to enhance gender competence in HEIs. To ensure the use and acceptance of such monitoring, relevant stakeholder groups were involved in the process and the link to a feminist discourse was ensured (Wroblewski et al. 2017). Broad acceptance of monitoring also seems important to enable its use as a steering tool for higher education processes in the context of gender equality policy (Wroblewski, Lipinsky 2018; Eckstein 2017; Wroblewski 2017). Furthermore, the monitoring should represent a further development or complement to the established gender monitoring (www.unidata.gv.at) and be based as far as possible on available or easily accessible data.

Indicators were developed for different levels: Input indicators provide information on the activities carried out. In the case of recommendations addressed to universities, this provides a basis for comparing implementation activities. Implementation indicators do not only show whether activities have been carried out, but also their scope and intensity. Impact indicators illustrate the achievement of objectives and the changes brought about by the measures implemented. Implementation and impact indicators are less suitable for making comparisons between higher education institutions, as they would have to consider differences in, for example, the size of the institution or its thematic focus.

Implementation and impact indicators are therefore intended as suggestions for universities that wish to know more about the effectiveness of the measures they have implemented. However, there are no plans to collect implementation and impact indicators for all HEIs on a comparative basis. This is also because it only makes sense to generate the relevant data if the institution is interested in promoting the evidence-based further development of specific measures to establish and expand gender competence.

THE PROCESS: INDICATOR DEVELOPMENT AND FEEDBACK

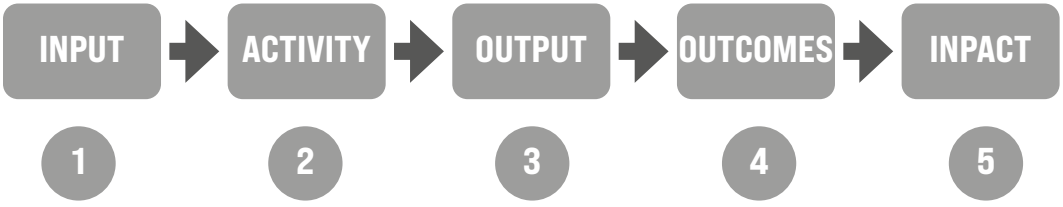
The indicators were developed in the following way: In a first step, an impact logic (based on the logic model developed by W. K. Kellog Foundation 2004) was formulated for each recommendation in the light of the current state of gender equality policy debate and research. The formulation of the impact logic is intended to adequately reflect the complexity and multidimensionality of the recommendations. Gender equality in higher education is not a goal that can be easily quantified, i.e. an adequate set of indicators will have to include both qualitative and quantitative indicators. For example, the proportion of women in

professorships alone is not a meaningful indicator of gender equality in universities (Wroblewski 2021). The proportion of women says nothing about potential sources of bias in appointment procedures, e. g. how appointment procedures are organised, how many women applied or were invited to interviews. In general, Indicators are only an approximation of a construct that is not directly measurable (Meyer 2004: 27). The construct can be defined theoretically, politically or ad hoc. In our case, the recommendations are a political goal, the operationalisation of which forms the basis for the development of the indicators.

Each of the recommendations addresses complex and, in some cases, multidimensional objectives. Recommendation 24, for example, suggests that universities develop a guideline on the use of gender-sensitive or gender-inclusive language for HEI members. The aim is to ensure that all types of teaching, research and administrative texts are formulated in a gender-sensitive or gender-inclusive manner. However, the existence of a guideline does not mean that members of the HEI are aware of it and use it, or that texts are actually formulated in a gender-equitable or gender-inclusive way.

The development of indicators to monitor the implementation of the recommendations was based on the ideal approach to developing gender indicators (Wroblewski et al., 2017). Accordingly, the monitoring system to support the implementation of the HSK recommendations should include indicators that can be derived from the objectives of the recommendations and provide information on their implementation. A distinction is made between input indicators, implementation indicators (activities) and impact indicators (short, medium and long term). These indicators are derived from the so-called logic model, i.e. the formulated assumptions as to why a measure should contribute to a desired result. Compared to logic models used in evaluations (Rossi et al. 1999), the model used for the development of indicators to monitor the implementation of the recommendations is rather simplified. For the sake of simplicity, a linear model is formulated, although in practice the implementation of interventions is often characterised by feedback loops or similar non-linear relationships.

Figure 1 Logic model



Source: own elaboration based on W.K. Kellogg Foundation 2004: 1.

The first two stages of the logic describe the measure and its implementation. Corresponding input indicators are, for example, the resources available for a measure, the planned training units or the number of participants. Implementation indicators describe the activities carried out and the results produced by a measure, such as the number of people trained, or the number of consultations carried out. Outcome indicators show short-term effects, for example when trained people are able to apply the newly acquired knowledge in their daily work. Impact indicators, on the other hand, are designed to reflect long-term changes and in the case of gender equality in HEI cultural change, such as the integration of the gender dimension in research projects as a crosscutting issue.

The examples already show that in some cases the relevant information is relatively easy to obtain from administrative data, such as the number of participants in a training course. In other cases, the information is more complex and needs to be generated through separate surveys, such as the skills acquired by participants in a training programme. The complexity of gender monitoring is compounded by the fact that many databases are not inherently suited to analysing gender issues. While gender equality policies tend to focus on social gender, i.e. different groups of women and men in specific life situations and with the role expectations placed on them, databases tend to represent biological gender (sex) (Döring 2013). Sometimes databases allow a differentiated view of social phenomena, e. g. when women and men can be represented as heterogeneous groups, for example by differentiating according to age, qualification level, childcare responsibilities, health impairments or sexual orientation. As such differentiations are rarely possible, reflection on the validity of existing databases for gender analysis should be embedded as a quality standard in the indicator development process (Hedman et al. 1996). The development of gender or equality monitoring is therefore a long-term process and monitoring itself becomes a 'living tool' that is subject to constant adaptation and development (Wroblewski, Leitner 2022). Hedman et al. (1996: 11) even state: "The production of gender statistics is a never-ending process".

The indicators were formulated knowing that not all indicators have corresponding databases or that existing information cannot be used for analysis without considerable effort. A second step was therefore to identify what data was available, what information was easily accessible (e. g. through documents) and what information should ideally be available. The indicators were deliberately formulated before data availability was checked in order to identify data gaps (D'Ignazio, Klein 2020; Criado-Perez 2019), to include them in the reflection on the indicators and to stimulate discussion on alternative ways of gen-

erating data (Wroblewski et al. 2017). Where possible, several indicators have been proposed for each recommendation. For each of these, the level of indicator addressed is explained and possible data sources are discussed.

Parallel to the development of the indicators at the theoretical level, a survey was conducted among the universities on the status of implementation of the HSK recommendations and a systematic search was conducted for available information on the implementation of measures to develop and expand gender competence (strategic documents of the universities, homepages, etc.). The survey of the universities took place in spring 2022, whereby the contact persons nominated by the universities for the implementation of the recommendations were contacted. A total of 66 of the 71 universities took part in the survey, representing a response rate of 93%. An internet search was carried out in April and May 2022. Publicly available documents (such as performance agreements, development plans, plans for the advancement of women or gender equality, gender reports, diversity strategies, statutes, checklists, position papers or information brochures) were included in the research. The information in the completed questionnaires was compared with that from our own research to check the validity and significance of the results.

Based on the results of the survey and the research, the indicators and the underlying data were analysed with a particular focus on their validity. The focus was on the institutions' understanding of gender competence in the implementation of the recommendations. What understanding of gender competence can be gleaned from the documents or the survey? To what extent can different concepts of gender competence be identified? To what extent does this limit the meaningfulness of the indicators? This reflection took place together with Dr Anke Lipinsky from GESIS in June 2022. The set of indicators has been revised on the basis of the reflection, e. g. by providing more precise definitions or by removing multidimensional indicators. This set of indicators was discussed with the BMBWF (June 2022). The set of indicators and the results of the survey and research were presented and discussed at the networking meeting on gender and diversity competence in October 2022. The first monitoring report was sent to all participants of the networking meeting in September 2022; the survey participants also received a summary of their answers and information in case the answers were interpreted differently by the monitoring team.

Four one-day workshops were held in February and March 2023 for representatives of the higher education sector. Each of the workshops addressed one of the four sectors which make up the higher education sector in Austria

(universities, universities of applied sciences, teacher training colleges and private universities).[1] The workshops took place at the Institute for Advanced Studies and were moderated by Dr Birgit Buchinger (Solution, Social Research & Development). The contact persons nominated by the universities for the implementation of the HSK recommendations and the respondents to the survey were invited to these workshops. A total of 40 universities provided feedback during the workshops.

The aim of the workshops was to discuss with HEI representatives the relevance of the recommendations in general and the possibilities of using the indicators of the monitoring to steer gender equality policy at HEI level. Strengths and weaknesses of the implementation in the respective sector as well as the meaning of the indicators from the perspective of the HEIs should be reflected upon.

MAIN RESULTS OF THE REFLECTION PROCESS

The workshops with HEI representatives were characterised by constructive and active discussions between the participants. The opportunity to give feedback was appreciated by the participants, as was the continuation of the exchange through the annual networking meetings.

In all four workshops, the recommendations were described as helpful and supportive. Participants saw them as a guide for action. They allow them to take stock of what has already been implemented at their own HEI and provide suggestions for future action. The recommendations were also used by the participants as an argumentation aid vis-à-vis the HEI management in the course of developing concrete measures. In this context, it was considered helpful that the recommendations were brought to the universities from outside and that the gender equality actors at the universities could refer to them. The recommendations were perceived by some participants as "pressure from outside" - especially from the BMBWF - which increased the willingness of the management level to deal with the issue of gender competence.

All the workshops pointed to the need to increase the "visibility" of the recommendations and to the fact that sustained communication of the recommendations is essential for their successful implementation. When managers change, there is no guarantee that new members of management will be aware of the recommendations or consider them relevant. There is also a certain turnover of gender equality stakeholders, so it is important to ensure that new stake-

1 This approach was chosen to take account of the different frameworks (e. g. legal requirements) for gender equality in the sectors.

holders have access to the recommendations as quickly and directly as possible.

At the same time, the discussions in all workshops showed that the recommendations' significance and binding nature remained unclear. On the one hand, the external pressure mentioned above was perceived, on the other hand, the word "recommendation" and the associated low level of commitment was pointed out. In none of the workshops were the recommendations perceived as a product of the Austrian Convention of Higher Education Institutions, and only a few of the participants were aware that the representatives of the higher education sectors themselves had adopted the recommendations. The BMBWF was perceived as the central actor, probably also because it took over the organisation of the annual networking meetings and the communication of the recommendations.

The complexity of the recommendations was also discussed in relation to their importance or binding nature. In all four workshops, it was discussed that the simultaneous implementation of all recommendations would overburden HEIs. This is particularly the case if the implemented measures are also to be monitored beyond the input level or even be subject to an evaluation.

The indicators were also considered helpful and supportive in principle. However, the discussion on the level of commitment focused mainly on the indicators. It was unclear to the participants why, in the context of "recommendations", reporting on their implementation should take place (input indicators). The lack of clarity was even greater for the implementation and impact indicators. The necessary data collection was seen as problematic, as it requires the commitment of the HEI management on the one hand and the provision of considerable resources on the other. Both seem realistic to the participants only if there is a corresponding demand and support (including budget) from the BMBWF.

Beyond the issues of commitment and resource allocation, participants saw a need for further concretisation - in the recommendations themselves and, subsequently, in the indicators. It was repeatedly emphasised that it is unclear how gender competence can be measured in concrete terms - at HEI level and at individual level. How can one know that a HEI has achieved the goal, or how can one know that an individual is gender competent? In general, there is a need for more concrete examples of how the recommendations or indicators can be implemented, e. g. how gender competence can be specifically asked of applicants in the application process and how this can be captured by indicators in monitoring.

In addition, the need for more concrete examples and handouts related to an intersectional approach was discussed. In practice, it is often unclear how other dimensions of diversity in addition to gender can and should be taken into account when designing and implementing policies.

It was noted in all the workshops that HEIs had developed specific tools for themselves to implement the recommendations. Participants felt that it would be more efficient and contribute to consistent standards if such tools were developed jointly by HEIs within a sector or across sectors (e. g. by setting up a working group) to avoid a situation in which the wheel is reinvented several times. Examples of issues raised in this context were a guide to gender-inclusive language or criteria for gender competence in application procedures (on the part of applicants and HEI staff involved in the application process). Other common challenges were the design of training courses on different topics, the development of quality standards for training courses or the creation of a pool of qualified trainers.

The discussion highlighted both sector-specific characteristics and challenges in implementing the recommendations that are common to all four sectors. For example, differences in HEI culture were repeatedly mentioned, although these can also be found within a sector. On the other hand, it became clear that the starting point for existing measures at public universities differs from that at private universities. For example, the recommendations for universities with a long tradition of gender equality policies contain little that is new. The different importance of research in the sectors and the different relevance of teacher training were also discussed.

The effective involvement of senior management in the implementation of recommendations was identified as a common challenge. Although the recommendations are addressed to the management level, in most cases they delegate implementation and rarely take an active role in the implementation process. In addition, all workshops discussed the extent to which the implementation of recommendations could be strengthened through incentives or sanctions.

CONCLUSIONS

The recommendations for strengthening gender competencies in higher education processes were developed by a working group of the Austrian Convention of Higher Education Institutions (HSK) and adopted by the HSK. Individual recommendations were addressed to the BMBWF, which was requested, among other things, to organise an annual networking meeting (recommendation 9), to develop a gender competence index for universities (recommendation 8) or to anchor the development of gender competence in existing management instruments or legal bases (recommendations 14, 15 or 29, among others). During the implementation of these recommendations, the BMBWF continuously emphasised their importance. Due to this visibility in the process, those involved in their implementation at HEIs see the recommendations as a requirement of the BMBWF and not as a product of the Austrian Convention of Higher Education Institutions. This would also result in the loss of the character of a voluntary commitment on the part of the universities that have undertaken to implement the recommendations.

The monitoring of the implementation of the HSK recommendations was initiated by the BMBWF and met with broad support from the universities. Almost all Austrian HEIs participated in the survey on the status of implementation of the recommendations on gender competence in higher education processes (participation rate 93%). The results of the survey show a wide range of implementation activities and individual priorities at the universities. They illustrate the differences between the higher education sectors resulting from the different framework conditions (e. g. legal requirements vs. recommendations).

The project commissioned by the BMBWF (Wroblewski, Englmaier 2023) implemented input-side monitoring for the year 2022. Feedback from the workshops has shown that it is currently not very realistic for universities to undertake monitoring and evaluation of gender equality measures on their own initiative. Rather, it is expected that the initiative will come from the BMBWF and that resources will be made available. It was also clear from the workshops that the limited commitment of management to implementing the HSK recommendations and their lack of embedding in existing control mechanisms hinders the use of the steering potential of monitoring.

The design and implementation of the monitoring was carried out in a complex and participatory process that aimed to create transparency, build competence among the actors involved and establish a common understanding of the

objectives of the monitoring and the indicators. This was largely successful and the project contributed to a cross-sectoral discourse on gender competence and gender equality. However, it has also become clear that this is a process that needs to continue in a planned and facilitated way.

A key aspect of the discourse on gender competence and gender equality is that the indicators implicitly set standards for the evaluation of the measures implemented. Establishing a common understanding of these standards would be an important objective for the continuation of the process. This would include making the concepts used in the recommendations more concrete. In the survey, but also in the workshops, it became clear that some of the concepts used are operationalised differently when it comes to the development of concrete measures. These include, for example, gender, gender competence, gender diversity or gender mainstreaming, but also the relationship between gender and gender competence and diversity.

REFERENCES

BMBWF (Hrsg.) (2018), Verbreiterung von Genderkompetenz in hochschulischen Prozessen. Empfehlungen der Hochschulkonferenz, Vienna. Online available: https://www.bmbwf.gv.at/Themen/HS-Uni/Gleichstellung-und-Diversit%C3%A4t/Aktuelles/Empfehlungen-der-Hochschulkonferenz-zur-Verbreiterung-von-Genderkompetenz-in-hochschulischen-Prozessen.html [30.11.2023].

BMWFW (2016), Austrian ERA Roadmap, Vienna. Online available: https://era.gv.at/era/era-2000-2021/era-roadmap/austrian-era-roadmap/ [30.11.2023].

Criado-Perez, Caroline (2019), Invisible Women: Data Bias in a World Designed for Men, New York: Abrams Press.

D'Ignazio, Catherine; Klein, Lauren F. (2020), Data Feminism, Cambridge: The MIT Press.

Döring, Nicola (2013), Zur Operationalisierung von Geschlecht im Fragebogen: Probleme und Lösungsansätze aus Sicht von Mess-, Umfrage-, Gender- und Queer-Theorie, in: GENDER – Zeitschrift für Geschlecht, Kultur und Gesellschaft (5): 94-113. Online available: https://www.ssoar.info/ssoar/handle/document/39660 [30.11.2023].

Eckstein, Kristin (2017), Gleichstellungsindikatoren an Universitäten – von der Berichterstattung zur Steuerung, in: Wroblewski, Angela; Kelle, Udo; Reith, Florian (2017), Gleichstellung messbar machen, Wiesbaden: Springer VS: 149-169.

Hedman, Brigitta; Perucci, Francesca; Sundström, Pehr (1996), Engendering statistics: A tool for change, Stockholm: Statistics Sweden. Online available: https://ams-forschungsnetzwerk.at/downloadpub/gender_1996_statistics_601.pdf [30.11.2023].

Lipinsky, Anke; Wroblewski, Angela (2021), Revisiting Gender Equality Policy and the Role of University Top Management, in: O'Connor, Pat; White, Kate (Hrsg.), *Gender, Power and Higher Education in a Globalised World*, Palgrave Studies in Gender and Education: 163-186. Online available: https://doi.org/10.1007/978-3-030-69687-0_8 [30.11.2023].

Löther, Andrea; Steinweg, Nina; Lipinsky, Anke; Meyer, Hannah (2021), Gut oder unzureichend?: Wirkung und Wirkungsmessung von gleichstellungspolitischen Maßnahmen, in: Forschung & Lehre (3): 188-189.

Markiewicz, Anne; Patrick, Ian (2016), Developing Monitoring and Evaluation Frameworks. L.A. et al.: SAGE.

May, Judith V.; Wildavsky, Aaron B. (Hrsg.). (1978), The Policy Cycle, Beverly Hills/London: Sage Publications.

Meyer, Wolfgang (2004), Indikatorenentwicklung: eine praxisorientierte Einführung (CEval-Arbeitspaper, 10), Saarbrücken: Universität des Saarlandes, Fak. 05 Empirische Humanwissenschaften, CEval – Centrum für Evaluation. Online available: https://www.ssoar.info/ssoar/handle/document/11124 [30.11.2023].

Rossi, Peter H.; Freeman, Howard E.; Lipsey, Mark W. (1999), Evaluation. A Systematic Approach, Thousand Oaks/London/New Dehli: SAGE.

W. K. Kellogg Foundation (2004), W. K. Kellogg Evaluation Handbook, Battle Cree. Online available: https://www.betterevaluation.org/sites/default/files/2022-07/EvaluationHandbook.pdf [30.11.2023].

Wroblewski, Angela (2017), Gender-Indikatoren in der Wissensbilanz – Grundlage für ein Gleichstellungsmonitoring oder Datenfriedhof?, in: Wroblewski, Angela; Kelle, Udo; Reith, Florian (2017), Gleichstellung messbar machen, Wiesbaden: Springer VS: 171-189.

Wroblewski, Angela (2021), Monitoring of ERA priority 4 implementation – update and final assessment, WP3 GENDERACTION report, Prag. Online avail-

able: https://genderaction.eu/wp-content/uploads/2022/03/741466_D3_3_Monitoring-of-ERA_web.pdf [30.11.2023]

Wroblewski, Angela; Palmén, Rachel (2022), A Reflexive Approach to Structural Change, in: Wroblewski, Angela; Palmén, Rachel (Hrsg.), Overcoming the Challenge of Structural Change in Research Organisations: A Reflexive Approach to Gender Equality, Emerald Publishing Limited: 15-32. Online available: https://library.oapen.org/handle/20.500.12657/57559 [30.11.2023].

Wroblewski, Angela; Kelle, Udo; Reith, Florian (2017), Gleichstellung messbar machen: Grundlagen und Anwendungen von Gender- und Gleichstellungsindikatoren, Wiesbaden: Springer VS.

Wroblewski, Angela; Leitner, Andrea (2022), Relevane of Monitoring for a Reflexive Gender Equality Policy, in: Wroblewski, Angela; Palmén, Rachel (Hrsg.), Overcoming the Challenge of Structural Change in Research Organisations: A Reflexive Approach to Gender Equality, Emerald Publishing Limited: 33-52. Online available: https://library.oapen.org/handle/20.500.12657/57559 [30.11.2023].

Wroblewski, Angela; Lipinsky, Anke (2018), Steering by Evaluation in Times of Impact Orientation and Monitoring?, in: fteval: Journal for Research and Technology Policy Evaluation (45): 22-27. Online available: http://irihs.ihs.ac.at/4836/. https://www.fteval.at/content/home/journal/aktuelles/ausgabe_45/fteval_Journal45_WEB.pdf.

Wroblewski, Angela; Englmaier, Victoria (2023), Monitoring der Umsetzung der HSK-Empfehlungen zur Verbreiterung von Genderkompetenz in hochschulischen Prozessen, Studie im Auftrag des BMBWF, Wien: Institut für Höhere Studien.

AUTHOR

ANGELA WROBLEWSKI
IHS – Institute for Advanced Studies, Josefstädter Straße 39
1080 Vienna, Austria
Email: wroblews@ihs.ac.at
Orcid: https://orcid.org/0000-0002-1233-4409

fteval JOURNAL
for Research and
Technology Policy Evaluation

April 2024, Vol. 55, pp. e3, 1-11
DOI: 10.22163/fteval.2024.635

ENHANCING EVALUATION TENDER QUALITY: INTEGRATING GENDER FOR ROBUST RESULTS

KARIN GRASENICK
DOI: 10.22163/FTEVAL.2024.635

ABSTRACT

Evaluations play a key role in assessing the effectiveness of strategies, programmes and projects and drawing conclusions for their future development. Standards have been introduced to improve the quality, focusing mainly on the evaluators. However, essential conditions for the evaluation are specified in advance, particularly in the calls for proposals. Tenders are expected to clearly communicate requirements and expectations, ensuring that proposals can be effectively compared, while allowing evaluators sufficient flexibility in their design and interpretation. However, evaluation standards often lack the integration of gender and diversity considerations, particularly in the initial stages such as in calls for proposals. This article assesses the inclusion of gender analysis and intersectionality within evaluation standards, with a focus on tender processes. The necessity for a systematic approach to embed gender and diversity considerations in evaluation practices is discussed, highlighting the gap in current standards. The article references a guideline developed by the Working Group on Gender Mainstreaming (AK GM) of the German Evaluation Society (DeGEval) as an example of efforts to address this gap. It exemplifies a framework aimed at guiding evaluation practitioners and those issuing tenders towards more inclusive evaluation practices. In this way, evaluations not only comply with legal requirements, but also fulfil societal commitments to promote gender equality, especially in the field of research, technology and innovation.

INTRODUCTION

Evaluations are instrumental in determining the success and effectiveness of strategies, programmes, or projects. The quality of evaluations is - among others - dependent on the requirements and standards established during the tendering process. A crucial aspect of this process is the recognition of stakeholders as diverse groups, each with specific needs and perspectives. It is essential to understand and address the varied impacts of a strategy or intervention on these different stakeholder subgroups. This involves a detailed approach to stakeholder engagement, careful data collection and analysis, and transparent reporting of potential discrepancies. Such a methodology starts by considering gender as important variable and thereby not only ensures evaluations of higher quality but also aligns with legal requirements and supports societal commitments to promoting gender equality.

The importance of gender-specific analysis in evaluations is emphasised by various frameworks, also relevant for research, technology and innovation. Already in 1979 the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW 1979, see also https://www.un.org/womenwatch/daw/cedaw) mandated the integration of gender perspectives in all policy areas, including research, technology, and innovation. In Europe frameworks relevant for gender-specific analysis in research and innovation include the European Commission Gender Equality Strategy 2020-2025 and the mandatory Gender Equality Plans (GEPs) for organisations applying for funding under Horizon Europe (see also https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/democracy-and-rights/gender-equality-research-and-innovation_en). The shift from policy formation to practical application underscores the necessity for robust evaluation mechanisms. Evaluations that incorporate gender perspectives ensure decisions are made considering the diverse needs and experiences of all genders, leading to more equitable policies and programmes (Wroblewski, 2019).

The “Gender-and diversity-sensitive tendering of evaluations – a guideline” (GeD-Call-Eval, 2023) offers a foundational framework for incorporating gender and diversity considerations into evaluation processes. This guideline underscores the significance of integrating gender and diversity as central quality criteria in tenders, proposals, and resulting evaluations. By doing so, evaluations should become more comprehensive, socially responsible, and impactful. They aim to provide a nuanced understanding of programme effectiveness and societal impact, potentially leading to more informed decision-mak-

ing and policy development that genuinely reflects and meets the needs of a diverse society.

By emphasising the importance of gender and diversity in evaluation tenders, the guideline addresses a crucial gap. It advocates a shift in evaluation practices, urging the integration of gender and diversity as central quality criteria in all stages of evaluation. The guideline thus plays a pivotal role in fostering comprehensive evaluations that capture the diverse experiences and needs of all involved stakeholders regardless of their power, direct involvement or indirect effected by a development or intervention.

EVALUATION STANDARDS FOR RESEARCH AND TECHNOLOGY POLICY EVALUATION

Evaluation societies are professional networks for organisations tendering evaluations, researchers, and practitioners. These organisations are typically formed to promote and advance the field of evaluation, and as such often establish and promote quality standards. The standards of the evaluation organisations take gender into account to varying degrees. For example, in the standards of SEVAL (Evaluation Society of Switzerland, 2016) gender is only included in the explanations on ethical aspects. The standards set by the Austrian Platform for Research and Technology Policy Evaluation (fteval, 2019) emphasise a comprehensive approach to RTI policy evaluation, integrating principles like joint understanding, commitment, participation, utilisation and benefits, feasibility, impartiality, independence, professionalism, and gender. A notable aspect of the fteval standards is their explicit focus on gender. They mandate the inclusion of gender-specific questions in the Terms of Reference (ToR), require the collection and interpretation of gender-specific data, and insist that the language used for reporting should be gender neutral. This means avoiding gender-specific language by using gender-neutral terms or including at least masculine and feminine forms. Additionally, they explicitly state that *"If there is no apparent gender dimension in the evaluation object, this fact should be indicated and explained".* This aspect ensures that gender considerations are not overlooked and are explicitly addressed, whether they are present or absent in the context of the evaluations. By outlining what constitutes professional and competent practice in RTI policy evaluation the fteval standards, aim at both, evaluators and those who tender evaluations in research, technology, and innovation (RTI) policy evaluation. **For organisations or individuals tendering**

evaluations, they provide a basis to ensure that the evaluations they commission are comprehensive, ethical, and consider gender. However, while the fteval standards acknowledge the relevance of integrating gender considerations into the evaluation of RTI policies as crucial for enhancing quality and maintaining relevance they do not provide detailed instructions on how to incorporate these aspects into the tendering process of evaluations.

THE GUIDELINE FOR GENDER AND DIVERSITY SENSITIVE TENDERING AND ITS RELATION TO EVALUATION STANDARDS

This is where the guidelines for "gender and diversity sensitive tendering" (GeD-Call-Eval-Guideline) fill a critical gap by offering practical steps and examples for including gender and diversity considerations in the evaluation process, right from the tendering stage. This ensures that the expectations and requirements around these aspects are clearly communicated to potential evaluators. While not explicitly focusing on technology and innovation, they offer a framework for systematically incorporating these aspects into evaluation tenders and include a template for systematically considering gender and diversity specific analysis Terms of Reference (ToR), which covers the following sections:

- **Object of Evaluation:** This section details general and specific objectives, time frames, phases, and locations. It offers insights into the cultural and social context, highlighting gender dimensions relevant to the evaluation's scope. Further, users of the ToRs are reminded to carefully identify the various stakeholders and their roles and influences on the evaluation object, such as decision-makers (e.g., ministries, supervisory boards), implementers (e.g., agencies, partner organizations), target groups/beneficiaries (primary audience), and unintentionally affected persons (indirectly impacted groups).

- **Underlying Understanding of Gender and Diversity:** According with DeGEval and fteval standards the ToR request that a tender Defines and explains how gender and diversity are conceptualized within the context of the evaluation object.

- **Aim, Purpose and Scope of the Evaluation:** These sections within ToR too, should be clearly set the expectations for gender- and diversity-

sensitive evaluation outcomes. This involves assessing how different genders or demographic groups are affected and ensuring all needs are addressed, aiming for equitable and effective evaluation results. Potential evaluation teams should be reminded that all relevant data and groups, especially marginalized ones, are considered in the methodological approach. To enable evaluators to develop a suitable design, potential constraints such as resource limitations, data availability, and stakeholder interactions must be outlined as early as possible.

- **Methodology and Evaluation Team**: To offer a gender and culturally sensitive methodological design the evaluation team must possess diverse skills and backgrounds, including gender, equality, and intersectionality expertise, with at least one member specializing in gender. – and the tenderers must request such expertise.

- **Evaluation Criteria**: the ToR included suggestions how to consider gender and diversity when applying standards like DeGEval or OECD-DAC to evaluate tenders.

It is important to emphasise that these guidelines are not intended to provide uniform definitions of gender, diversity and intersectionality or inclusion. Rather, it is the responsibility of the tenderers to clarify which definitions guide their actions and to demand this from the bidders. For those not familiar with the relevant gender and diversity research, the guidelines provide short explanations and further references. It refers to the evolving understanding of gender as a non-binary concept and the importance of intersectionality and broader perspective on diversity. Intersectionality, a term coined by Kimberlé Crenshaw (Crenshaw 1989), refers to the interconnected nature of social categorisations such as race, class, and gender, which can lead to overlapping and interdependent systems of discrimination or disadvantage. This broader understanding is crucial in evaluations, ensuring that programmes and projects do not inadvertently exclude or misrepresent non-binary and transgender individuals or marginalised groups (Kubdqvist et.al 2019; UN Women 2018)

This GeD-Call-Eval-Guideline was aligned with the DeGEval standards for evaluations (DeGEval, 2016), which emphasise best practices in evaluation methodology and ethics. It reflects several key DeGEVal standards, including “G3 – Description of Purposes and Procedures”, which underlines the importance of clearly defining the evaluation’s objectives and methods. Additionally, “N1 – Identification of Involved and Affected Parties” underscores the need to consider the perspectives and needs of all stakeholders, a principle echoed in this guideline. The guideline also aligns with “G8 – Substantiated Assessments and

Conclusions", emphasising the analysis of how evaluation subjects contribute to gender equality and diversity. By adhering to these standards, the guideline emphasises that Gender and Diversity are integral components of the evaluation process, promoting inclusivity and rigorous evaluation practices.

The explanatory introduction can easily be adapted to explain how gender and diversity should be considered for tendering RTI evaluations by referring to the fteval standards which already mandate the inclusion of gender-specific questions, data collection and interpretation. The following paragraphs relate the original explanation to the fteval standards accordingly:

1. **Clearly defined aim and purpose of the evaluation:** This corresponds to the fteval principle "Utilisation and Benefits" (Section 5.4). The recommendation emphasises the need to describe the understanding of gender and diversity in the measure. The ToR should clearly describe the object of the evaluation and the specific objectives and purposes. The ToR should also address the understanding of gender and diversity with which the evaluation object was developed and/or implemented.

    a. **Differentiated consideration of participants and those affected:** This corresponds to the fteval principle "Participation" (Section 5.3). The specific perspectives and needs of those directly and indirectly affected must be presented in a differentiated manner. All social and gender groups should be named or included.

    b. **Analysis of effectiveness:** What contribution do evaluation objects make to the promotion of equal opportunities and gender equality? This corresponds to the fteval principle "Feasibility" (Section 5.5).

    c. If evaluators are unable to establish a gender and/or diversity reference, this must be explicitly justified.

2. **Time frame and resources:** This corresponds to the fteval principle "Professionalism" (Section 5.8). The ToR should contain a clear timetable for the implementation of the evaluation as well as information on the available resources, such as financial resources, personnel, and data – this applies in particular to resources for sufficient consideration of the gender dimension as a quality criterion.

3. **Indication of the methodological approach:** This corresponds to the fteval principle "Professionalism" (Section 5.8) and "Transparency" (Section 5.10). The ToR should contain information on which evaluation methods are expected. This includes methods and tools for considering gender and diver-

sity: evaluators should explain how they will incorporate gender and, where appropriate, other relevant dimensions into their methodological approach and how gender and diversity sensitivity will be taken into account in the evaluation.

4. **Clear information on the available database:** This corresponds to the fteval principles "Credibility" (Section 5.12) and "Transparency" (Section 5.10). It is important to explain the extent to which the available sources and data enable a differentiation of the gender dimension and which data gaps need to be taken into account.

5. **Clearly defined responsibilities, roles, required expertise:** The ToR should clearly define the responsibilities of the actors involved, including the evaluators, commissioning parties, and other stakeholders.

    a. **Commitment to consider gender and diversity:** This is in line with the fteval principle "Ethics" (Section 5.11) and the legal basis, which is why tenderers and evaluators should make their commitment to gender-sensitive and diversity-conscious evaluation clear.

    b. **Expertise and experience:** This corresponds to the fteval principle "Professionalism" (Section 5.8). Evaluators should demonstrate their competence and experience in gender-sensitive, inclusive, and diversity-conscious evaluation. This can be done through references, case studies, or training in this area.

    c. **Inclusion of gender and diversity in the evaluation team:** When selecting the evaluation team and conducting the evaluation, a diverse composition should be sought that takes into account different genders and experiences. This should ensure that different perspectives are brought in and that a comprehensive assessment takes place.

6. **Communication and reporting:** This corresponds to the fteval principles "Transparency" (Section 5.10) and "Utilisation and Benefits" (Section 5.4). The ToR should define the expectations regarding communication and reporting during and after the evaluation. This includes the type of interim reports, the timing of reporting, and the target groups for the communication of results.

The GeD-Call-Eval-Guideline includes on pages 6–13 a template for tendering organisations (Grasenick et al 2023). It aims to cover all essential aspects like objectives, methodologies, expected outcomes, and stakeholder engagement. This ensures a thorough approach to tendering, including gender and diversity

considerations. Generally, the template is sector independent as the guiding questions on how to consider gender and diversity are relevant across various fields. However, the template will always need to be adapted to specific terminologies and challenges. The key is to maintain the essence of gender and diversity sensitivity while making it relevant to the specific context of the sector.

CRITICAL REFLECTION

The GeD-Call-Eval-Guideline was presented and discussed at the DeGEval annual conference 2023 in a focus group setting attended by 6 evaluators from various sectoral backgrounds. Additionally, 4 evaluators provided written feedback directly in the first version of the document or via email. Overall, all evaluators were members of the DeGEval, and 5 evaluators were members of the AK-GM. The feedback can be summarised as follows:

1. The GeD-Call-Eval-Guideline primarily focuses on the practical application of gender and diversity considerations for tendering evaluations. However, a more profound exploration into the theoretical basis of these concepts could significantly enhance its depth and applicability.

2. The GeD-Call-Eval-Guideline, especially the template for ToR, is structured in a way that prompts clear and detailed responses. However, it could benefit from more explicit instructions or practical examples to guide users, especially those less familiar with gender and diversity issues on how to practically implement these principles. While the template is comprehensive, it will need to be adapted to different sectors, like e.g. research and technology. Moreover, the broad scope of intersectionality is complex and difficult in its practical application.

3. The GeD-Call-Eval-Guideline is clearly linked with the DeGEval Evaluation Standards However, some sectors refer to different standards. This could be addressed and further, internationally recognised evaluation standards could be incorporated, such as those set by the OECD.

CONCLUSION AND RECOMMENDATIONS FOR FURTHER RESEARCH

This article has demonstrated that the GDT-EVAL guidelines align well with the fteval standards, while complementing the scope to include tendering. They extend the fteval standards by incorporating a non-binary understanding of gender, recognising intersectionality, and broadening the scope of diversity to ensure more inclusive evaluations. This is particularly important as the technology and innovation sectors are generally less informed about gender bias and often assume that programmes and projects are gender neutral. However, implementation challenges may arise, particularly in contexts less familiar with these issues.

The feedback received at the DeGEval Annual Conference has so far only partially been considered. The next version should expand the theoretical background, add more references to international standards, as well as incorporate case studies or examples, e.g. from the technology sector, to further enhance the usability of the guideline. These adaptations could significantly support policies and practices in promoting more equitable and socially responsive RTI outcomes.

Future research should focus on the practical implementation of the guidelines and assess their impact on the inclusiveness and effectiveness of RTI evaluation tenders. This is essential for the development of more bias-conscious, equitable and inclusive RTI policies, programmes and projects.

REFERENCES

Austrian Platform for Research and Technology Policy Evaluation (2019): Evaluation Standards for Research, Technology and Innovation Policy https://fteval.at/standards/ checked on 17 April 2024

Crenshaw, Kimberlé (1989): "Demarginalising the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." University of Chicago Legal Forum.

DeGEval - Society for Evaluation e.V. (2016): Standards for Evaluation, last checked on 17 April 2024. https://www.degeval.org/degeval-standards/

Dill, Katja and Völkle, Hanna (2023): Zur Gratwanderung Intersektionalität und Technikgestaltung Discussion Paper No. 47, 02/2023 des Harriet Taylor Mill-Institut für Ökonomie und Geschlechterforschung der Hochschule für Wirtschaft und Recht Berlin ISSN 1865-9806 Download: https://www.htmi.hwr-berlin.de/publikationen/discussion-papers/

Eckstein, Kirstin; Wroblewski, Angela (2017): Gender Mainstreaming and Evaluation 16 (2), pp. 267–269. Available online at https://www.waxmann.com/artikelART102267.

European Agency for Fundamental Rights: https://fra.europa.eu/en/theme/anti-racism/compendium-practices/about-compendium

European Institute for Gender Equality: https://eige.europa.eu/ in particular the explanations on "gender responsive evaluation", which also contains further literature https://eige.europa.eu/gender-mainstreaming/tools-methods/gender-responsive-evaluation

Grasenick, Karin; Wroblewski, Angela; Rodenberg, Birte; Kohlweg, Karin; Eckstein, Kirstin; Willig, Michaela C.; Staeck, Nicola; Kiel, Petra; Neymeyer, Susanne (2023, 2024): Gender- und diversitätssensible Ausschreibungen von Evaluationen - ein Leitfaden https://repository.fteval.at/id/eprint/704/

Gutknecht-Gmeiner, Maria; Wroblewski, Angela (2015): Gender competence of evaluators. Position paper of the Gender Mainstreaming Working Group of DeGEval, last checked on 17 April 2024.

Heinrich Böll Foundation: Words have power. A lexicon on international LGBTIQ+ activism, gender and sexuality (2023): https://www.boell.de/sites/default/files/2023-05/sprache_wissen_lexikon_intersektionalem_lgbtiq_aktivismus_geschlecht_sexualitaet.pdf

Lindqvist A., Gustafsson S´M. Renstroem E. (2019): What is gender, anyway: a review of the options for operationalising gender https://www.tandfonline.com/doi/full/10.1080/19419899.2020.1729844

OECD *Applying a Human Rights and Gender Equality Lens to the OECD Evaluation Criteria:* https://www.oecd-ilibrary.org/development/applying-a-human-rights-and-gender-equality-lens-to-the-oecd-evaluation-criteria_9aaf2f98-en

Swiss Evaluation Society SEVAL (2016): Evaluation Standards of the Swiss Evaluation Society Adopted (in German and French) by the General Assembly on 9 September https://www.seval.ch/en/standards-competences/standards/

United Nations (UN) Entity for Gende Equality and the Empowerment of Women (1979): Convention on the Elimination of all Forms of Discrimination against Women https://www.unwomen.org/en

United Nations (UN) Women (2018): Inclusive Systemic Evaluation for Gender equality, Environments and Marginalized voices (ISE4GEMs): A new approach for the SDG era https://www.unwomen.org/en/digital-library/publications/2018/9/ise4gems-a-new-approach-for-the-sdg-era

Wroblewski, A. (2019): Gender Mainstreaming und die DeGEval-Standards in: Jan Ulrich Hense, Wolfgang Böttcher, Michael Kalman, Wolfgang Meyer (Hrsg.) Evaluation: Standards in unterschiedlichen Handlungsfeldern Einheitliche Qualitätsansprüche trotz heterogener Praxis? Waxmann Lehrbuch

AUTHOR

KARIN GRASENICK
convelop
cooperative knowledge design gmbh
Kaiserfeldgasse 7/II
8010 Graz
Email: karin.grasenick@convelop.at
ORCID: https://orcid.org/0000-0002-3130-0773

fteval JOURNAL
for Research and
Technology Policy Evaluation

November 2024, Vol. 55, pp. e4, 1-35
DOI:10.22163/fteval.2024.660

THE INTERSECTION OF AI AND RTI (POLICY) EVALUATION: PRINCIPLES AND CONSIDERATIONS

AUTHORS: STEFANIE SCHUERZ, CHARLOTTE D'ELLOY, MICHAEL DINGES

CONTRIBUTORS: SABINE CHAI, THOMAS PALFINGER, MICHAEL STRASSNIG, ISABELLA WAGNER, MAGDALENA WICHER
DOI: 10.22163/FTEVAL.2024.660

PREAMBLE

In May 2023, the Austrian Platform for Research and Technology Policy Evaluation (fteval) established a working group on Artificial Intelligence (AI). The working group members agreed to split into three subgroups with distinct tasks: The first subgroup would collect proprietary AI systems relevant to the evaluation phases and activities. The second subgroup would consider the evaluation system and model of interaction between its actors, as well as the possible effects that can arise from AI, and make reasonable recommendations for a strategy. The third subgroup would screen existing guidelines on the use of AI and adapt them to the fteval community, integrating good practice examples and including input and feedback from the other two groups. The final deliverables of the working group are three papers – one for each subgroup – to be shared as resources with the fteval community via the fteval journal. The working group members met between November 2023 and January 2024 to share their progress and discuss questions that had arisen during the subgroups' work. The document below is the outcome of subgroup 3 on guidelines, supplemented by information elaborated in subgroup 1 on AI systems, and intends to provide comprehensive guidance to the fteval community.

INTRODUCTION

As the digital transformation changes our society at an ever-accelerating pace, Artificial Intelligence (AI) promises to impact organisational and business processes, boost productivity, and support the analysis and visualisation of large amounts of data in an unprecedented manner. Generative AI (GenAI) systems and services show staying power beyond media hype and are in the process of establishing themselves as integral parts of research, technology, and evaluation work. At the same time, many important questions remain. Within the fteval working group on AI, we discussed various difficulties we encountered, including: How do we effectively employ AI systems while not negatively impacting the quality and rigorousness of our work? How do we (need to) disclose the specifics of our employment of AI systems in the context of our work and whom do we (need to) disclose this to? In other words, what standards of use and reporting do we need in order to integrate AI systems in our work? And how may we make informed decisions on all of these questions?

Against this background, this document is structured as follows: After an overview on the employed methodology, it provides working definitions for the specific types of AI covered in this paper and delineates the scope of AI subfields to establish a joint understanding on the multitude of AI application areas. It then outlines principles and considerations for effectively integrating AI into RTI evaluation contexts, ensuring informed decision-making and maximising the benefits of AI-driven approaches. Finally, the document describes the diverse application areas of AI within the context of research, technology and innovation (policy) evaluation, offering some insights into its multifaceted roles in daily evaluation practices of the fteval community. Annex I provides a translation of a survey sent to the fteval community, while Annex II entails a list of proprietary AI systems, including potential areas of application, pricing models, available languages, and possible alternatives.

The document focuses mostly on practical and ethical considerations of employing AI systems and solutions, not for their development and deployment. The use of AI systems for evaluating scientific research proposals and the use of AI processes for student or teacher evaluation in educational contexts are also out of scope for this paper. The main target group are evaluation practitioners working as evaluators, programme owners and managers in the Austrian RTI community.

METHODOLOGY

The present Principles and Considerations have been developed by a working group established in the context of the fteval community shortly after the release of ChatGPT-4 with the aim of exploring the impact of this highly disruptive technology on the Austrian RTI community. Split into three thematic subgroups, various aspects of this emergent cluster of technologies were explored, including a mapping of specific systems, changing roles within the RTI evaluation ecosystem, as well as existing principles and guidelines and their applicability to our RTI evaluation context. In both subgroup and plenary meetings, study design, methodologies and findings were discussed and aligned to create several documents, including the one at hand.

For these Principles and Considerations, the work plan started with a systematic literature review of existing AI guidelines from RTI, business and policy contexts. These were reviewed and core information was extracted according to a predefined set of questions to synthesise the principles and considerations laid out below. Specifically, the following information was extracted from each document:

- Country of origin, publishing organisation, year of publication
- Core focus of the text: Development or use of AI?
- What definition(s) of AI are used (explicitly and implicitly)
- Which topics on AI are addressed (e.g., trustworthiness of AI, etc.)
- Which processes and tasks are addressed?
- Which problems / problem areas are addressed?
- Which solutions are suggested?
- What topic areas are (still) missing?

Additional information was sought through existing publications to fill gaps in understanding and information. Furthermore, a short survey was developed and shared within the fteval community, to gather inputs on challenges, information gaps, attitudes, practices, as well as existing institutional guidelines regarding the use of AI within fteval member organisations. The survey employed both open and closed question formats (see Annex I) and led to a total of 20 valid responses representing Austrian RTI, RTI funding, and RTI policy organisations. Findings from this survey were used to validate and enrich the

Principles and Considerations. Finally, the first draft of the text was shared with the fteval community to review the contents and allow for adjustment. The concluding meeting of the fteval AI working group on March 15., 2024, allowed for a final discussion of the paper, which was subsequently complemented with insights from the second subgroup – mapping AI systems – to round out the text for publication.

WORKING DEFINITIONS

AI spans multiple evolving and often overlapping domains, making it difficult to identify what does and does not qualify as AI in any given context: "Most definitions of AI centre on the concept of emulating intelligent behaviour through machines, where intelligence refers to the capacity to perform complex tasks in real-world environments and learn by experience. […] definitions of AI converge in their focus on machines or AI systems that (i) possess learning capabilities, (ii) can make intelligent decisions, (iii) influence the environment, (iv) improve tasks autonomously, and (v) exhibit human-like cognitive functions." (EC DG RTD 2023: 31)

These systems are capable of learning from data, recognise patterns, make decisions, and solve problems without explicit programming for each task. Building on a definition of the US Department of Defense, the Austrian Council for Robotics and Artificial Intelligence (Österreichische Rat für Robotik und Künstliche Intelligenz – ACRAI) characterises them as autonomous cognitive systems. They work through rule knowledge created by experts or on the basis of statistical models derived from data (machine learning, e.g. deep learning). AI systems are based upon extensive training data to analyse and identify correlations or patterns, enabling them to make predictions. This training process enhances their ability to produce improved, precise, and realistic responses, or to generate actionable insights. (ACRAI 2018)

Following Regona et al. (2022), in terms of AI research objectives, AI comprises major subfields such as machine learning, knowledge-based systems, computer vision, robotics, natural language processing, automated planning and scheduling, and optimisation. Figure 1 showcases the breadth of fields and understandings connected to AI and outlines their various components.

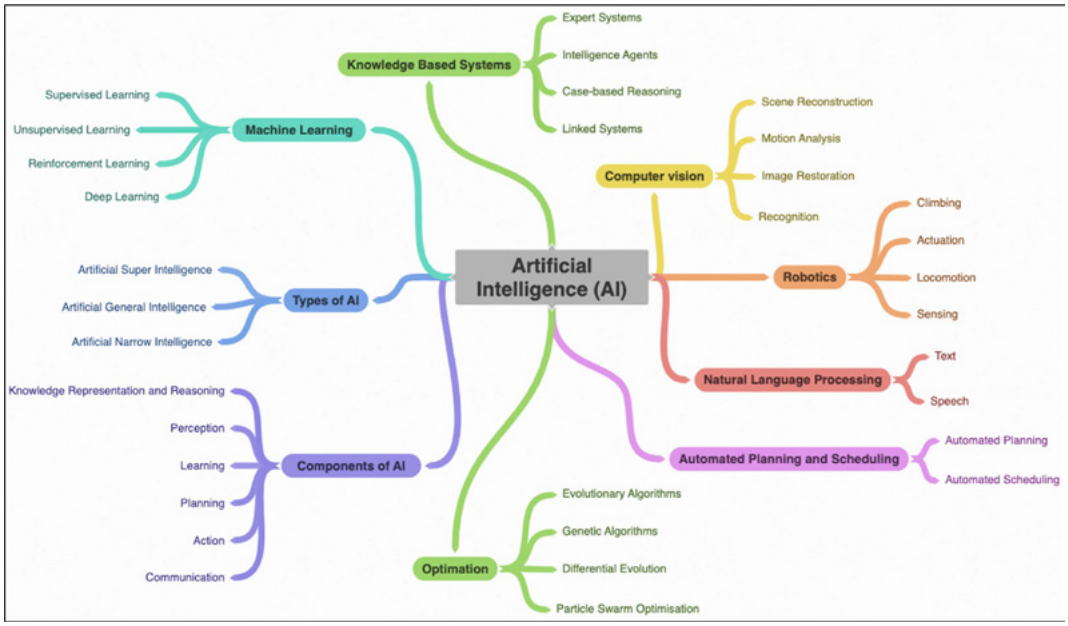


Figure 1: Components, types, and subfields of AI. (Regona et al. 2022)

Among the subfields presented above, the following are of potentially high relevance for evaluation purposes:

Machine Learning (**ML**) focuses on the development of algorithms that enable computers to learn from and make predictions or decisions based on datasets, without being explicitly programmed. Rather, they are build to receive input and predict outputs via statistical analysis, based on the respective training data. ML includes supervised learning, unsupervised learning, deep learning and reinforcement learning. (Pant 2019)

Natural Language Processing (**NLP**) is concerned with the interaction between computers and humans via "natural" human language. It involves tasks such as speech recognition, language translation, sentiment analysis, and text generation, enabling machines to interpret, generate and simulate human language. (Nancholas 2023)

Knowledge based systems (**KBS**) or **Expert Systems** (**ES**) are trained on high-level, domain-specific knowledge from human experts and are intended to mimic their decision-making abilities. They use knowledge bases, inference engines, and reasoning mechanisms to provide advice, solve problems, and make decisions. (Segreto 2019)

For this paper, we primarily focus on **Generative AI** or **GenAI**, which refers to "a type of machine learning architecture that uses AI algorithms to create novel data instances, drawing upon the patterns and relationships observed in the

training data" and is thus "capable of generating seemingly new, meaningful content such as text, images, or audio from training data." (Feuerriegel et al. 2024) These new data points are generated in such a way that they "plausibly could be part of the original dataset". (Pinaya et al. 2023)

PRINCIPLES AND CONSIDERATIONS

In reviewing the relevant current literature on AI and its application in contexts adjacent to RTI, a number of principles are brought up repeatedly, intended to ensure a responsible use of AI systems that minimises harm while exhausting their potential. The following section will discuss these principles in detail.

RESPECT FOR HUMAN AUTONOMY

Various high-level guidelines on AI stress the importance of maintaining human autonomy when employing AI systems (see e.g. EC DG CNCT 2019, EC 2020, APA 2022, and OECD 2024[2019]). This means ensuring human self-determination in the sense that AI systems do not "unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans" (EC DG CNCT 2019: 12). Following Prunkl (2022), autonomy has an internal dimension, authenticity, which refers to beliefs, values and motivations that are an authentic reflection of a person's 'inner self'; and an external dimension, agency, which refers to a person's effective capacity to make and enact decisions and take charge of important aspects of their lives. For authenticity to be preserved, AI systems must not be used to manipulate or deceive people. For agency to be preserved, humans must maintain control over whether, when and how decisions concerning their lives are relegated to AI systems.

RESPONSIBILITY

When using AI systems, humans remain responsible for the accuracy, transparency, and accountability of outputs created with AI support and shared in a professional context. AI systems can serve as a valuable research tool, but they do not replace critical thinking, human expertise, and rigorous scientific methodology. Care needs to be taken to critically review the outputs created by an AI with a researcher's own expertise, to check for correctness, ethical aspects, and plausibility.

Research results derived from the analysis of AI systems should not be considered as a sole source of information for developing conclusions or recommendations. For developing these, multiple data sources, methods, and perspectives should be taken into account in order to create comprehensive and robust evaluation results. A triangulation of methods and results helps to identify inconsistencies and conflicting viewpoints and helps mitigate potential limitations of AI systems in terms of bias, limited scope of analysis, and challenges regarding explainability. Individuals should be able to understand and challenge the outcome while respecting personal data protection obligations, if relevant.

TRANSPARENCY AND EXPLAINABILITY

The Evaluation Standards for Research, Technology and Innovation Policy demand that "the evaluation process, its findings, and the subsequent recommendations are conducted and completed in a manner that is transparent and accountable for all those involved and affected." (Kohlweg 2019: 15) This also holds true when AI systems are employed in evaluation processes, meaning there is a need to be transparent about the use of AI systems and its purposes.

Transparency is particularly important when 1) AI applications are being used for analytical purposes; 2) reproducibility of results is required; and 3) the output created is likely to have an impact on decision-making processes. The OECD notes that "disclosure [of AI use] should be made with proportion to the importance of the interaction. The growing ubiquity of AI applications may influence the desirability, effectiveness, or feasibility of disclosure in some cases." (OECD 2024)

Following the OECD AI Principles, explainability pertains to "enabling people affected by the outcome of an AI system to understand how it was arrived at. This entails providing easy-to-understand information to people affected by an AI system's outcome that can enable those adversely affected to challenge the outcome, notably – to the extent practicable – the factors and logic that led to an outcome." (OECD 2024)

Tacit and explicit standards for clarifying which systems were used, for what purposes and at which point of the process will most likely be negotiated and settled in the upcoming years, depending on which systems and AI use cases are evolving. The following measures might be taken to ensure transparency in the evaluation process:

- Clearly state in the methods section which AI systems have been used for which purposes of the evaluation. Specify the respective system and relevant parameters used.
- When using AI systems for analytical purposes, we suggest keeping a record of AI-generated outputs including prompts, parameters, and outputs created by the system. Keeping a record helps to revisit the analytical process and allows for future revisions.

PREVENTION OF HARM

The fteval Standards prescribe that "activities undertaken in connection with evaluations are carried out in a manner that is ethically responsible, gender aware, and with openness toward social and cultural diversity (e.g. age, background, language)." and that "all those involved in or affected by an evaluation are treated with respect and fairness." (Kohlweg 2019: 15) Similarly, the EU-HLEG on AI's[1] Ethics Guidelines for Trustworthy AI state that "AI systems should neither cause nor exacerbate [collective or individual] harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity." (EC DG CNCT 2019: 12) As such, with regard to the application and development of all research tools and methods, the responsibility for ethical development and ethical use of AI models lies both with the developer and user. When training AI models with training data for specific evaluative purposes, but also when prompting and reviewing the outputs generated by AI models, special care must be taken to identify, minimise, actively counteract and be transparent regarding the biases inherent in the training data and subsequent outputs. This is especially true when AI is used for decision-making processes, which in turn means that human verification is a must for central activities such as writing manuscripts and data scripts, peer review, proposal evaluation, and so on.

Potential discriminatory or unfair outcomes must be actively addressed and counteracted to ensure human, environmental and ecosystem safety and security. Some guidance for developers and decision-makers is provided e.g. by the UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO 2021a) and its companion document Ethical impact assessment: a tool of the Recommendation on the ethics of artificial intelligence (UNESCO 2021b). Furthermore, in its report on Regulating AI in the UK, the Ada Lovelace Institute (2023) identifies the need for regulatory frameworks that, among others,

1 The independent High-Level Expert Group on Artificial Intelligence was set up by the European Commission in June 2018.

contain legal rights and protections for people, establish routes to redress, and introduce reporting requirements.

In addition, technical security checks need to be employed at relevant points within an AI life-cycle to address security vulnerabilities in AI models or codes developed with AI systems, in order to prevent models from manipulation, data breaches, unauthorised access, hidden malware, etc. Data collected and used by AI systems need to be handled responsibly, adhering to relevant data privacy regulations and user consent.

FAIRNESS AND NON-DISCRIMINATION

All existing AI guidelines point towards the necessity of fairness and non-discrimination in the development, deployment and use of AI systems. For the EU-HLEG on AI (EC DG CNCT 2019), both the substantive and the procedural dimension of fairness are of importance. The substantive dimension pertains to the general commitment to fairness, which in various documents entails e.g. the equal and just sharing of benefits and costs (see e.g. APA 2022); the promotion of social justice and safeguarding of fairness and non-discrimination of any kind, making all reasonable efforts to minimise and avoid reinforcing or perpetuating discriminatory or biassed applications and outcomes (see UNESCO 2021a); and that the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice (EC DG CNCT 2019). The EU-HLEG on AI also points out that "fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives." (EC DG CNCT 2019: 12f)

The procedural dimension of fairness, on the other hand, ties into the steps laid out above, and entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable. As such, disclosure obligations in the development and application cycle and the attributability of responsibility and accountability are necessary requirements of a fair AI system (Heesen et al. 2020). It also means that human agency and oversight in the development and deployment of AI systems are essential components of fair AI systems.

ENSURING PRIVACY

The privacy standards as defined by the GDPR remain an important guide also in the context of GenAI systems. It lays out the need to safeguard individual rights regarding the use of personal data, and to provide easily accessible and understandable information about the data collection practices and the processing practices employed. In practice, this has a number of implications:

Since most GenAI applications are cloud-based, data must be uploaded to the cloud-infrastructure of the service provider in order to use these systems. Through this, a third party is given direct access to this data. By default, most cloud applications use uploaded data to further train their AI systems. Exploitable information includes e.g. text inputs, file uploads, and feedback, all of which might be triangulated with other information at a provider's disposal and used for purposes other than training the system. In addition, providers usually collect and store personal information on the user, which can also be linked to other data sources.

Not all data owners/sources permit AI-supported processing by use of an external AI system. Broadly speaking, published information such as texts on websites, publications, open data sets, patents, and codes can be used as input. However, it is important to verify whether the owner of the data allows or prohibits the input of these data, e.g. according to copyright or licensing agreements.

For personal data, informed consent for this type of processing must be sought. When planning to employ external AI systems e.g. for processing recordings or transcripts, informed consent sheets must reflect how and for what purposes these data are being processed and stored, and what measures will be taken to safeguard personal and sensitive data. Similarly, third party data such as non-published monitoring data of funding programmes provided by clients and project proposals submitted to funding agencies may only be used as an input if informed consent has been given. In general, confidential information, trade secrets, and data covered by non-disclosure-agreements may not be used as an input in external AI systems.

The same might not be true for local instances of an AI system with no connectivity to an external server or third-party service provider, where data security can be ensured. Still, these questions need to be clarified before employing such services, and relevant rules and regulations regarding a specific data handling task should be reviewed in advance. Likewise, any AI system's Terms of Service and Privacy Policy must be reviewed before inputting data, while

organisational guidelines and the wishes of contractual parties/collaborators must be taken into account when deciding whether and under which conditions an AI system may be used.

As such, privacy necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols, as well as the capability to process data in a manner that protects privacy throughout the entire AI lifecycle. This is also true when employing AI in the implementation of evaluations, as it is crucial to ensure a proper handling of data and user privacy. In any case, GDPR-compliance must be ensured along every step of the process.

A NOTE ON INTELLECTUAL PROPERTY RIGHTS (IPR)

GenAI impacts innovation and creation and thus raises various questions regarding intellectual property (IP), including how IPR might intersect with AI models, their training data, their inputs and their outputs. The legal situation regarding the intellectual property rights of AI-generated outputs is currently very much contested and remains unsolved. The Terms of Use of a GenAI system may provide information where and how its outputs can be used. If the output of an AI system is used without significant changes, the origin of the outputs might need to be disclosed.

According to the World Intellectual Property Organization (WIPO 2024), incorporating AI systems into the innovation process can complement human creativity in generating ideas and solutions, while humans remain responsible for defining problems, setting goals, and determining how AI-derived insights are applied. However, the integration of generative AI in this context may complicate the licensing or patenting of new solutions. Copyright law, traditionally centred on human creators, generally restricts ownership to humans in many countries. Yet, this concept is being contested in some jurisdictions. For instance, early US rulings assert that AI-generated content is not eligible for copyright unless there is clear evidence of human creative contribution (Brittain 2023). In contrast, the Beijing Internet Court has recognised images generated by Stable Diffusion as original works, citing the human input through prompts as sufficient for copyright eligibility (Lai et al. 2024).

IPR is also not synonymous with Austrian copyright or "Urheberrecht", leading to further potential confusion. An added layer of complexity comes from the numerous ambiguities regarding human involvement and intentions (in terms of authorship) and the training data the AI system could have used in generating the output, often scraped from the internet without the consent of the

original author(s) (El Atillah 2023). Finally, in the context of RTI (policy) evaluation, all evaluation results usually belong to the client, which makes the use of AI systems and ownership of results even more tricky.

CURRENT PRACTICES IN THE USE OF AI IN RESEARCH AND EVALUATION

AI will most likely have a considerable impact on the routine work of researchers and other experts engaged in the field of RTI evaluation. Overall, the percentage of researchers reporting extensive use of AI in their research increased from 12% in 2020 to 16% in 2021 across all fields – even before the wide release of ChatGPT-4 in March 2023. Bibliometric analysis shows consistent increase in the share of research papers mentioning AI or machine-learning terms across all fields over the past decade, reaching around 8% in total (van Noorden & Perkel 2023). A foresight study of the European Research Council (ERCEA 2023), conducted among ERC grantees, explored the concrete uses of AI in their scientific practice. The study shows that the use cases for AI systems are far-reaching, although they are still early in the process of being explored. Overall, respondents were positively inclined towards the opportunity for GenAI to handle repetitive or labour-intensive tasks (85%) and reduce language barriers (75%), although 93% of respondents found the implementation of ethical guidelines for AI a likely requirement. This tracks with serious concerns expressed regarding a lack of transparency and replicability (71% of respondents), but also the intrusiveness, manipulation and discrimination of AI systems (79% of respondents). AI systems also seem to be seen mostly as assistant technologies, although on average only about 54% of respondents were sceptical about the possibility of AI-based scientific publication and peer-review, while 50% saw AI affecting research integrity. Future perspectives on the use of AI are mostly collaborative and range from scientific discovery purposes, brainstorming on scientific ideas with increasingly responsive AI companions, to generating new scientific hypotheses. Only a limited number of researchers expect fully autonomous research processes, in which AI also takes actions such as selecting the best resources in the lab or accessing additional data if needed.

In the public policy arena, Tangi et al. (2022) report that policymakers recognise that the integration of AI in the public services could provide large benefits and public value to citizens (reduced administrative burden, personalised

services, etc.), but they are also accompanied by some serious societal, ethical, regulatory, and technological challenges to be addressed. With respect to analytical purposes, potential use cases for analysis, monitoring and regulatory research, include 1) the process of inspecting, transforming, and modelling information into actionable knowledge (e.g. dashboard to support decision-making), 2) monitoring of policy implementation, and 3) processes for management of resources based on prediction models, to support planning. Furthermore, the study highlights potentials to increase public services and engagement management.

This makes apparent a number of benefits and challenges associated with the use of AI systems for evaluation purposes. In terms of benefits, mentioned dimensions include better cleaned data, possibility to analyse large quantities of data, more fluid processes through AI automation, real-time monitoring in rapidly changing contexts, removal of language barriers widening the access of STI research and policy with non-native speakers, better articulation of thoughts and enhanced communication, creative input of the AI, improvement of prompting, coding, programming skills, and most prominently, efficiency gains. (Ferretti 2023) Potential challenges, however, may reduce any such efficiency gains, for instance if there is a high dependence on a system and it fails to perform either in terms of content or technical aspects. Other challenges in the field of AI include ethical issues such as biases embedded in algorithms as well as their traceability, data security on servers beyond EU jurisdiction or data protection of personal information, as well as authorship and copyright concerns related to data used to train Large Language Models (LLMs). Additionally, incorporating AI into STI policy evaluation may present challenges due to its limited contextual understanding and hallucinations, potential reinforcement of the digital divide, and associated infrastructure and expertise costs. (Odumbe 2023)

AI USE WITHIN THE FTEVAL COMMUNITY

The survey developed by the working group for the fteval community led to a total of 20 responses, covering 14 of the 26 fteval member institutions. Within this group of respondents, funding agencies, research performing organisations and evaluators are relatively well represented, while ministries are covered the least. Still, these responses give important indications on attitudes and practices regarding the use of AI across relevant RTI organisations in Austria as of spring 2024. For instance, only 2 organisations indicated that no AI systems were in use at the time, neither at individual nor at organisation level. At the same time, only 4 respondents indicated that there were standard AI sys-

tems in use within their organisation – all of which were research performing organisations. 40% of respondents (8 out of 20) indicated that their institutions supplied them with certain AI systems for experimental use (representing 7 fteval members, including a ministry), while 75% of respondents (15 out of 20) feel the use of AI is mainly pushed by individuals. This seems to indicate that interest in AI was considerable, while use cases still needed to be established – according to respondents, no organisation used AI for decision-making processes. Only 3 of the responding organisations had no guidelines in place for the use of AI, with a fourth having not yet put them into writing. Of the remaining organisations, half had guidelines in development and half had them already in place.

Since adoption of AI was still rather early and there was at the time no clarity on relevant laws and regulations, any institutional guidelines must be seen as preliminary and could not cover all eventualities. While no responding organisation indicated a systematic use of AI systems within their institution, individual use of GenAI systems seemed widespread and hard to govern. The responses indicated that within the various organisations, some staff were already using AI systems for a lot of tasks, while others did not consciously use them at all. Complexities arise since AI systems pop up in different places unprompted, being integrated in ever more sophistication in search engines (such as Google and Bing), word processing software (as in Word assistant technologies), online meeting services (such as the Zoom AI Companion), graphic design and layout software (such as in the Adobe Suite), and operating systems (such as Microsoft Copilot). Nevertheless, the survey responses gave insight into typical scenarios in which AI systems are used or might be used in the future, broadly grouped by the authors into the following categories:

- **Text generation**, for instance for job vacancies, emails, programming/ coding, terms of reference (ToR), but also the development of survey items / questionnaires.
- **Proofreading and quality check**, including for accessibility/ plain language, coding, but also e.g. for the review of the fulfilment of all ToR or potentially for controlling on the use of AI for text and image generation.
- **Text and data processing**, including for the extraction, analysis and cleaning of data, summarising and synthesising, and for categorisation and text-mining.
- **Transcription** of audio, both live and recorded.

- **Translation of text**, which might also support simultaneous interviewing in different languages.
- **Knowledge gathering**, including desk research of relevant literature but also brainstorming, as an initial information tool to gain an overview, and for linking and mapping of thematic areas within a field.
- **Image generation**, including the creation of copyright-free images and document templates, visualisation of data and results, but also potentially to detect image manipulation.
- **Automatisation** especially of repetitive, monotonous, easy to control tasks, but potentially in the future also of processes such as peer-review.
- **Generalised support**, e.g. to enlarge participative formats, to serve as search tools in existing file systems, or for formatting reports.

Systems for language checking and translation – including DeepL – seem to be in regular use across institutions, while transcription tools were also mentioned a lot. More involved applications of GenAI, e.g. for summarising, structuring, mining, analysing or reworking texts, supporting desk research, developing codes for analysis, and phrasing of emails, were mentioned by most respondents as regular occurrences.

At the same time, a strong need for more guidance was expressed by most respondents. While the existing guidelines are mostly described as rudimentary – and dependent on future regulation – responsibility for exploring the use of AI systems, as well as for the downstream effects of said use, seemed at the time to be mostly passed onto the individual users. In this context, a provision of training was named as essential, especially on topics such as privacy and (GDPR) compliance, AI ethics, technical implementation, transparency, but also new developments in the field of AI and how they might be effectively employed. In this context, clear guidance is required both from ministries and employers that also outline responsibilities within the institution and beyond.

While a lot of existing institutional guidelines ask for AI generated outputs to be marked as such, there is no security on the extent to which AI is truly in use, and no standardisation in how such contents can and should be marked.

AI APPLICATIONS IN THE EVALUATION CONTEXT

The following section outlines some potential use cases for the application of AI systems for RTI evaluation by exploring synergies across the evaluation phases and discussing the feasibility of project and proposal evaluation, before proposing a self-assessment list to support the consideration process when adopting proprietary AI systems.

EXPLORING SYNERGIES AT THE INTERSECTION OF PROPRIETARY AI SYSTEMS AND EVALUATION PHASES

Experts working in the area of RTI (policy) evaluation perform a multitude of tasks in different organisational contexts. As researchers and evaluators, they are responsible for analysing RTI policies and programmes and their framework conditions within international, national, regional or sectoral innovation systems. As programme experts in national ministries, they are responsible for setting up R&I policies and instruments, delineating evaluation plans, and launching and overseeing R&I policy evaluation processes. Within funding agencies, RTI experts are responsible to design and execute transparent project selection mechanisms and ensure appropriate monitoring systems and portfolio analyses to guarantee efficient and effective funding, prevent fraud, and lay the foundations for impact measurement.

According to the 2019 fteval Standards (Kohlweg 2019), an evaluation typically comprises a preparation and planning phase, an implementation phase, and a management response phase. The implementation phase is further divided, with the term 'inception phase' commonly used when refining the evaluation design and reviewing its scope, purpose, methodology, and other aspects. The data collection phase, then, involves gathering relevant information to establish a basis for analysis and assessment. The data analysis and triangulation phase entail examining data from various sources, using diverse analytical techniques, and cross-referencing findings to improve the validity and reliability of conclusions. In the reporting phase, the findings, conclusions, and recommendations derived from data analysis are synthesised and communicated in a comprehensive and accessible manner to stakeholders.

To implement their tasks, experts from the evaluation community draw upon quantitative and qualitative research methods and tools to design policy inter-

ventions, set up monitoring and evaluation frameworks, implement monitoring and evaluation actions, and facilitate learning processes through engagement with diverse stakeholders. All these tasks and skills can be supported by AI to various degrees. According to Canadian evaluator and researcher Steve Jacobs, current literature suggests that AI could benefit every stage of the evaluation process (Jacobs 2023). In 2023, a group of nine evaluation experts analysed the evaluation activities and stages that could be affected by the use of artificial intelligence. On average, the study found that two thirds of evaluation activities can be impacted by artificial intelligence. (Head et al. 2023) The current level of support that AI can provide for evaluations is limited to narrow AI, which is task-specific. Some of these translate into applications which can be used at different stages in the implementation of RTI evaluations: 1) to systematically explore large or growing data sources (such as archives or document repositories) that are prohibitively time-consuming to process for humans; 2) to continuously improve assessments with new data being introduced (new categories and implementation challenges can be added, updated, etc. as the body of data to be assessed changes); and 3) to perform quality control and check for accuracy of methods employed. (Raimondo et al. 2023) Still, it is important to point out that at this stage, quality control remains a human task, and the AI user remains the responsible entity for ensuring the veracity and validity of any AI-generated outputs employed in their work. This is also an explicit requirement for the inclusion of AI applications in evaluative work for many institutions, including research funding organisations. The table below selectively maps use cases of AI capabilities and applications with evaluation phases and task examples. Since AI capabilities are constantly evolving, these must be seen as a snapshot.

AI capability	Application	Evaluation phase	Evaluation task examples
Natural Language Processing	Text editing, automated text generation	Inception, data collection & analysis, reporting & communication	Written deliverable (data collection tools, report, etc.), correspondence with evaluation stakeholders, coding, excel formulas, etc.

Natural Language Processing	Summarisation and synthesis	Data collection & analysis, reporting & communication	Tailoring evaluation results to cater to different audiences, synthesis of policies or previous studies for a review or a context section, etc.
Natural Language Processing and Machine Learning	Speech to speech, text to speech, text to text, speech to text	Data collection & analysis, reporting & communication	Automated dictation of interviews, focus groups, observations, evaluation meetings, live interpretation with end beneficiaries when no common language is spoken, translation of written deliverables
Natural Language Processing and Machine Learning	Document processing, data extraction	Data collection & analysis	Meta analyses, etc.

Natural Language Processing and Machine Learning	Chatbots, knowledge sharing	Inception, data collection & analysis	Refining, brainstorming on the best methodological approach, peer opinion, research companionship, collective intelligence gathering, desk research, design of interaction formats (training of evaluators, surveyors) sensitive to end beneficiaries. Social experiments such as through the simulation of social interactions, substitute or testing ground for human participants. (Grossman 2023)
Machine Learning	Predictive analytics	Data collection & analysis	Foresight and risk assessments based on historical data, etc.
Machine Learning	Anomaly detection	Data collection & analysis	Analysis of large sets of data such as financial transactions, outliers in results data, etc.

Table 1: Correspondence table of AI applications and evaluation tasks examples

PROJECT AND PROPOSAL EVALUATION

While the use of AI systems for the evaluation of projects and proposals can be seen as analogous to the higher-level evaluation outlined in the previous section, there are as of yet no established systems employed on a broader basis to support evaluation efforts. However, potential areas of application have been pointed out specifically to detect biases, inconsistencies and discrepancies in an evaluation process (Divasón et al. 2023a, 2023b), while also enhancing the speed and potentially overall scope of an evaluation, allowing for new insights to be drawn (Raimondo et al. 2023).

As the exploration of the potential use cases of AI for systematic employment in project and proposal evaluation are still at an early stage, only a few further considerations can be sketched out at this point:

- Legal and regulatory frameworks are still very much in the making. This needs to be taken into account when setting up new processes, in that they might need to be adjusted after the fact to fit new regulation.
- The costs of setting up new AI systems may be considerable, especially if they are to adhere to existing and developing regulation, including privacy standards and standards for scientific rigour and integrity.
- Big tech companies are thus in a favourable position, being able to invest large amounts of resources to the development of new technologies, while regulation is always necessarily lagging behind.
- The downstream effects of employing AI systems at a larger scale, especially to support decision-making processes, are currently hard to foresee. This might pertain to the development of inherently biassed systems, impacts on labour markets and work practices, and the overall quality of outputs in closed-loop systems.
- As such, it is important to proceed with care and commit to AI systems only if sufficient data is available on the consequences of such an institutional change.

PROPOSING A SELF-ASSESSMENT LIST TO CONSIDER PROPRIETARY AI SYSTEM ADOPTION

A semi-systematic evaluation of 22 guidelines (Hagendorff 2020) highlighted significant shortcomings in the field of AI ethics, noting that ethical guidelines often lack enforcement mechanisms, leading to no real consequences for de-

viations. At the time, ethics in AI was frequently viewed as a marketing tool rather than a core aspect of technology development, with guidelines having little impact on software developers' decision-making. There was a general perception of AI ethics as an optional addition, and a distributed responsibility model diluted accountability. Economic incentives often outweighed ethical commitments, suggesting a misalignment between AI development purposes and societal values. Despite these challenges, there were efforts to address specific ethical concerns in AI through technical solutions, such as enhancing privacy, anti-discrimination measures, and explainability. However, many ethical issues remain under-addressed or ignored in guidelines, spanning a wide range of topics from the potential for malevolent AI to the social and ecological impacts of technology. The study suggested the need for more comprehensive and effectively enforced ethical guidelines in AI development, which remain relevant today as the range of stakeholders using AI systems has expanded to include the wider public. Following this, in 2020 the European Commission published a guide on ethical principles for designers and developers of AI systems, data scientists, procurement officers or specialists, front-end staff working with AI systems, legal and compliance officers, and managers, against which they can self-assess the trustworthiness of AI systems (EC DG CNCT 2020).

Beyond ethical considerations, more concrete operational steps should be envisaged as part of a process using an AI system or purchasing AI services. These are outlined for consideration in the figure below:

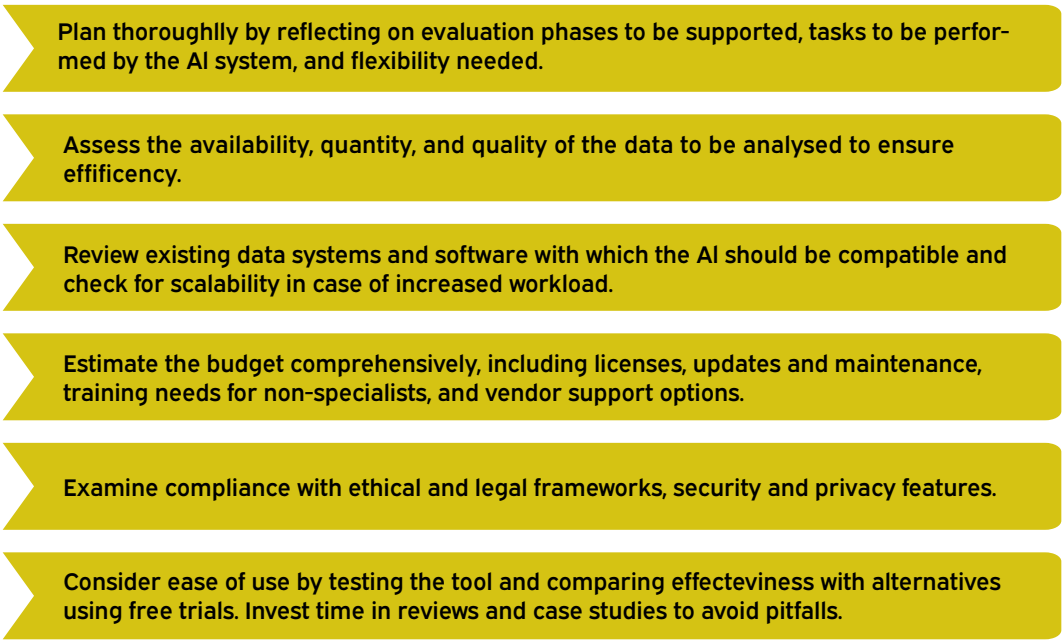


Figure 2: Self-assessment considerations for adopting proprietary AI systems

OUTLOOK

The technical, legal and ethical competencies needed in the evaluation community to utilise AI systems responsibly will increase. To explore the opportunities and the ways to employ AI systems responsibly for better evaluations, AI skills training and open online resources should be built with a focus on technical, legal and ethical questions. As outlined in the section on AI use in the fteval community, there is a strong need for more guidance and training to support the responsible and effective application of AI systems for evaluation purposes. Key topics for which training is needed broadly cover best practices on the development, deployment and application of AI systems, again focussing on technical, legal and ethical aspects such systems imply. In particular, questions on privacy and (GDPR) compliance, transparency, as well as quality assurance, ethics, and bias loom large. Here, the fteval platform might provide support to member organisations by inviting external experts or continuing exchange meetings, allowing the community to keep track of this quickly changing field.

REFERENCES AND FURTHER READING

ACRAI – Österreichischer Rat für Robotik und Künstliche Intelligenz. (2018). Die Zukunft Österreichs mit Robotik und Künstlicher Intelligenz positiv gestalten. White Paper des Österreichischen Rats für Robotik und Künstliche Intelligenz. https://www.bmk.gv.at/dam/jcr:f2f7a973-8aa4-4be8-9a6b-0c7c44e73ce4/white_paper_robotikrat.pdf.

Ada Lovelace Institute. (2023). Regulating AI in the UK: Strengthening the UK's proposal for the benefit of people and society. https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/.

AI-Kompetenznetzwerk. (2023). Kompass für den dienstlichen Umgang mit generativer Künstlicher Intelligenz (KI). Magistratsdirektion der Stadt Wien. Geschäftsbereich Organisation und Sicherheit. Gruppe Organisation. Wien.

APA – Austria Presse Agentur. (2022). Leitlinie zum Umgang mit künstlicher Intelligenz. Wien.

Biegelbauer, P., Lackinger, C., Schlarb, S., Subak, E., Weinlinger, P. (2023). Leitfaden Digitale Verwaltung und Ethik. Praxisleitfaden für KI in der Verwaltung, Version 1.0. Bundesministerium für Kunst, Kultur, öffentlicher Dienst und Sport. Wien.

BMK. (2021a). Strategie der Bundesregierung für Künstliche Intelligenz. Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie. Wien.

BMK. (2021b). Strategie der Bundesregierung für Künstliche Intelligenz: Annex. Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie. Wien.

BMWE. (2020). Ethik und Künstliche Intelligenz. Was können technische Normen und Standards leisten? Whitepaper. DIN e.V. und DKE. Berlin.

Bockting, C. L., et al. (2023). Living guidelines for generative AI — why scientists must oversee its use. In: Nature Vol. 622.

Brittain, B. (2023). AI-generated art cannot receive copyrights, US court says. Reuters. https://www.reuters.com/legal/ai-generated-art-cannot-receive-copyrights-us-court-says-2023-08-21/ (last accessed 24.09.2024).

BVDW. (2019). Acht Leitlinien für künstliche Intelligenz. Leitlinien des BVDW. Bundesverband Digitale Wirtschaft. Berlin.

de Witt, C., Rampelt, F., Pinkwart, N. (Hrsg.). (2020). Künstliche Intelligenz in der Hochschulbildung – Whitepaper. Berlin: KI-Campus. https://doi.org/10.5281/zenodo.4063722.

Der Bundesrat. (2020). Leitlinien «Künstliche Intelligenz» für den Bund: Orientierungsrahmen für den Umgang mit künstlicher Intelligenz in der Bundesverwaltung. Schweizerische Eidgenossenschaft.

Der Bundesrat. (2022). Monitoring der Leitlinien «Künstliche Intelligenz» für den Bund: Evaluation der Anwendung und Aktualität der Leitlinien. BAKOM-Bericht. Schweizerische Eidgenossenschaft.

DFG. (2023). Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG.

Divasón, J., Martínez-de-Pisón, F.J., Romero, A., Sáenz-de-Cabezón, E. (2023a). Artificial Intelligence Models for Assessing the Evaluation Process of Complex Student Projects. In: IEEE Transactions on Learning Technologies 16(5). 694-707. https://doi.org/10.1109/TLT.2023.3246589.

Divasón, J., Martínez-de-Pisón, F.J., Romero, A., Sáenz-de-Cabezón, E. (2023b). Robustness Analysis of a Methodology to Detect Biases, Inconsistencies and Discrepancies in the Evaluation Process. In: García Bringas, P., et al. International Joint Conference 16th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2023) 14th International Conference on EUropean Transnational Education (ICEUTE 2023). CISIS ICEUTE 2023. Lecture Notes in Networks and Systems, vol 748. Springer, Cham. https://doi.org/10.1007/978-3-031-42519-6_31.

Europäische Kommission, Generaldirektion Kommunikationsnetze, Inhalte und Technologien. (2019). Ethik-Leitlinien für eine vertrauenswürdige KI. Amt für Veröffentlichungen der Europäischen Union. https://doi.org/10.2759/22710.

Europäische Kommission. (2022). Ethische Leitlinien für Lehrkräfte über die Nutzung von KI und Daten für Lehr- und Lernzwecke. Amt für Veröffentlichungen der Europäischen Union. https://doi.org/10.2766/494.

EC – European Commission. (2020). White Paper on Artificial Intelligence—A European approach to excellence and trust. COM(2020).

EC DG CNCT – European Commission, Directorate-General for Communications Networks, Content and Technology. (2019). Ethics guidelines for trustworthy AI. Publications Office. https://data.europa.eu/doi/10.2759/346720.

EC DG CNCT – European Commission, Directorate-General for Communications Networks, Content and Technology. (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. Publications Office. https://data.europa.eu/doi/10.2759/002360.

EC DG RTD – European Commission, Directorate-General for Research and Innovation, Arranz, D., Bianchini, S., Di Girolamo, V. et al. (2023). Trends in the use of AI in science – A bibliometric analysis. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/418191.

El Atillah, I. (2023, July 10). Copyright challenges in the age of AI: Who owns AI-generated content? Euronews. https://www.euronews.com/next/2023/07/10/copyright-challenges-in-the-age-of-ai-who-owns-ai-generated-content (last accessed 24.09.2024).

ERCEA – European Commission, European Research Council Executive Agency. (2023). Use and impact of artificial intelligence in the scientific process – Foresight. Publications Office of the European Union. https://data.europa.eu/doi/10.2828/10694.

Ferretti S. (2023). Hacking by the prompt: Innovative ways to utilize ChatGPT for evaluators. In: New Directions for Evaluation 178-179. 73-84. https://doi.org/10.1002/ev.20557.

Feuerriegel, S., Hartmann, J., Janiesch, C., Zschesch, P. (2024). Generative AI. In: Business & Information Systems Engineering 66. 111-126. https://doi.org/10.1007/s12599-023-00834-7.

Google. (2022). AI Principles Progress Update.

Grossmann, I. (2023, July 3). Beyond the hype: How AI could change the game for social science research. The Conversation. https://theconversation.com/beyond-the-hype-how-ai-could-change-the-game-for-social-science-research-208086 (last accessed 24.09.2024).

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. In: Minds and Machines 30. 99-120. https://doi.org/10.1007/s11023-020-09517-8.

Head, C. B., Jasper, P., McConnachie, M., Raftree, L., Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical impli-

cations. In: New Directions for Evaluation 178-179. 33-46. https://doi.org/10.1002/ev.20556.

Heesen, J. et al. (Hrsg.). (2020). Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen – Whitepaper aus der Plattform Lernende Systeme. München.

Heidrich, J. (2023). Gut ausgerichtet: Richtlinien für die Nutzung von KI im Unternehmen schaffen. In: c't 2023, Heft 19.

Heller-Schuh, B., Kasztler, A., Leitner, K.-H. (2019). Künstliche Intelligenz als thematische Herausforderung für österreichische Universitäten. AIT-IP-Report Vol. 21.

Hense, J., Rädiker, S. (2023). ChatGPT & Co. – Anwendungsszenarien künstlicher Intelligenz in der Evaluation. Foliensatz. https://www.degeval.org/fileadmin/content/Arbeitskreise/AK_Methoden/KI/JT2023_AK-Session_A4_KI_Doku.pdf.

Kohlweg, K. (2019). Evaluation Standards for Research, Technology and Innovation Policy. Technischer Bericht. fteval - Österreichische Plattform für Forschungs- und Technologiepolitikevaluierung. Wien. https://doi.org/10.22163/fteval.2019.344.

Jacobs S. (2023, December 12). Webinaire sur l'intelligence artificielle et le futur de l'évaluation. Société canadienne d'évaluation (SCÉ) & Perfeval.

Lai, S. Lim, D. Shi, L., Tay, J. (2024, March). Legal implications – Beijing Internet Court grants copyright protection to AI-generated artwork. National University of Singapore – Centre for Technology, Robotics, Artificial Intelligence & the Law. https://law.nus.edu.sg/trail/legal-implications-beijing-internetcourt-copyright/ (last accessed 2024.09.24.).

MacLaren, I., O'Brien, G., et al. (2023). Generative Artificial Intelligence: Guidelines for Educators. National Academic Integrity Network. Quality & Qualifications Ireland.

Nancholas, B. (2023, October 12). Natural language processing: An explainer. University of Wolverhampton. https://online.wlv.ac.uk/natural-language-processing-an-explainer/ (last accessed 2024.09.24.).

NSAI. (2023). AI Standards & Assurance Roadmap. National Standards Authority of Ireland. Dublin.

Odumbe, K. O. (2023, December 11). Integrating AI into Monitoring and Evaluation: A Pathway for Enhanced Efficiency in Development Work. LinkedIn. https://www.linkedin.com/pulse/integrating-ai-monitoring-evaluation-pathway-enhanced-ken-odak-odumbe-u3pef/ (last accessed 2024.09.24.).

OECD. (2022). Artificial Intelligence in Science. Challenges, Opportunities and the Future of Research. OECD Publishing. Paris. https://doi.org/10.1787/a8d820bd-en.

OECD. (2024 [2019]). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449.

OECD. (2024). Transparency and explainability (Principle 1.3). OECD AI Principles. https://oecd.ai/en/dashboards/ai-principles/P7 (last accessed 24.09.2024).

Ostendorf, A., Peters, M. (2003). Handreichung für Studierende an der Fakultät für Betriebswirtschaft zum Einsatz von KI-Tools im Studium. Leopold-Franzens-Universität Innsbruck.

Pant, A. (2019, January 7). Introduction to Machine Learning for Beginners. Towards Data Science. Medium. https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08 (last accessed 2024.09.24.).

Pinaya, W. H. L., Graham, M. S., Kerfoot, E., Tudosiu, P.-D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P. F., Patel, A., Chung, H., Zhao, C., Peng, W., Liu, Z., Mei, X., Lucena, O., Ye, J. C., Tsaftaris, S. A., Dogra, P., Feng, A., Modat, M., Nachev, P., Ourselin, S., Cardoso, M. J. (2023). Generative AI for Medical Imaging: extending the MONAI Framework. In: Electrical Engineering and Systems Science. https://doi.org/10.48550/arXiv.2307.15208.

Poretschkin, M. et al. (2021). Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. KI-Prüfkatalog. Fraunhofer IAIS. Sankt Augustin.

Projektgruppe „Künstliche Intelligenz“ des VDMA Bayern. (2020). Leitfaden Künstliche Intelligenz – Potenziale und Umsetzungen im Mittelstand. VDMA Bayern. München.

Prunkl, C. (2024). Human Autonomy at Risk? An Analysis of the Challenges from AI. In: Minds & Machines 34, 26. https://doi.org/10.1007/s11023-024-09665-1.

Raimondo, E., Anuj, H., Ziulu, V. (2023, August 16). Setting up Experiments to Test GPT for Evaluation. Independent Evaluation Group – World Bank Group. https://ieg.worldbankgroup.org/blog/setting-experiments-test-gpt-evaluation (last accessed 24.09.2024).

Regona, M., Yigitcanlar, T., Xia, B., & Li, R. Y. M. (2022). Opportunities and Adoption Challenges of AI in the Construction Industry: A PRISMA Review. In: Journal of Open Innovation: Technology, Market, and Complexity 8(1). 45. https://doi.org/10.3390/joitmc8010045.

Russell Group. (2023). Russell Group principles on the use of generative AI tools in education.

Segreto, T. (2014). Knowledge-Based System. In: Laperrière, L., Reinhart, G. (eds) CIRP Encyclopedia of Production Engineering. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-20617-7_6557.

Tangi L., van Noordt C., Combetto M., Gattwinkel D., Pignatelli F. (2022). AI Watch. European Landscape on the Use of Artificial Intelligence by the Public Sector. EUR 31088 EN. Publications Office of the European Union. Luxembourg. https://doi.org/10.2760/39336.

U.S. Department of Education, Office of Educational Technology. (2023). Artificial Intelligence and the Future of Teaching and Learning. Insights and Recommendations. Washington, DC.

UNESCO. (2021a). Recommendation on the ethics of artificial intelligence. UNESCO France.

UNESCO. (2021b). Ethical impact assessment: a tool of the Recommendation on the ethics of artificial intelligence. UNESCO France. https://doi.org/10.54678/YTSA7796.

Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. In: Nature 621(7980). 672–675. https://doi.org/10.1038/d41586-023-02980-0.

WIPO. (2024). WIPO Conversation IP and Frontier Technologies: Generative AI. https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_conv_ge_2_23/wipo_ip_conv_ge_2_23_summary.pdf.

Wischmann, S. et al. (2022). Leitfaden für das Qualitätsmanagement bei der Entwicklung von KI-Produkten und -Services. Begleitforschung des Technologieprogramm KI-Innovationswettbewerb des Bundesministeriums für Wirtschaft und Klimaschutz (BMWK). Institut für Innovation und Technik (iit) in der VDI/VDE Innovation + Technik GmbH. Berlin.

AUTHORS

STEFANIE SCHUERZ
ZSI - Zentrum für Soziale Innovation GmbH
Linke Wienzeile 246
1150 Vienna
ORCID 0000-0002-3816-1294

CHARLOTTE D'ELLOY
Technopolis Group
Rudolfsplatz 12/11
1010 Vienna

MICHAEL DINGES
AIT Austrian Institute of Technology GmbH
Giefinggasse 4
1210 Vienna
ORCID 0000-0003-0433-4318

ANNEX I: QUESTIONNAIRE TO THE FTEVAL COMMUNITY

The following questionnaire was translated from the original German into English.

1. Information on you and your organisation

- What is your name? (Open question)
- In what organisation are you employed? (Open question)
- What is your role within the organisation? (Open question)

2. Relevance of AI in your organisation and your area of work

(Likert scale: 1. not relevant; 2. scarcely relevant; 3. neither/nor; 4. relevant; 5. very relevant)

- How relevant is the topic of AI for your organisation at this time?
- How relevant is the topic of AI for you in your area of work?
- How relevant is the topic of "Guidelines for the use of AI" in your organisation?
- How relevant are "Guidelines for the use of AI" in your area of work?

3. In what stage of AI use is your organisation?

- AI tools are not used in our organisation – neither by individuals nor on an organisational level.
- AI tools are mainly used by individuals for their own purposes.
- Our organisation has provided certain AI tools for trial in specific work areas.
- There are AI tools that are used by default in our organisation.
- Other:

4. Are there already guidelines for the use of AI in your organisation?

- Yes
- No
- Currently in development
- I don't know

5. **If yes or in development: What experiences are there in your organisation regarding "Guidelines for the use of AI"? (Open question)**

6. **Are AI tools used in your organisation or in your area of work for tasks that are part of decision-making processes?**

   - Yes
   - No
   - I don't know

7. **If yes, which tasks? (Open question)**

8. **What experiences are there with the use of AI in your organisation or in your area of work?**

9. **What is a typical scenario in your area of work where you are using AI tools, or might do so in the future? (Open question)**

10. **Please list typical AI tools that you are using or that you know are being used in your organisation. What are these tools used for?**

11. **Which kind of support would you like to have to ensure a safe and reliable application of AI as an employee?**

12. **How do you handle receiving AI-generated outputs from external stakeholders (e.g. project partners, beneficiaries)?**

13. **What are the biggest challenges you or your organisation are currently facing with regard to the use of AI tools?**

14. **Any other comments?**

ANNEX II: AI SYSTEMS

Name	Description	Evaluation step	Pricing model	Language support	Alternative	Review
consensus	Consensus is a search engine that uses language models to surface papers and synthesize insights from academic research papers. The current source material used in Consensus comes from the Semantic Scholar database, which includes over 200M papers across all domains of science. More papers are meant to be added. The dataset is updated on a monthly cadence. Consensus is not a chatbot, but uses the same technology to help make the research process more efficient.	- Proposal writing/ context - Assessment of the evaluation criteria "relevance of programme design"	Freemium	English only	Elicit, Research GPT, ScholarAI, Scisummary, HeyScience	https://www.youtube.com/watch?v=YIow-IQFS9rg
fireflies.ai	Fireflies.ai is an AI transcription tool that records and transcribes meetings across various web-conferencing platforms like Teams, Zoom and Google Meet, actively listening and noting key insights converting spoken words into text and generating summaries. Besides basic transcription, Fireflies.ai also analyzes meeting sentiments, categorizing them into positive, negative, and neutral segments for easier review and team sharing.	- Data collection: interviews/ focus groups - Project management: client meeting	Freemium	60+ languages, incl. German, Dutch, French, Spanish, Portuguese, Italian and three English accents: UK, Australian, and US.	Airgram.io, Tactiq, Whisper	https://www.notta.ai/en/blog/fireflies-ai-review
Curie	Curie is an AI-powered writing assistant designed specifically for academic papers. It provides intelligent suggestions, improves writing structure, enhances flow, and assists with citations to help produce high-quality academic content. Curie uses advanced artificial intelligence and natural language processing algorithms to assist you in editing and translating scholarly writing. Curie analyses the input provided by users and suggests edits, helping with tasks like drafting articles, polishing grant applications, or improving writing style. Curie has been specifically designed for research writing, trained on a specialized collection of manuscripts edited by professional subject editors.	Report drafting	Free trial for 14 days Premium $11.25 Custom price for a preferred group of people	English only translations from Chinese, Portuguese and Spanish	PaperPal, Grammarly, Quillbot, writefull X, compose.ai	-

Name	Description	Evaluation step	Pricing model	Language support	Alternative	Review
scite_	Scite_ is a citation index tool that takes advantage of recent advances in artificial intelligence to produce “Smart Citations.” Smart Citations reveal how a scientific paper has been cited by providing the context of the citation and a classification system describing whether it provides supporting or contrasting evidence for the cited claim, or if it just mentions it.	Report drafting	Monthly €15,92, yearly €114,63. Additionally there are a few free applications (e.g. browser extension)	English (no information on other languages)	Zotero, Mendeley, Citavi, Jotbot	MIT Press
Browse AI	Browse AI is a web-based tool that allows you to extract and monitor data from any website without coding. With Browse AI, you can train a robot by clicking on the elements you want to extract from a website. The robot will simulate your actions and extract your desired data on your chosen schedule, giving you a live web data pipeline within minutes.	Data collection and monitoring	Freemium	English (no information on other languages)	Fivetran, hevo DATA, Apify, Dataddo, Bright data	Browse ai Erfahrung: Automatisieren mit der KI. Ist es gut? (seo-tech.de)
CHATPDF	ChatPDF simplifies reading and interacting with PDF documents by allowing users to ask real-time questions and receive context-specific answers based on the document’s content. Utilizing natural language processing and deep learning, it enables a conversational interface with PDFs by analyzing the text within uploaded documents to understand and extract key concepts. ChatPDF provides detailed answers by matching questions to relevant information in the document, supporting follow-up questions for an interactive and efficient experience. This tool combines scanning, natural language understanding, and AI-driven response generation to enhance document comprehension and interaction.	Proposal writing, data collection	Freemium	ChatPDF accepts PDFs in any language and can chat in any language.	Elephas APP, PDFgear	ChatPDF Review (2024): Should You Try It? | Great Software

Name	Description	Evaluation step	Pricing model	Language support	Alternative	Review
	Quivr is designed as a personal AI database or “second brain,” helping users store and effortlessly retrieve unstructured information through an AI knowledge base built on their data. Quivr is an exception in this list since it’s open-source, and can be run locally, ensuring transparency in data storage and security. Powered by generative AI, Quivr automatically organizes and categorizes uploaded information, simplifying access and eliminating the need for manual sorting. This tool is ideal for those inundated with information daily, helping clear mental clutter for better focus. Quivr offers a demo, access to its GitHub for customization, and a Discord channel for community support, emphasizing its role in streamlining information management with AI.	Data collection, data analysis	Freemium	No information	Obsidian	https://medium.com/@manaaki.walker-tepania/quivr-your-personal-conversational-knowledge-base-harnessing-the-power-of-large-language-3785ec8e3f96
Rows	Rows is the only spreadsheet with the capabilities of an AI Analyst and a native OpenAI integration. Both are available to everyone and free to use. What makes the AI features in Rows special is the combination of the spreadsheet interface with the AI capabilities. This means users can leverage AI to drag automations across ranges of cells, combine the AI actions with any other spreadsheet formula, and use cell references to create dynamic AI experiences. Addiationlly, the ability to use AI to summarize and help you make sense of data is special. Rows offers the only product that allows users to ask questions about a dataset using plain language.	Data collection	Freemium	English (no information on other languages)	Sheet AI	-

Name	Description	Evaluation step	Pricing model	Language support	Alternative	Review
formularizer	Formularizer is an AI-powered assistant that streamlines the creation and understanding of formulas, scripts, and Regex patterns for Excel, Google Sheets, Notion, and other platforms. It ensures user data security with the latest technologies, without storing any data on its servers. Leveraging OpenAI's advanced GPT-4 model, this tool simplifies turning text instructions into precise formulas rapidly and for free, supporting various operations and allowing for regular expression use.	Data analysis	Formularizer is free for 150 uses each month (5 uses per day), which is enough for most users. There also is a Premium version.	No information	Rows, GPTExcel, Excelformulabot	-
iThenticate 2.0	iThenticate 2.0 is the new plagiarism detection programme by turnitin. This programme is characterised by its high sensitivity and accuracy in detecting plagiarism and AI-generated text. iThenticate 2.0 is designed to help with quality assurance in science and in the production of content in high-stakes areas, such as government institutions. iThenticate 2.0 introduces a revitalized and contemporary interface, crafted with the research community in mind, and enabling effortless navigation from the initial onboarding stage through to the final similarity check.	Quality assurance	$100 starting costs	English, German, Spanish, Japanese and more	Grammarly Business, GPTZero	iThenticate Pricing, Alternatives & More 2024 | Capterra

fteval JOURNAL
for Research and
Technology Policy Evaluation

April 2025, Vol. 55, pp. e5, 1-35
DOI: 10.22163/fteval.2024.644

WIE WIRKT KÜNSTLICHE INTELLIGENZ IM EVALUATIONSSYSTEM?

DISKUSSIONSANSTÖSSE FÜR DIE GESTALTUNG DES EVALUATIONSSYSTEMS VON MORGEN.

THOMAS PALFINGER, FELIX GAISBAUER, ISABELLA WAGNER AND SUSANNE BECK
Beiträge von: Elisabeth Froschauer-Neuhauser, AQ Austria, Michael Strassnig, WWTF, Vitaliy Soloviy, AIT
DOI: 10.22163/FTEVAL.2024.644

ABSTRACT

In diesem Diskussionspapier werden erste Überlegungen zu möglichen Veränderungen durch künstliche Intelligenz (KI) in einem Evaluationssystem am Beispiel der Forschungs- und Innovationspolitikevaluierung angestellt. In den bisherigen Diskursen über Künstliche Intelligenz (KI) standen hier vor allem Datenschutz, Ethik sowie wissenschaftliche und methodische Zuverlässigkeit im Vordergrund. Dieses Diskussionspapier zielt darauf ab, eine zusätzliche, systemische Perspektive einzubringen: Es untersucht, wie sich die Beziehungen zwischen Akteur:innen in einem Evaluationssystem verändern, wenn dieses System mit generativer KI konfrontiert wird.
Eine Arbeitsgruppe der Österreichischen Plattform für Forschungs- und Technologiepolitikevaluierung (fteval) hat sich dieser Frage von Juli 2023 bis März 2024 gewidmet. Auf Grundlage interner Diskussionen und begleitender Literaturrecherche wurde ein Modell entwickelt, das als Instrument dient, um die Handlungspraxis der Akteur:innen im FTI-Evaluationssystem zu reflektieren und gemeinsam zu diskutieren. Ausgangspunkte sind die Annahmen, dass generative KI als disruptives Element in das bestehende Evaluationssystem eintritt und dass diese Werkzeuge die Beziehungen zwischen den Akteur:innen verändern können. Obwohl der Fokus auf das österreichische Evaluationssystem gerichtet ist, wird davon ausgegangen, dass das Modell auch in anderen Evaluationssystemen oder Branchen mit ähnlichen Strukturen nützlich sein könnte.

EINFÜHRUNG

Ansätze künstlicher Intelligenz (KI) haben die Fähigkeiten, menschenähnliche Arbeit zu produzieren, zuletzt rapide ausgebaut und die Zugänglichkeit zur Technologie, etwa in Form von OpenAI's ChatGPT, für eine breite Öffentlichkeit massiv erhöht (vgl. Dell'Acqua, et al. 2023). Dies führt dazu, dass sich derzeit starke Narrative aus unterschiedlichen gesellschaftlichen Bereichen (von Wissenschaft bis Politik), rund um diese Technologie spinnen, denen man sich unabhängig von den realen Fähigkeiten und Einsatz der Technologie kaum entziehen kann.

In der Konsequenz zeichnet sich für das Evaluationswesen, wie viele andere Systeme auch, die Notwendigkeit ab, einen konstruktiven Umgang mit der sich rapide entwickelnden Technologie zu finden. Die Auswirkungen generativer künstlicher Intelligenz auf den wissenschaftlichen Prozess und damit auch auf verschiedene forschungsbasierte Erhebungsmethoden des Evaluationsprozesses sind potenziell weitreichend (vgl. Van Noorden et al. 2023) und deutet auf größere Umbrüche in den kommenden Jahren hin (Haupt et al 2022; Chapinal-Heras und Díaz-Sánchez 2023; Stahl 2023; Konya und Nematzadeh 2024).

In den Diskursen um KI stehen bislang meist Aspekte des Datenschutzes, der Ethik und der wissenschaftlichen bzw. methodischen Zuverlässigkeit beim Einsatz von KI im Vordergrund. Mit diesem Diskussionspapier hoffen wir in diese Diskurse eine weitere, systemische Perspektive einzubringen, in dem wir die Frage aufwerfen, wie KI die Beziehungen zwischen Akteur:innen in einem Evaluationssystem auswirkt. Konkret Fragen wir: Wie verändern sich die Beziehungen zwischen Akteur:innen in einem Evaluationssystem, wenn dieses System mit generativer künstlicher Intelligenz[1] (in der Folge einfach „KI") konfrontiert ist?

Diese Frage hat sich eine Arbeitsgruppe der österreichischen Plattform für Forschungs- und Technologiepolitikevaluierung (fteval) in und mit ihrer Community in den Monaten von Juli 2023 bis März 2024 gestellt. In dem vorliegenden Diskussionspapier wird auf Grundlage der internen Diskussion und begleitender Literaturrecherche ein Denkmodell entwickelt und als Instrument angeboten, mit denen die Akteur:innen im FTI-Evaluationssystem

1 Generative (tiefe) Künstliche Intelligenz (KI) bezeichnet algorithmische Modelle, die darauf trainiert sind, eigenständig neue Daten zu erzeugen, indem sie die Wahrscheinlichkeitsverteilung eines gegebenen Datensatzes approximieren. Anders als diskriminative Modelle, die nur darauf abzielen, Klassenlabels vorherzusagen, lernen generative Modelle die zugrunde liegende Struktur der Daten und können daraus ganz neue Instanzen (Texte, Bilder, Audio, etc.) erschaffen. (z.B.: Ruthotto & Haber, 2021)

ihre Handlungspraxis reflektieren und gemeinsam diskutieren können. Ausgangspunkt hierfür sind die Grundannahmen, dass (1) generative KI in Form von Sprachmodellen als disruptives Element in das bestehende Evaluationssystem eintritt und (2) diese Werkzeuge künstlicher Intelligenz die Beziehungen zwischen den Akteur:innen in diesem System verändern können. Wir werden uns dabei zwar stark auf das österreichische Evaluationssystem beziehen, weil das Erfahrungswissen in der Arbeitsgruppe hier am größten war, allerdings gehen wir davon aus, dass das Modell auch in anderen Evaluationssystem oder Branchen mit ähnlichen Strukturen nützlich sein dürfte.

DAS DENKMODELL ZUM REFLEKTIEREN VON KI IM EVALUATIONSSYSTEM

Um eine strukturierte Auseinandersetzung mit den möglichen Auswirkungen von KI auf das Evaluationssystem zu erleichtern, wurde ein praxisnahes Denkmodell entwickelt. Es ermöglicht Akteur:innen und Entscheidungsträger:innen, die Veränderungen schrittweise zu reflektieren und auf dieser Grundlage in den Dialog mit anderen Beteiligten zu treten. Das Modell sieht drei Schritte vor: Zunächst definieren die Anwender:innen, welche Eigenschaften sie der KI zuschreiben. Anschließend identifizieren sie die relevanten Akteur:innen sowie deren Beziehungen untereinander. Im dritten Schritt werden jene Dimensionen des Evaluationssystems bestimmt, die im jeweiligen Kontext besonders relevant sind. In dem Modell wirken die zugeschriebenen Eigenschaften der KI auf die Beziehungen zwischen den Akteur:innen ein, was wiederum Auswirkungen auf die unterschiedlichen Dimensionen des Evaluationssystems haben kann. Dabei können durch den Einsatz von KI sowohl erwünschte als auch unerwünschte Veränderungen entstehen – zunächst auf der Beziehungsebene, in weiterer Folge aber auch im System und dessen Produktion selbst.

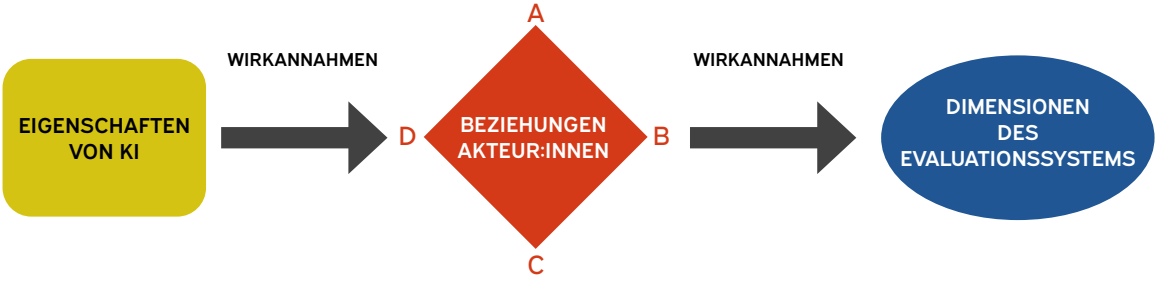


Abbildung 1: Schematische Darstellung des Wirkmodells von KI auf ein Evaluationssystem: KI hat Eigenschaften, die die Akteur:innenbeziehungen verändern, was wiederum Auswirkungen auf die unterschiedlichen Dimensionen des Evaluationssystems hat.

In den folgenden Abschnitten erläutern wir, welche Charakteristika wir den einzelnen Elementen unseres Modells zugeschrieben haben und zeigen schlaglichtartig, welche Reflexionspunkte sich daraus ergeben haben. Entlang der Modellstruktur bewegen wir uns dabei Schritt für Schritt von links nach rechts und beginnen daher mit den Eigenschaften von KI. Eine detaillierte Darstellung des Modells mit Leitfragen findet sich im Anhang.

ANGENOMMENE EIGENSCHAFTEN VON KI

Bezüglich der Eigenschaften (generativer) KI mussten wir zur besseren Nachvollziehbarkeit eine starke Vereinfachung vornehmen. Für unser Modell haben wir uns auf drei zentrale, möglichst allgemein gehaltene Eigenschaften verständigt. Diese abstrahieren bewusst von konkreten Anwendungsfällen – etwa dem automatisierten Schreiben von Texten – um nicht an kurzfristige Leistungsstände gebunden zu sein. Das Modell bleibt damit offen für künftige Entwicklungen und lässt sich bei Bedarf auch mit spezifischeren KI-Eigenschaften ausfüllen.
Die Auswahl der Eigenschaften orientiert sich an der Arbeit von Dell'Acqua et al. (2023). Basierend auf dieser zeichnet sich generative KI im Unterschied zu klassischem maschinellen Lernen durch drei Merkmale aus: (1) überraschende, ungeplante Anwendungsmöglichkeiten, (2) das Potenzial, die Leistung von Mitarbeiter:innen direkt zu steigern, und (3) eine gewisse Undurchsichtigkeit, da sie auch plausible, aber fehlerhafte Ergebnisse erzeugen kann. Zusammengenommen bergen diese Eigenschaften das Potenzial, soziale und institutionelle Beziehungen zu verändern und weitreichend auf Arbeitsweisen einzuwirken. Im Folgenden wird auf diese drei Eigenschaften näher eingegangen.

Generative KI hat erstens überraschende Anwendungsmöglichkeiten, für die sie nicht speziell entwickelt wurde, und die sich im Laufe der Zeit schnell weiterentwickeln können. Einerseits wenn sich die Größe und Qualität des Modells verbessert, andererseits auch wenn sich das Nutzungsverhalten der Anwender:innen ändert. Obwohl sie als allgemeine Modelle ausgebildet sind, zeigen generative KIs spezialisiertes Wissen und Fähigkeiten als Teil ihres Trainingsprozesses sowie während des normalen Gebrauchs (Singhal et al., 2022; Boiko et al., 2023). Während auf technologischer Ebene weiterhin erhebliche Debatten über das Konzept der emergenten Fähigkeiten bestehen (Schaeffer et al., 2023), sind die **effektiven Fähigkeiten von KIs neuartig und unerwartet, weit verbreitet anwendbar und nehmen in kurzen Zeitspannen erheblich zu**. Aktuelle Arbeiten zeigen, dass KI auf einem hohen Niveau

in professionellen Kontexten von der Medizin bis zum Recht eingesetzt werden kann (Ali et al., 2023; Lee et al., 2023) und damit auch Einfluss auf diese Kontexte nimmt. Bisher ist außerdem mit jeder Generation von KI-Modellen eine erhebliche Verbesserung der Fähigkeiten festzustellen, womit mit jeder Iteration immer neue überraschende und ungeplante Anwendungsmöglichkeiten aus der praktischen Anwendung heraus entstehen können. Dieser Umstand bedeutet aber auch, das generative KIs und deren Gestaltungsfähigkeiten sowohl für Individuen als auch für Institutionen schwer fassbar sind und abseits von prinzipiellen Umgangsformen schwer zu verwalten sein können.

Weil generative KI grundsätzlich dazu in der Lage ist domänenspezifische Probleme zu lösen, hat sie zweitens die Fähigkeit, die Leistung von Personen direkt zu steigern, die diese Systeme nutzen, ohne dass umfangreiche organisatorische oder technologische Investitionen oder technische Expertise auf Seiten der Nutzenden erforderlich sind. Studien deuten auf direkte (individuelle) Leistungssteigerungen durch die Nutzung von KI hin, insbesondere bei Schreibaufgaben (Noy und Zhang, 2023) und Programmierung (Peng et al., 2023), sowie bei Ideenfindung und kreativer Arbeit (Boussioux et al., 2024; Girotra et al., 2023). KI dürfte daher gerade auch auf Berufsgruppen, die mit komplexen Aufgaben konfrontiert sind, Einfluss haben und hier Individuen die Möglichkeit bieten, rasch und unabhängig von den jeweiligen Arbeitgeber:innen Leistungssteigerungen zu erzielen – die sie dann entweder für sich oder das Unternehmen nutzen können.

Als dritte relevante Eigenschaft kennzeichnet generative KIs eine **relative Undurchsichtigkeit,** die darin besteht, dass KI-Modelle **inkorrekte, aber plausible Ergebnisse** produzieren können (Halluzinationen oder Fabulationen) und damit faktisch falsche Angaben machen. Diese relative Undurchsichtigkeit geht dabei über ein technisches oder methodisches Verständnis von Undurchsichtigkeit hinaus, weil selbst durch das Offenlegen von Quellcodes die Herleitung der Ergebnisse nicht komplett erklärt werden kann. Somit kann KI bei bestimmten Aufgaben zielführend sein, während sie **unter anderen Umständen auf vorher schwer vorhersehbare Weise versagt**. Das hat zur Folge, dass die Möglichkeiten, diese KI-Systeme sinnvoll zu nutzen, nicht von vorneherein abschließend von ihren Entwickler:innen dargestellt werden können, sondern durch fortlaufendes Ausprobieren und Fehlermachen der Benutzer:innen erarbeitet werden müssen – zum Beispiel durch den Austausch von Erfahrungen und Heuristiken über verschiedene Online-Foren (wie Benutzergruppen, Hackathons, Twitter-Feeds und YouTube-Kanäle). Für Organisationen kann diese Eigenschaft von KI bedeuten, dass sie niemals

genau wissen können, wie genau das Tool wirkt, das sie in ihre Arbeitsprozesse einführt und mit dieser relativen Undurchsichtigkeit umgehen lernen muss.

ERWEITERUNG: KI TRITT ALS ACTANT INS EVALUATIONSSYSTEM EIN

Für die Analyse möglicher Veränderungen im Evaluationssystem verstehen wir generative Künstliche Intelligenz (KI) als *Actant* und beziehen uns dabei lose auf Konzepte der Actor-Network-Theory nach Latour (2005). Zentral an diesem Zugang ist die Idee, dass auch nicht-menschliche Entitäten, etwa Technologien, Dokumente oder eben KI, als wirksame Elemente in sozialen Netzwerken auftreten können. Sie agieren nicht im klassischen Sinn, beeinflussen jedoch durch ihre Eigenschaften, Konfigurationen und Einsatzkontexte das Handeln anderer Akteur:innen. Vor diesem Hintergrund schreiben wir KI bestimmte Eigenschaften zu, die, ohne KI zu vermenschlichen, das Potenzial haben, bestehende Beziehungen zwischen Akteur:innen im Evaluationssystem zu verändern. Als Actant kann KI damit auf das Beziehungssystem und die darin angelegten Handlungsmöglichkeiten einwirken.

Technisch verstehen wir KI im weitesten Sinne als „die Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren“ [2]. Dabei basieren viele KI-Anwendungen auf statistischen Verfahren, insbesondere auf dem maschinellen Lernen. In diesem Diskussionspapier liegt der Fokus auf *generativer KI*, also jenen Systemen, die neue Texte, Bilder oder andere Inhalte erzeugen können – typischerweise gestützt durch große Sprachmodelle (Large Language Models, LLMs), wie sie beispielsweise in ChatGPT, Gemini, Claude oder LLama zum Einsatz kommen.

Im Unterschied zu anderen Formen von KI (etwa in der Bilderkennung oder Prozessoptimierung) erlauben generative KI-Modelle eine unmittelbare, interaktive Nutzung. Sie reagieren in Echtzeit auf Eingaben, kombinieren Informationen neu und erzeugen Inhalte, die direkt in die Kommunikation zwischen Akteur:innen eingebunden werden können. Gerade diese Eigenschaft macht generative KI für unsere Fragestellung besonders relevant: Als interaktiv einsetzbares Werkzeug kann sie nicht nur Aufgaben übernehmen, sondern auch kommunikative Praktiken, Deutungsmuster und Entscheidungsprozesse mitprägen - und damit Beziehungsgefüge nachhaltig beeinflussen.

2 https://www.europarl.europa.eu/topics/en/article/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used

Die im Folgenden zugeschriebenen Eigenschaften verstehen wir daher nicht als technische Beschreibung eines konkreten Tools, sondern als konzeptionelle Verallgemeinerung auf Basis gegenwärtig verfügbarer Systeme. Unser Ziel ist es, generative KI in dem Denkmodell als Typus zu fassen, der in seiner Handlungsrelevanz für Evaluationssysteme analysierbar und diskutierbar wird.

ANGENOMMENE BEZIEHUNGEN UND AKTEUR:INNEN

In diesem Diskussionspapier konzentrieren wir uns auf die Auswirkungen von KI auf die Beziehungsebene zwischen den unterschiedlichen Akteur:innen im Evaluationssystem. Um die möglichen Veränderungen durch den Eintritt generativer KI reflektieren zu können, müssen wir zunächst die bestehenden Strukturen und Interaktionen innerhalb des Systems beschreiben - vereinfacht, aber hinreichend differenziert, um plausible Veränderungen durch KI denkbar zu machen.

Im Zentrum des hier betrachteten Evaluationssystems stehen vier zentrale Gruppen: Auftraggebende, Auftragnehmende, Evaluierende sowie die Evaluierten. Diese Gruppen sind in sehr unterschiedlichem Maße formalisiert. So stellen die Evaluator:innen meist Individuen dar, die in institutionelle Kontexte eingebunden, aber nicht durch diese vollständig repräsentiert sind. Die Auftraggebenden und Auftragnehmenden hingegen sind in der Regel Organisationen, die formale Rollen und Mandate innehaben. Das Evaluationsobjekt selbst, also das, was evaluiert wird, kann wiederum auf ganz unterschiedlichen Ebenen angesiedelt sein: von einem einzelnen Programm bis hin zu einer gesamten Institution. Häufig handelt es sich um eine kollektive Leistung von Individuen innerhalb institutioneller Strukturen.

Diese Heterogenität bringt ein hohes Maß an Komplexität in die Beziehungsstruktur des Evaluationssystems. Um diese dennoch analysierbar zu machen, bedarf es gewisser konzeptueller Vereinfachungen. Unsere Darstellung basiert daher auf grundlegenden Annahmen, die die Komplexität der Akteur:innenbeziehungen reduzieren, ohne ihren wesentlichen Aussagegehalt zu verlieren. Konkret gehen wir von folgendem vereinfachten Beziehungsmuster aus: Die direkte Verbindung zwischen Evaluator:innen und der auftraggebenden Institution ist schwach ausgeprägt, da sie durch eine Vertreter:in der auftragnehmenden Institution vermittelt wird. Ebenso besteht keine unmittelbare Beziehung zwischen der auftragnehmenden Institution und den Evaluierten; vielmehr sind es die Evaluator:innen selbst, die mit den Evaluierten interagieren. Dieses vereinfachte Beziehungsgeflecht bildet die Grundlage unseres Modells und ist in Abbildung 2 visualisiert.

Auf dieser Basis kann anschließend untersucht werden, wie generative KI als neuer *Actant* in diese Struktur eintritt und welche Veränderungen dies für die Interaktionen und Handlungsmöglichkeiten der beteiligten Akteur:innen mit sich bringt.

Abbildung 2: Schematische Skizze der zentralen Akteur:innen und ihrer Beziehungen

DIE AKTEUR:INNEN IM ANGENOMMENEN EVALUATIONSSYSTEM

Für die Analyse möglicher Veränderungen durch generative KI ist es notwendig, die zentralen Akteur:innen im angenommenen Evaluationssystem zu benennen. Unsere Betrachtung beruht auf einer vereinfachten, aber analytisch hilfreichen Differenzierung in vier Hauptakteur:innen(gruppen), die jeweils spezifische Rollen, Perspektiven und Handlungsspielräume im Evaluationsprozess einnehmen:

Auftraggebende sind in diesem Kontext die Programmeigner oder Verantwortlichen einer Intervention - und damit tendenziell auch die Auftraggebenden deren Evaluation. Der Auftraggebende benötigt effiziente und präzise Evaluierungsergebnisse, um fundierte Entscheidungen treffen zu können. Der Bedarf an klaren Kommunikationswegen und einer verständlichen Interpretation von Evidenz ist sehr hoch. Neben der Datenqualität ist auch das Vertrauen in die Evaluationspraxis ein wichtiger Aspekt, weil die Auftraggebenden sich nicht nur auf die Daten selbst, sondern auch auf eine sinnvolle Kontextualisierung im Politikfeld, einen guten Umgang mit Stakeholdern und in gewissem Maße auf die Fähigkeit, eine noch nicht messbare Zukunft zu antizipieren und Szenarien oder Empfehlungen zu entwickeln, verlassen können müssen.

Auftragnehmende werden hier als die Institutionen verstanden, die Evaluationen in Auftrag nehmen und für deren Durchführung verantwortlich sind. Auftragnehmende benötigen Ressourcen und möglichst umfassenden Zugang zu möglichst hochwertigen Daten. Jede Evaluationsaktivität hat

zum Ziel, die zur Verfügung stehenden Personal- und Budgetressourcen wirtschaftlich einzusetzen oder sogar zu vermehren, indem etwa Wissen und Know-How aufgebaut oder Gewinne erzielt werden. Effizienzgewinne sind daher willkommen, bei gleichzeitigem Anspruch an die Qualität der Arbeit, da die Reputation hinsichtlich des erwarteten Vertrauens in die Evidenzproduktion und die Zuverlässigkeit der Methoden essenziell ist.

Evaluierende sind jene Akteur:innen, die die Evaluation als Individuen selbst durchführen, die Daten generieren bzw. aufbereiten und analysieren und Empfehlungen hinsichtlich der Evaluierungsfragen geben. Sie sind sowohl dem Auftraggebenden als auch den Auftragnehmenden (meist deren Arbeitgebende) verpflichtet und können gleichzeitig Erwartungen an die Arbeitsweisen und Inhalte ihrer Tätigkeit haben, was Aspekte wie Karriereambitionen, Team-Dynamiken, inhaltliche Eigeninteressen und individuelle Weiterentwicklung, usw. impliziert. Aufgrund ihrer Expert:innenrolle und Kontextwissen haben diese Akteur:innen trotz des Abhängigkeitsverhältnisses zum Arbeitgeber häufig ein sehr hohes Maß an Freiraum und Verantwortung und können daher auch Ausgangspunkt für Entwicklungen sein.

Bei **Evaluierten/Eigner:innen des Evaluationsobjekts** handelt es sich, je nach Kontext, um die Intervention bzw. das Programm selbst, um eine Institution oder Organisationseinheit, oder um die Performance einer einzelnen oder mehrerer Personen, die evaluiert werden. Evaluierte müssen nachvollziehen können, wie KI in den Bewertungsprozess integriert wird. Transparenz und Kommunikation können entscheidend sein, um das Vertrauen der von Evaluation Betroffenen zu gewährleisten.

Akteur:innen-gruppe	Typische Rolle	Institutionali-sierungsgrad	Typische Interessen	Berührungs-punkte mit KI
Auftraggebende	Definition von Zielen, Finanzie-rung, Steuerung des Evaluwations-prozesses	Hoch (Organisationen mit klaren Mandaten)	Relevante Erkenntnisse, Steuerungswis-sen, Legitimation	Analyseunter-stützung, Entscheidungs-vorlagen durch KI-Outputs etc.
Auftrag-nehmende	Formelle Auf-tragserfüllung, Ressourcen-management, Schnittstelle zu Evaluator:innen	Mittel bis hoch (Organisationen mit Evaluations-kompetenz)	Qualitätsvolle Durchführung, Projekterfolg, Reputation	Integration von KI in Arbeitsprozesse, Koordination von KI-Einsatz etc.
Evaluator:innen	Datenerhebung, Analyse, Bewer-tung und Bericht-erstattung	Gering bis mittel (meist Einzelper-sonen oder kleine Teams)	Fachliche Integrität, methodische Qualität, Unabhängigkeit	Nutzung von KI für Textgenerie-rung, Datenana-lyse, Synthese etc.
Evaluierte	Bereitstellung von Daten, Interaktion mit Evaluator:innen, Objekt der Unter-suchung	Sehr unterschiedlich (individuell, projektbezogen oder institutionell)	Faire Darstellung, Einfluss auf Bewertung, Transparenz	Interaktion mit KI-gestützten Erhebungsinst-rumenten oder Bewertungen

Tabelle 3: Übersicht über die Akteur:innen eines Evaluationssystems und deren zentralen Eigenschaften

ANGENOMMENE DIMENSIONEN DES EVALUATIONSSYSTEMS

Das Evaluationssystem verstehen wir als ein Beziehungsnetzwerk mit wiederkehrenden Interaktionen zwischen bekannten Akteur:innen, das sich über die Zeit hinweg stabilisiert hat. Die etablierten Rollen und Interaktionsmuster bilden die Grundlage einer „Evaluationskultur", die das System prägt. Für dieses Diskussionspapier gehen wir stark vereinfachend von einem in sich geschlossenen System aus und berücksichtigen nur jene Akteur:innen, die in unserem Modell explizit benannt wurden. Diese Akteursgruppen können aber von Anwender:innen beliebig erweitert oder weiter eingeschränkt werden.

In jedem Fall bildet die Summe der Akteur:innen, ihrer Aktivitäten und Interaktionen das Evaluationssystem. Wir nehmen für unser Modell zumindest **vier Eigenschaftsdimensionen des Evaluationssystems** an, die alle Akteur:innen betreffen bzw. deren Handlungsrahmen definieren:

- Dimension 1: Hoher Anspruch an **Vertrauen** in die Prozesse, die Evidenz erzeugen. Hohe Erwartungen an den vertrauenswürdigen Umgang mit Daten und Evidenz
- Dimension 2: **Zuverlässigkeit** der Wissensproduktion und Reproduzierbarkeit
- Dimension 3: **Geschwindigkeit** und (Kosten-)Effizienz der Evidenzproduktion für rasche Entscheidungsfindung
- Dimension 4: Gleichzeitig werden schwer objektivierbare **Kontextualisierung** und Branchenwissen benötigt, das seit je her eine Art „Blackbox" im Evaluationssystem begriffen werden kann

Wie im Modell beschrieben gehen wir davon aus, dass die Anwendung von KI Auswirkungen auf einzelne oder alle Akteur:innegruppen und deren Beziehungen hat, die in weiterer Folge Auswirkungen auf diese (oder andere) Dimensionen des Evaluationssystems haben kann.

KI ALS DISRUPTION FÜR DAS EVALUIERUNGSSYSTEM

Die zuvor für das Modell definierten Eigenschaften von KI, entfalten in ihrer Kombination ein erhebliches Störpotenzial für etablierte Systeme. Dell'Acqua et al. (2023) sprechen in diesem Zusammenhang von einer „jagged technological frontier", einer ausgefransten Grenze, an der sich die Fähigkeiten von KI ungleichmäßig über verschiedene Bereiche der Wissensarbeit erstrecken. Während KI in manchen Feldern menschliche Leistungen übertrifft, bleibt sie in anderen weit hinter den Erwartungen zurück und dies oft auf schwer vorhersehbare Weise.

Dieses asymmetrische Leistungsprofil schafft neue Handlungsspielräume für einzelne Nutzer:innen oder Organisationen. Wer KI entlang dieser Grenze geschickt einsetzt, kann produktive Vorteile erzielen - etwa durch schnellere Analyseprozesse oder effizientere Textproduktion. Solche Vorteile können jedoch systemische Nebenwirkungen erzeugen, wenn sie für andere Akteur:innen intransparent bleiben oder nicht replizierbar sind. Misstrauen, Unsicherheit und ungleiche Machtverhältnisse können die Folge sein. Ebenso kann ein unbedarfter oder unangemessener Einsatz von KI die Qualität von Beziehungen und Bewertungen untergraben.

Vor diesem Hintergrund begreifen wir KI wie beschrieben als **Actant**, der in ein bestehendes Beziehungssystem eintritt und dieses durch seine Eigenschaften verändert. Die Effekte entstehen nicht durch intentional agierendes Verhalten,

sondern durch die Art und Weise, wie KI als Werkzeug genutzt, interpretiert und in soziale Prozesse eingebunden wird.

Gerade das Evaluationssystem ist in dieser Hinsicht ein besonders sensibler Bereich: Es basiert stark auf etablierten Interaktionsmustern, Vertrauen und Reputationslogik. Gleichzeitig steht es unter dem Druck, Effizienz zu steigern und neue technologische Möglichkeiten zu nutzen. Die Spannung zwischen Stabilität und Innovationsdruck macht das System besonders anfällig für die disruptiven Effekte generativer KI und gerade deshalb erscheint wichtig, diese genauer zu analysieren.

MODELL ZUM VERSTÄNDNIS DES EINTRETENS VON KI IN DAS EVALUATIONSSYSTEM

In dem Modell stehen die drei zuvor hergeleiteten Eigenschaften von KI, (1) unerwartete Anwendungsmöglichkeiten, (2) direkte Leistungssteigerung der Mitarbeiter, (3) relative Undurchsichtigkeit, den vier von uns definierten Zieldimensionen des Evaluationssystem, (1) Vertrauen, (2) Zuverlässigkeit, (3) Geschwindigkeit und (4) Kontextualisierung - gegenüber. In dieser Betrachtung können die zuvor vorgestellten Akteur:innen sowie ihre Handlungs- und Reaktionslogiken im Umgang mit KI zu einem Gesamtmodell des Evaluationssystems ergänzt werden. Das Modell soll Anwender:innen dabei als Reflektionsinstrument unterstützen, über relevante Fragen Wissenslücken und Gestaltungsmöglichkeiten zu identifizieren.

In unserer eigenen Betrachtung gehen wir davon aus, dass das Disruptionspotenzial von KI als Actant für das Evaluationssystem variieren kann und davon abhängt, wie stark bestimmte Eigenschaften der KI ausgeprägt sind. Wie diese Ausprägung aussieht, hängt einerseits mit den Eigenschaften und Fähigkeiten der KI zusammen, andererseits aber auch davon, wie sie eingesetzt wird oder werden darf. Deswegen ist es denkbar, dass die Einführung von denselben KI-Tools zu unterschiedlichen Zwecken und in unterschiedlichem Ausmaß erfolgt und daher das Disruptionspotenzial von KI nicht einförmig, sondern vielmehr in Szenarien zu denken ist. Um diesen Umstand zu berücksichtigen, wird von zwei idealisierten Positionen von KI als Actant in einem System ausgegangen – einer schwächeren und einer stärkeren:

- **„einfache Anwendnung von KI“**: KI-Tools, werden dazu genutzt einfache Aufgaben zu lösen und Fragen, die bisher schon bearbeitet wurden, effizienter lösbar machen
- **„komplexe Anwendung von KI“**: KI Modelle, werden dazu genutzt komplexe Aufgaben und Fragen zu lösen, die bisher nicht stellbar waren, und ermöglichen es, gänzlich neue Wege der Problemlösung zu beschreiten.

Die Begriffe „einfach“ und „komplex“ werden hier nicht als Eigenschaften für die angewendete Technologie selbst verwendet. In beiden Anwendungsfällen können im Grunde die gleiche Technologie zur Anwendung kommen, aber die Anforderungen an den Umgang anders sein: Während bei „einfachen“ Anwendungen zur Effizienzsteigerung routinemäßige Aufgaben an die KI ausgelagert werden können und die Reliabilität der Ergebnisse von erfahrenen Evaluierenden einigermaßen abschätzbar bleiben sollte, ist die Annahme, dass bei Fragestellungen, die bisher ohne KI nur mit großem Aufwand oder gar nicht gelöst werden konnten, die Möglichkeiten zur Reliabilitätsprüfung geringer sind.

AUSWIRKUNGEN DES EINTRETENS VON KI AUF DIE AKTEUR:INNEN UND REAKTIONSMÖGLICHKEITEN

Dadurch, dass die Akteur:innen im Evaluationssystem zueinander in Beziehung stehen, werden sie selbst dann von KI betroffen sein, wenn sie diese selbst gar nicht einsetzen (oder einsetzen dürfen). Dadurch entsteht eine Situation, in der alle Akteur:innen unabhängig von ihrer eigenen Position zu KI gefordert sind, sich mit KI auseinanderzusetzen und Standpunkte bzw. Strategien zu entwickeln. Damit dies positiv gelingt, gilt es sich gemeinsam über die notwendigen, die erwünschten und die unerwünschten Veränderungen und Anpassungen zu verständigen.

Um diesen Prozess zu unterstützen, stellen wir im nächsten Abschnitt basierend auf dem entwickelten Modell und der ausformulierten Charakteristika beispielhaft Annahmen zur Auswirkung von KI auf die einzelnen Akteur:innen auf. Hierbei stand das Formulieren von möglichen Strategien der jeweiligen Akteur:innen im Fokus, wobei grundsätzlich keine Wertung zwischen „offensiver“ oder „defensiver“ Strategie vorliegt. Am Beginn

des jeweiligen Abschnitts werden die getroffenen Annahmen offengelegt und auch welche Dimension(en) des Evaluationssystems, welche Eigenschaft von KI und welche Beziehung zwischen Akteur:innen für den jeweiligen Fall als besonders relevant eingeschätzt wurden und deswegen besonders in die Analyse eingeflossen sind. Das Ergebnis sind dabei ausgefüllte Tabellen mit möglichen positiven und negativen Auswirkungen (jeweils für die einfache und die komplexe Anwendung von KI) und daraus resultierenden strategischen Optionen für die jeweilige Akteursgruppe. Dabei handelt es sich nicht um definitive Aussagen, sondern um Anregungen für einen Diskurs. Anwender:innen können (basierend auf anderen Annahmen) zu anderen Schlussfolgerungen gelangen, die aber durch das Modell strukturiert und explizit gemacht werden können, was für eine lösungsorientierte Debatte notwendig ist.

AUFTRAGGEBENDE

Für Politik- oder Interventionseigner:innen ist die Herausforderung besonders groß, dass sie keine direkte Kontrolle darüber haben, welche Akteur:innen KI an welcher Stelle und auf welche Weiße einsetzen. Allerdings sind sie in besonderem Maße davon abhängig, dass das **Vertrauen in das Evaluationssystem** hoch ist, da die Ergebnisse der Evaluationen die Basis für Entscheidungen legen (z.B. für oder gegen die Weiterführung eines Programms). Dieses Vertrauen ist für die **Auftraggebenden gerade gegenüber den Evaluierten** wichtig, da diese beiden Gruppen in der Regel über lange Zeiträume hinweg miteinander zusammenarbeiten müssen, womit der falsche Einsatz von KI für diese Beziehung ein hohes Risiko birgt. Die Auftraggebenden könnte sich daher gezwungen sehen, Regeln oder Kontrollen einzuführen, was wiederum Kosten erzeugt, oder die Entwicklungspotenziale der neuen Technologie hemmt bzw. wiederum bei den anderen Akteur:innen wie den Evaluator:innen im System negativ aufgenommen wird.

Auswirkungen für die Akteursgruppe „**Auftraggebende**“ bei Eintritt von KI ins Evaluationssystem			
		POSITIV	**NEGATIV**
Disruptionspotenzial	**„einfache“ Anwendung von KI**	Durch die **direkte Leistungssteigerung** Individuen – hier insbesondere bei Evaluator:innen – können Evaluationsergebnisse schneller bereitgestellt werden. Dabei können auch größere Datenmengen effizienter verarbeitet werden, was in Summe zu Kosteneinsparungen im Evaluationssystem führen kann.	Auch bei einer einfachen Anwendung von KI kann die identifizierte **relative Undurchsichtigkeit** bei den Ergebnissen zu Vertrauensverlusten führen. Gerade bei größer werdenden Datenmengen wird es für Auftraggebende und Evaluierte auch bei einfachen Anwendungen immer schwerer nachzuvollziehen wie die Ergebnisse zustande gekommen sind. Auftraggebende könnten dadurch in die Evaluationspraxis eingreifen und hier Regelsysteme vorgeben, mit unbekannten Konsequenzen für die langfristige Einführung von KI in das Evaluationssystem.
	„komplexe“ Anwendung von KI	KI ermöglicht bisher schwer fassbare Aspekte besser zu verstehen und zu bewerten. Dadurch können Programme aus gänzlich anderen, aber relevanten Perspektiven beleuchtet werden, wodurch deren Impact erhöht werden kann. Hierfür muss der Auftraggebende entweder ein gutes Verständnis über die Anwendungsmöglichkeiten entwickeln (und hier auch aktiv **unerwartete Anwendungsmöglichkeiten** suchen), oder darauf vertrauen, dass Auftragnehmende diese Informationen teilen mit ihm teilen.	Die hohe Bedeutung **unerwarteter Anwendungsmöglichkeiten**, zusammen mit der **relativen Undurchsichtigkeit** können dazu führen, dass die Integrität der Aussagen dieser Modelle nicht ausreichend gegeben ist und daher nicht als Entscheidungsgrundlage für Auftraggebende genutzt werden können/sollten. Diese müssen selbst hohe Kapazitäten zu Interpretation und Einordnung der Ergebnisse aufbauen, um diese sicher nutzen zu können.

Strategien der Akteursgruppe „**Auftraggebende**“ im Umgang mit KI	
OFFENSIV	**DEFENSIV**
Auftraggebende schaffen aktiv Anreize, um die Auseinandersetzung im eigenen Haus mit KI anzuregen und dadurch selbst aktive Kapazitäten bei der Anwendung von KI aufzubauen. Es werden außerdem Foren mit Evaluierten geschaffen, um sich über die Einsatzmöglichkeiten und Grenzen von KI auszutauschen, aber auch um besonders interessante Anwendungsgebiete für das jeweilige Programm zu identifizieren. Dabei nehmen Auftraggebende eine gestalterische Rolle ein und definieren beispielsweise Erfolgsindikatoren und Ziele für den Einsatz von KI im Evaluationssystem. Diese werden über die nächsten Jahre (gemeinsam mit den anderen Akteur:innen) im Feld kontinuierlich beobachtet, um etwaige Schwierigkeiten früher identifizieren zu können. Auftraggebende setzen sich für den Einsatz von KI ein und schaffen transparente Handlungsoptionen für die anderen Akteur:innen falls dieser in einzelnen Fällen nicht ideal gelingt.	Auftraggebende nehmen eine passive Rolle bei der Einführung von KI im Evaluationssystem ein und überlassen es anderen Akteursgruppen diese Technologie in die aktive Anwendung zu bringen. Solange das Vertrauen in das Evaluationssystem durch den Einsatz von KI nicht erodiert. Hierfür ist aber ein Austausch mit den anderen Akteursgruppen notwendig, um hier insbesondere ein Bild zur Lage des Vertrauens im Evaluationssystem zu haben. Auftragnehmende nehmen zwar keine aktive gestalterische Rolle ein, schulen ihre Mitarbeiter:innen aber regelmäßig über die jeweils bekannten Anwendungsmöglichkeiten von KI.

AUFTRAGNEHMENDE

Auftragnehmende befinden sich in einer besonderen Position, da sie sowohl in Konkurrenz zu anderen Auftragnehmenden stehen, die Vorgaben der Auftraggebenden erfüllen müssen und sich in Bezug auf KI aber auch in ein neues Abhängigkeitsverhältnis zu ihren Angestellten begeben könnten. **Geschwindigkeit** (effiziente Evidenzproduktion) und die Fähigkeit zur **Kontextualisierung** (Verfügbarkeit von Expert:innen/ Branchenwissen) sind für sie zwei relevante Dimensionen, um im Evaluationssystem bestehen zu können. Allerdings können Mitarbeiter:innen KI-Modelle oft schneller nutzen als ihre Organisationen, da diese frei zugänglich sind und ohne Einbindung der Arbeitgebenden angewendet werden können. Dadurch kann sich die Abhängigkeit von spezialisierten Fachkräften weiter verstärken. Auf der anderen Seite können Auftragnehmende strukturelle Kapazitäten in Bezug auf KI aufbauen, die es ihnen ermöglicht, spezielle kontextualisierte Aufgaben besser zu lösen (entweder durch das Trainieren eigener Modelle oder dem Aufarbeiten von Daten extra für KIs), wofür ein Ressourceneinsatz notwendig ist, der für Individuen nicht zu stemmen ist. In der Konkurrenz zu anderen Auftragnehmenden dürfte es essenziell sein, KI effektiv in Arbeitsprozesse

zu integrieren und auch die dadurch gewonnen Effizienzgewinne für sich nutzbar zu machen. Zugang zu hochwertigen Daten und die Fähigkeit, die KI-Ergebnisse zu interpretieren, können entscheidend für deren Erfolg am Evaluationsmarkt sein, und daher zu Konkurrenz und weniger zu Kooperation auf dieser Ebene führen.

Auswirkungen für die Akteursgruppe „**Auftragnehmende**“ bei Eintritt von KI ins Evaluationssystem			
		POSITIV	**NEGATIV**
Disruptionspotenzial	**„einfache“ Anwendung von KI**	Die **direkte Leistungssteigerun**g durch KI bei Mitarbeiter:innen – etwa beim Schreiben der Berichte – führt zu einer effizienteren Nutzung der vorhandenen Ressourcen. Diese können entweder in andere Teile der Evaluation fließen, oder aber auch in den Aufbau zusätzlicher Kapazitäten/Fähigkeiten fließen, wodurch sich die Evaluationsqualität dauerhaft erhöht.	Die **direkte Leistungssteigerung** durch KI wird von Mitarbeiter:innen nicht mit der Organisation geteilt, sondern zum eigenen Vorteil genutzt. Dadurch steigt das Misstrauen der Auftragnehmenden gegenüber den eigenen Mitarbeiter:innen was zu neuen Überwachungsmaßnahmen führen kann. Eine Qualitätssicherung ist unter solchen Bedingungen für Auftragnehmende nur mehr erschwert möglich, obwohl diese durch die **relative Undurchsichtigkeit** besonders notwendig wäre. Möglicherweise muss die Organisation auch zwischen den unterschiedlichen Interessen der Mitarbeiter:innen mediieren, was zusätzliche Ressourcen bindet.
	„komplexe“ Anwendung von KI	Die komplexe Anwendung von KI bietet die Möglichkeit, hochwertige Daten für anspruchsvolle Analysen zu nutzen. Weil Fachwissen vermehrt durch die KI abgedeckt werden kann, können neue Berufsbilder entstehen, die vermehrt über Querschnittskompetenzen verfügen. Das erleichtert die Zusammenarbeit über unterschiedliche Bereiche in und außerhalb einer Organisation und ermöglicht neue Herangehensweisen und Evaluationsformen.	Die Anpassung an den Einsatz komplexer KI-Modelle erfordert möglicherweise starke Umstrukturierungen in einer Organisation (Neue Berufsbilder benötigen beispielsweise neue Arbeitsverträge und/oder Schulungen). Zusätzlich dazu muss die technische Infrastruktur und Expertise aufgebaut werden, ohne Gewissheit über den letztlichen Nutzen zu haben. Es bleibt unklar, ob Mitarbeiter:innen **unerwartete Anwendungsmöglichkeiten** (ähnlich wie direkte Leistungssteigerungen) für ihren eigenen Vorteil nutzen, wodurch die Organisationen ein erhöhtes Risiko haben, dass sich ihre Investition nicht rentiert.

Strategien der Akteursgruppe „**Auftragnehmende**“ im Umgang mit KI	
OFFENSIV	**DEFENSIV**
KI wird als Möglichkeit verstanden, aktiv mit neuen Methoden und Ansätzen zu experimentieren und Mitarbeiter:innen ein entsprechendes Umfeld geboten. Im Zentrum der Bemühungen steht das Entwickeln von besseren Lösungen für vorhandene Probleme und das Identifizieren von neuerdings erreichbaren Lösungen für bisher nicht bearbeitbare Probleme. Auftragnehmende bauen dabei auf eine gute Beziehung zu Expert:innen. Es werden deswegen vertrauensbildende Maßnahmen getroffen, um den Stress, den diese Entwicklung unter Mitarbeiter:innen auslösen kann, zu reduzieren (und dadurch Widerstand zu vermeiden). Dabei wenden Auftragnehmende Ressourcen auf, um eine entsprechende Expertise in der Organisation aufzubauen. Die vertrauensbildenden Maßnahmen sind außerdem so angelegt, dass sie zum aktiven Teilen neuer Anwendungsmöglichkeiten in der Organisation ermutigen.	Auftragnehmende überlassen die Anwendung von und den Umgang mit KI weitestgehend ihren Mitarbeiter:innen und vermitteln minimale Prinzipien für deren Anwendung oder auch Verbote. Die aktiven Investitionen in diesem Bereich werden begrenzt gehalten und man setzt weiterhin auf „traditionelle“ Formen der Evaluation. Dieser Ansatz wird dabei auch als Unterscheidungsmerkmal zur Konkurrenz im Feld gesehen und soll Auftraggebenden und Evaluierten das notwendige Vertrauen in die angewandten Methoden geben und Mitarbeiter:innen das Vertrauen in die Organisation. Das führt dazu, dass es bei entsprechenden Aufträgen auch ein KI Nutzungsverbot für die Mitarbeiter:innen geben kann.

EVALUIERENDE

Für Evaluierende bietet KI die Möglichkeit, ihren Arbeitsalltag selbstständig effizienter zu Gestalten und Aufgaben an diese auszulagern. Je besser Evaluierende die KI navigieren können, desto größer ist der potenzielle Nutzen für sie, weshalb sie ein hohes Interesse haben können, **unerwartete Anwendungsmöglichkeiten** zu identifizieren. Dabei fällt ihnen aber potenziell viel Verantwortung zu, da sie (wenn es hierfür keine institutionellen Lösungen gibt), auch mit der **relativen Undurchsichtigkeit** von KI umgehen müssen und hier entscheiden müssen, ob ein KI generiertes Ergebnis den Standards des Evaluationssystems ausreicht (und das teilweise in Fällen, wo es noch keine genauen Standards und Praktiken gibt). Eine unstrukturierte Einführung von KI durch die Evaluierenden selbst kann also dazu führen, dass diese in hohen Maß über die **Zuverlässigkeit** und das **Vertrauen** im Evaluationssystem entscheiden und mit dieser Verantwortung überlastet werden könnten.
Eine gute Beziehung zwischen Auftragnehmenden und den Evaluierenden und der Aufbau entsprechender Strukturen kann hier Abhilfe schaffen und diese Umverteilung der Verantwortung abmildern. Allerdings müssen

dabei auch die persönlichen Interessen der Evaluierenden, berücksichtigt werden, die beispielsweise Effizienzgewinne nicht mit den Arbeitgebenden teilen wollen, weil sie diese für relative Vorteile innerhalb des Teams oder für Karrierepfade verwenden können. Umgekehrt könnten die Mitarbeiter:innen ihre Arbeitsplätze vom Einsatz von KI-Werkzeugen bedroht sehen, was zu Konflikten mit einem sehr der KI zugeneigten Arbeitgebenden führen kann.

Auswirkungen für die Akteursgruppe „**Evaluierende**“ bei Eintritt von KI ins Evaluationssystem			
		POSITIV	**NEGATIV**
Disruptionspotenzial	**„einfache“ Anwendung von KI**	Die **direkte Leistungssteigerung** kann für Evaluierende zu Entlastung von Routineaufgaben durch Automatisierung, die diese auch noch weitestgehend selbstständig und auf ihre Bedürfnisse zugeschnitten umsetzen können. Dadurch bestünde die Möglichkeit zur Fokussierung auf interessantere Aufgaben wie beispielsweise komplexere Analysen und Interpretationen.	**Direkte Leistungssteigerungen** sind nur für Individuen greifbar, die sich mit der Technologie auseinandersetzen, was zu „Generationenkonflikten“ in einem Team oder einer Organisation führen kann. Personen die KI für ein effizienteres Arbeiten nutzen wollen, stoßen auf Personen, die sich verweigern, wodurch erstere aufgehalten werden und letztere sich bedroht fühlen können.
	„komplexe“ Anwendung von KI	Die komplexe Anwendung von KI kann von Evaluierenden dazu genutzt werden, um weit über ihre Kernexpertise hinaus Unterfangen durchführen zu können und sich so laufend neues Branchenwissen aneignen können. Durch einen daraus resultierenden andauernden Austausch mit Kolleg:innen können Herausforderungen wie die **relative Undurchsichtigkeit** von KI gemildert werden und **unerwartete Anwendungsmöglichkeiten** breiter geteilt werden. Es entsteht ein neues Verständnis der Arbeitsgestaltung mit neuen Freiräumen und Möglichkeiten.	Die komplexe Anwendung von KI führt zu Überforderung, wodurch die **relative Undurchsichtigkeit** von KI vermehrt zu schwer zu identifizierenden Fehlern führt. Weil Evaluierende sich nicht austauschen können (etwa bei einem KI-Verbot durch die Organisation) oder wollen (eigene Vorteile), fällt dieses Korrektiv weg. Die Unklarheit, wer die Verantwortung bei der Einführung und Verwendung von KI-gestützten Analysen trägt, führt zu keiner, oder nur versteckter Nutzung.

Strategien der Akteursgruppe „**Evaluierende**“ im Umgang mit KI	
OFFENSIV	**DEFENSIV**
Evaluierende werden zu Treiber:innen und Gestalter:innen der Anwendung von KI. Sie treiben dabei sowohl die Entwicklung von kodifizierten Richtlinien aber auch impliziten Verhaltensregeln basierend auf der realen Praxis in den Institutionen und dem System voran. Dadurch können Routinearbeiten sicher an KI ausgelagert werden, womit mehr Zeit für anspruchsvolle Arbeiten bleibt. Das Eigeninteresse und die Gestaltungsfähigkeit führen dazu, dass regelmäßige Schulungen und Updates über neue Entwicklungen im Bereich der KI gerne angenommen werden und Austauschformate mit Kolleg:innen gepflegt werden.	Evaluierende geraten im Effizienzwettbewerb unter Druck und fühlen sich gezwungen KI als „Abkürzung“ ad-hoc und unreflektiert zu nutzen, um mit dieser Herausforderung umzugehen. Die Anwendung von KI findet nur dort statt, wo es aufgrund des entstandenen Effizienzdrucks nicht mehr anders geht. Ein Austausch zwischen den Individuen kann durch den Zeitdruck, aber auch dadurch erschwert werden, dass es keinen offiziellen Einsatz der Technologie gibt und jene, die Vorteile durch die KI haben, diese nicht aufgrund von Verboten verlieren wollen. Mangelnde Bereitschaft unter diesen Bedingungen Schulungen anzunehmen beeinträchtigt die Qualität der Evaluierung weiter.

EVALUIERTE

Für die Evaluierten könnten durch erwartete und **unerwartete Anwendungsmöglichkeiten** von KI neue Formen der Evaluation entstehen, die eine vielschichtige und breitere Evaluation ihrer Tätigkeiten zulässt. Das kann das **Vertrauen** in das Evaluationssystem steigern, weil die Erfolge eines Programms differenzierter dargestellt werden können. Hierfür müssen die Evaluierten aber von der **Zuverlässigkeit** dieser neuen Evaluationsformen überzeugt sein. Gleichzeitig ermöglicht KI potenziell auch das interne Einführen eigener robuster, neuartiger Monitoringsysteme, die Evaluierten einerseits eine Reflexion der eigenen Tätigkeiten ermöglichen und andererseits eine gute Grundlage für eine etwaige Evaluation bieten. Diese internen Monitoringsysteme können von den Evaluierten aber auch dazu genutzt werden, „Gegendarstellungen“ zu externen Evaluationen zu erstellen und damit die Beziehung zwischen Evaluierten und Evaluierenden erschweren. Außerdem kann so auch das **Vertrauen** in das Evaluationssystem gesenkt werden, wenn etwa Auftraggebende nicht einschätzen können, welcher Evidenz sie trauen können. Hier können also neue Gräben geöffnet werden, weshalb ein guter Ausgleich zwischen den Interessen der Evaluierten und der Auftraggebenden getroffen werden muss, der eine gute Zusammenarbeit mit Auftragnehmenden und Evaluierenden ermöglicht.

Auswirkungen für die Akteursgruppe „**Evaluierte**“ bei Eintritt von KI ins Evaluationssystem			
		POSITIV	**NEGATIV**
Disruptionspotenzial	**„einfache“ Anwendung von KI**	Neben der schnelleren Verfügbarkeit von Ergebnissen, kann KI durch die **direkte Leistungssteigerung** Evaluierte dabei unterstützen die Ergebnisse von Evaluationen in eine leichter operationalisierbare Form zu Übersetzten, die damit schneller auch ihren Weg in neue Handlungspraktiken findet.	Die **relative Undurchsichtigkeit** von KI führt dazu, dass Evaluierte bereits durch die einfache Anwendung von KI nicht mehr nachvollziehen können, wie es zu dem jeweiligen Ergebnis gekommen ist. Dieser Effekt kann durch die von KI erzeugte pseudoobjektive Sprache weiter verstärkt werden. Um mit dieser Situation umgehen zu können, sind die Evaluierten gezwungen, eigene Kapazitäten in dem Bereich aufzubauen, die Ressourcen von den eigentlichen Aktivitäten abziehen, ohne einen größeren Mehrwert zu generieren.
	„komplexe“ Anwendung von KI	Das Zusammenspiel von internem Monitoring und externer Evaluation, die durch eine KI mediiert wird, eröffnet gänzlich neue Formen der Zusammenarbeit zwischen Evaluierten und Evaluierenden. Dabei können beide Seite von noch unentdeckten, **unerwarteten Anwendungsmöglichkeiten** der anderen profitieren und gemeinsam das Risiko der relativen Undurchsichtigkeit reduzieren.	Die komplexe Anwendung von KI führt zu Überforderung, wodurch die **relative Undurchsichtigkeit** von KI vermehrt zu schwer zu identifizierenden Fehlern führt. Weil Evaluierende sich nicht austauschen können (etwa bei einem KI-Verbot durch die Organisation) oder wollen (eigene Vorteile), fällt dieses Korrektiv weg. Die Unklarheit, wer die Verantwortung bei der Einführung und Verwendung von KI-gestützten Analysen trägt, führt zu keiner, oder nur versteckter Nutzung.

Strategien der Akteursgruppe „**Evaluierte**“ im Umgang mit KI	
OFFENSIV	**DEFENSIV**
Die Evaluierten nutzen KI einerseits als Möglichkeit die eigene Reflektions- und Lernfähigkeit zu steigern und bauen so ein Verständnis für die Einsatzmöglichkeiten der Technologie auf. Andererseits öffnen sich die Evaluierten gegenüber externen Bemühungen, KI zur Evaluation einzusetzen, und gestalten die Rahmenbedingungen dafür aktiv mit. Hierfür wurde ein partizipativer und transparenter Prozess entwickelt, um KI in den Evaluationsprozess einzubeziehen. Es gibt eine Kultur der klaren Kommunikation über den Einsatz von KI und wie die Ergebnisse verwendet werden.	Die Evaluierten nutzen KI teilweise zum internen Monitoring, teilen diese Ergebnisse aber nicht mit externen Evaluierenden. So senken sie die eigene Nachfrage/ den Bedarf nach Lernen durch externe Evaluationen – riskieren damit aber auch eine institutionelle Blindheit bzw. eine höhere Abhängigkeit von den KI-affinen Mitarbeiter:innen. Bei externer Anwendung bleiben die konkreten Prozesse aber unklar und wenig definiert. Dadurch kommt es zwar zu weniger externen Evaluationen mit komplexem Einsatz von KI, allerdings wird so auch nach Außen hin kein positiver Umgang mit dieser Technologie entwickelt. Die passive Haltung, kann zu Missverständnissen über den Einsatz von KI führen, und die Offenheit in der Organisation gegenüber der Technologie insgesamt reduzieren. Sollten Auftraggebende auf eine KI-basierte Evaluation bestehen, müsste die Organisation sehr schnell Prozesse dafür entwickeln, die möglicherweise fehleranfällig sind.

Die hier dargestellten Überlegungen sind keineswegs abgeschlossen, oder vollständig. Sie zeigen aber wie vielfältig die Überlegungen der unterschiedlichen Akteur:innen sein können und dass es hilfreich sein kann sie in simple Einzelaspekte herunterzubrechen. Unser Modell bietet dabei eine Möglichkeit dafür, dass dies strukturiert gelingen kann. Basierend auf den hier angerissenen Strategien, ließe sich nun beispielsweise in einem nächsten Schritt betrachten, wo es strategische Überschneidungen zwischen den Akteur:innen gibt und wo sie wohl sehr unterschiedliche Interessen in Bezug auf den Einsatz von KI haben. Dadurch entsteht eine gute Grundlage, um in einen Diskurs zu gemeinsamen Lösungsansätzen zu treten.

RELEVANTE FRAGEN

Wie bereits beschrieben, lässt sich unser Modell nutzen, um systemrelevante Fragen zum Einsatz von generativer KI in Evaluationskontexten zu identifizieren. In der Auseinandersetzung mit dem Modell hat sich insbesondere die Perspektive auf das Evaluationssystem als soziales Gefüge mit stabilen Beziehungen als zentral erwiesen. Der Eintritt von KI als *Actant* verändert dieses Gefüge, oft subtil und nicht für alle Akteur:innen unmittelbar sichtbar.

Besonders relevant erschien dabei die Eigenschaft generativer KI, zu unerwarteten Anwendungsmöglichkeiten zu führen. Im Unterschied zur oft diskutierten Undurchsichtigkeit oder Leistungssteigerung wird dieses Potenzial bislang wenig beachtet – obwohl es gerade auf individueller Ebene erhebliche Auswirkungen entfalten kann. Durch kreative oder nicht intendierte Nutzung kann Handlungsmacht von der institutionellen auf die individuelle Ebene verschoben werden. Dies bringt neue Chancen, aber auch Risiken für Vertrauen, Transparenz und Gleichbehandlung mit sich. Evaluator:innen könnten ihre Entdeckungen von unerwarteten Anwendungsmöglichkeiten von KI verstecken, um daraus resultierende Effizienzgewinne für sich zu nutzen, besonders in einem Umfeld, das KI skeptisch gegenübersteht. Gleichzeitig würden sie aber stark von einem Austausch mit ihren „Peers" profitieren, was eine höhere Organisationsstruktur und entsprechende Rahmenbedingungen beispielsweise seitens der Auftragnehmende erfordert. Hierbei würde es helfen, wenn Prinzipien für die Nutzung von KI, darauf Rücksicht nehmen und Experimentierfreude erlauben, während sie aber gleichzeitig Transparenz im sozialen System (aus einer nicht-technischen Perspektive) gewährleisten. Für Auftragnehmende ergibt sich daraus eine doppelte Herausforderung: Sie sollen Potenziale von KI nutzen, sind dabei jedoch auf die informelle Nutzung ihrer Mitarbeiter:innen angewiesen, tragen aber institutionell die Verantwortung. Eine fehlende Abstimmung über Normen und Prinzipien der KI-Nutzung könnte langfristig das gesamte Evaluationssystem destabilisieren.

Das Beispiel zeigt, wie sich technologische Innovationen auf soziale Beziehungen auswirken können, auch ohne dass dies beabsichtigt ist. Vor diesem Hintergrund ergeben sich zentrale Fragen, mit denen sich die fteval-Community weiter beschäftigen sollte:

- Wie stellen wir uns ein positives Bild von KI durchzogenes Evaluationssystem vor?
- Auf welcher Ebene ist eine Konkurrenz um unerwartete Anwendungsmöglichkeiten bzw. KI im Allgemeinen im Evaluationssystem wünschenswert und innovationsfördernd? Auf welcher Ebene ist sie schädlich?
- Wer soll letztlich von den Anwendungsmöglichkeiten von KI im Evaluationssystem profitieren? Evaluator:innen? Auftragnehmende? Auftraggebende? Die Evaluierten? Wie kann das etabliert werden?
- Was benötigt eine vertrauensvolle Beziehung zwischen Evaluator:innen und Auftragnehmenden bzw. den Akteursgruppen im Allgemeinen?

- Wie können Austauschprozesse zu neuen Anwendungsmöglichkeiten von KI insgesamt im Evaluationssystem organisiert werden?
- Welche Rahmenbedingungen müssen Auftragnehmende erschaffen, damit Evaluator:innen offen mit den von ihnen gefundenen unerwarteten Anwendungsmöglichkeiten von KI umgehen?
- Wie können Institutionelle Akteur:innen von der Experimentierfreude ihrer individuellen Mitarbeiter:innen mit KI am besten profitieren, ohne zu hohes Risiko einzugehen, was das Vertrauen in die Ergebnisse betrifft?
- Wie muss Verantwortung organisiert werden, damit Risiken für das Evaluationssystem oder bestimmte Akteur:innen minimiert werden, ohne die Experimentierfreude zu zerstören?
- Wie kann eine heimliche Ausnutzung von neuen Anwendungsmöglichkeiten reduziert und Individuen bzw. Institutionen gleichzeitig für ihren Einsatz belohnt werden?
- Wie soll das Teilen von neuen Praxiswissen zu KI im Evaluationssystem gestaltet sein? Wie lässt sich das angestrebte Ziel erreichen?
- Woher stammt das Feedback von Kolleg:innen oder Vorgesetzten?

Alle diese Fragen haben ethische und rechtliche Implikationen, gehen aber weit über diese hinaus. Gelingt es den Akteur:innen im Evaluationssystem nicht eine vertrauensvolle Arbeitsweise im Umgang mit der neuen Technologie aufzubauen, werden die negativen Externalitäten für das Gesamtsystem zunehmen. Damit dürfte das Vertrauen in die Ergebnisse sinken und es würde weniger attraktiv für Auftraggebende auf diese Art der Entscheidungsgrundlage zurückzugreifen. Der gelungene Umgang mit diesen scheint daher ein sehr relevantes und in seiner Gänze noch gar nicht erfasstes Thema für die kommenden Jahre zu sein.

AUSBLICK: WAS MUSS JETZT BEDACHT WERDEN?

Dieses Diskussionspapier soll dazu bewegen, sich strukturiert mit den Auswirkungen von KI auf die Beziehungen im Evaluationssystem auseinanderzusetzen. Wir alle werden von den Entwicklungen in diesem Feld betroffen sein. Gleichzeitig scheint es auch ein inhärenter Bestandteil generativer KI zu sein, dass wir alle genauso Treibende dieser Entwicklung sind. Darin liegt viel Potenzial aber auch viel Risiko für das Evaluationssystem. Aus unserer Perspektive wird die übergeordnete Frage der nächsten Jahre daher die folgende sein: Welche Beziehungen zwischen Akteur:innen wollen wir beibehalten, wie sie gerade sind, und welche sollen sich im Sinne eines verbesserten Evaluationssystems verändern?

Dabei darf KI nicht nur als Technologie verstanden werden, sondern auch als gesellschaftlicher Wille zum Wandel. Während man sich KI als Technologie wahrscheinlich noch Jahre – wenn nicht länger – verschließen kann, wird die KI als gesellschaftlicher Wille sich mit aller Kraft (und unabhängig von den Folgen) Zugang zu den vielen unterschiedlichen Systemen schaffen. Der rapide technische Wandel zusammen mit dem gesellschaftlichen Willen zur Anwendung machen es notwendig eine Zukunftsvorstellung zu erstellen, unabhängig von den konkreten Fähigkeiten der KI im Jetzt und auch losgelöst von der eigenen Haltung zur Technologie. Sonst können selbst grundlegend „positive“ Aspekte dieser KI die Beziehungen von Akteur:innen in einem System nachhaltig stören und hier zu ungewollten Disruptionen führen.

Diese Arbeit in Form eines schematischen Modells mit begleitenden Reflexionsfragen bietet eine Grundlage für die strukturierte Diskussion über die Integration von KI in die Evaluationslandschaft. Als Akteur:in ist es entscheidend, sich frühzeitig mit den potenziellen Veränderungen auseinanderzusetzen und an der Gestaltung eines zukunftsfähigen Evaluationsprozesses teilzunehmen. Gerade weil es im Evaluationssystem keine zentralen Entscheidungsträger:innen gibt, die das System als Ganzes überblicken, sondern es sich vielmehr aus der Summe einzelner Akteur:innen zusammensetzt, die sich gegenseitig - bewusst oder unbewusst – beeinflussen, ist der bewusste Umgang wichtig. Wir laden alle Betroffenen und Interessierten ein, aktiv an dieser Diskussion teilzunehmen und gemeinsam die Weichen für eine effektive und ethische Evaluationspraxis unter Verwendung von KI zu stellen.

DANKSAGUNG

Diese Arbeit basiert auf konzeptuellen Überlegungen der Kolleg:innen Thomas Palfinger und Susanne Beck vom Open Innovation in Science Center der Ludwig Boltzmann Gesellschaft, bzw. letztere mittlerweile Warwick Business School, University of Warwick. Diese wurden innerhalb der Arbeitsgruppe zu künstlicher Intelligenz[3] in der Evaluation der Österreichischen Plattform für Forschungs- und Technologiepolitikevaluierung (fteval) für den Evaluierungskontext weiterentwickelt. In der Untergruppe, die sich Juli 2023 bis May 2024 mit den veränderten Akteur:innenbeziehungen beschäftig hat, waren Alexander Daminger (WIFO), Charlotte D'Elloy (Technopolis Group | Austria), Elisabeth Froschauer-Neuhauser (AQ Austria), Felix Gaisbauer (DLR Projektträger), Tina Olteanu (FWF), Thomas Palfinger (LBG-OIS), Vitaliy Soloviy (AIT), Michael Strassnig (WWTF) und Isabella Wagner (fteval). Die gesamte Arbeitsgruppe bestand zusätzlich aus (weiteren) Vertreter:innen von AIT, aws, FFG, FWF, KMU Forschung Austria, ÖAWI, Technopolis Group | Austria, WIFO und ZSI. Vielen Dank für alle Ideen, Beiträge und das Feedback!

REFERENZEN

Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Sullivan, P. L. Z., ... & Asaad, W. F. (2022). Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery, 10-1227.

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V. & Lakhani, K. R. (2024). The Crowdless Future? Generative AI and Creative Problem Solving (July 01, 2024). Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-005, Available at SSRN: https://ssrn.com/abstract=4533642 or http://dx.doi.org/10.2139/ssrn.4533642

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality (September 15, 2023). Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, Available at SSRN: https://ssrn.com/abstract=4573321 or http://dx.doi.org/10.2139/ssrn.4573321

3 https://fteval.at/fteval-arbeitsgruppe-zu-kuenstlicher-intelligenz-in-der-evaluierung/

Europäische Commission (2023). Artificial Intelligence Act – Nicht finaler Gesetzesvorschlag: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Fisher, M., Smiley, A. H., & Grillo, T. L. H. (2022). Information without knowledge: the effects of Internet search on learning, Memory, 30:4, 375-387, DOI: 10.1080/09658211.2021.1882501

Girotra, K., Meincke, L., Terwiesch, C. & Ulrich, K. T., Ideas are Dimes a Dozen (2023). Large Language Models for Idea Generation in Innovation (July 10, 2023). The Wharton School Research Paper Forthcoming, Available at SSRN: https://ssrn.com/abstract=4526071 or http://dx.doi.org/10.2139/ssrn.4526071

Haupt, S. E., Gagne, D. J., Hsieh, W. W., Krasnopolsky, V., McGovern, A., Marzban, C., ... & Williams, J. K. (2022). The history and practice of AI in the environmental sciences. Bulletin of the American Meteorological Society, 103(5), E1351-E1370.

Konya, A., & Nematzadeh, P. (2024). Recent applications of AI to environmental disciplines: A review. Science of The Total Environment, 906, 167705.

Latour, B. (2005). An introduction to actor-network-theory. Reassembling the social.

Lee, P., Bubeck, S. & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 388:13), 1233–1239.

Noy, S. & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Available at SSRN: https://ssrn.com/abstract=4375283.

Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). The impact of ai on developer productivity: Evidence from github copilot. arXiv preprint arXiv:2302.06590.

Ruthotto, L., & Haber, E. (2021). An Introduction to Deep Generative Modeling. arXiv:2103.05180. https://arxiv.org/abs/2103.05180 arxiv.org

Schaeffer, R., Miranda, B. & Koyejo, S. (2023). Are emergent abilities of Large Language Models a mirage?. arXiv preprint arXiv:2304.15004.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2022). Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138.

Stahl, B. C. (2023). Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems. Scientific Reports, 13(1), 7586.

Stockmann, R. (2004). Was ist eine gute Evaluation? Einführung zu Funktionen und Methoden von Evaluationsverfahren. (CEval-Arbeitspapier, 9). Saarbrücken: Universität des Saarlandes, Fak. 05 Empirische Humanwissenschaften, CEval - Centrum für Evaluation. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-118018

Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. A Nature survey finds that scientists are concerned, as well as excited, by the increasing use of artificial-intelligence tools in research, NEWS FEATURE in: Nature 621, 672-675 (2023), DOI: https://doi.org/10.1038/d41586-023-02980-0

AUTHORS

THOMAS PALFINGER
Open Innovation in Science Center, LBG
Nußdorferstraße 64, 1090 Vienna, Austria
Email: Thomas.Palfinger@lbg.ac.at
ORCID: 0000-0001-6933-5114

FELIX GAISBAUER
DLR Projektträger
Sachsendamm 61, 10829 Berlin
Email: felix.gaisbauer@dlr.de
ORCID: 0000-0002-1285-1246

ISABELLA WAGNER
Plattform fteval
Linke Wienzeile 246, c/o ZSI, 1150 Vienna, Austria
Email: wagner@fteval.at
ORCID: 0000-0002-2772-6771

SUSANNE BECK
Warwick Business School
Scarman Road
Coventry, CV4 7AL
Email: Susanne.Beck@wbs.ac.uk
ORCID: 0000-0002-2448-6194

ANHANG – THEMENSPEICHER DER ARBEITSGRUPPE

DETAILLIERTE DARSTELLUNG UND LEITFRAGEN DES MODELLS

Zur Vereinfachung gehen wir in unserem Modell von einem einseitigen Einwirken von KI auf das Evaluationssystem und die darin enthaltenen Beziehungen aus. KI ist in diesem Modell ein neuer Actant, der in ein existierendes System tritt, und hier potenziell Veränderungsprozesse oder Disruption für bestehende Beziehungen auslöst. Wir fokussieren dabei auf den Einfluss von KI auf die Beziehungen zwischen den Akteur:innen und die darauf aufbauenden Einflüsse auf Teilaspekte des Evaluationssystems.

Abbildung 4 Zusammenwirken von KI, Beziehungen und System

In Abbildung 2 sind die grundlegenden Mechanismen des Modells grafisch dargestellt. Ausgehend von einer definierten Eigenschaft von KI (im abgebildeten Beispiel: unerwartete Anwendungsmöglichkeiten), werden Wirkannahmen dieser Eigenschaft für ausgewählte Akteur:innen getroffen und betrachtet, wie sich diese Wirkannahmen auf deren Beziehung auswirken können. Davon ausgehend kann anschließend die Auswirkung auf eine oder mehrere der vier von uns beschriebenen Dimensionen des Evaluationssystems (im Beispiel: Vertrauen) herausgearbeitet werden. So kann man ausgehend von den Eigenschaften von KI über Zwischenschritte die möglichen Auswirkungen des Systemeintritts von KI ableiten. In der Anwendung dieses Models haben wir uns immer wieder die folgenden Fragen gestellt, die uns durch den gesamten Prozess in der Auseinandersetzung mit KI begleitet haben:

1. Geht man davon aus, dass eine bestimmte Eigenschaft von KI Einfluss auf die existierende Beziehungspraxis von relevanten Akteur:innen in einem System hat?

2. Welche Annahmen lassen sich auf die Wirkung von KI aufstellen und welche davon erscheinen besonders relevant?
3. Wie beeinflussen die Wirkungsannahmen das vorhandene Beziehungsgefüge zwischen den Akteur:innen?
4. Welche Konsequenzen haben die Änderungen im Beziehungsgefüge auf eine oder mehrere Dimensionen des Evaluationssystems?
5. Sind diese Konsequenzen aus der eigenen Perspektive wünschenswert? Wie müssten sie gestaltet sein, um wünschenswert zu sein?

Dabei leiten diese Fragen Schritt für Schritt von links nach rechts durch das Modell und unterstützen dabei sich jeweils auf den zentralen Aspekt zu fokussieren. Die fünfte Frage leitet dann schon wieder aus dem Model hinaus in den Diskurs und regt dazu an sich mit der im Modell skizzierten Vorstellung auseinanderzusetzen und, wie wir es im nächsten Abschnitt selbst tun werden, Positionen und Strategien für einen gewünschten Zustand zu formulieren.

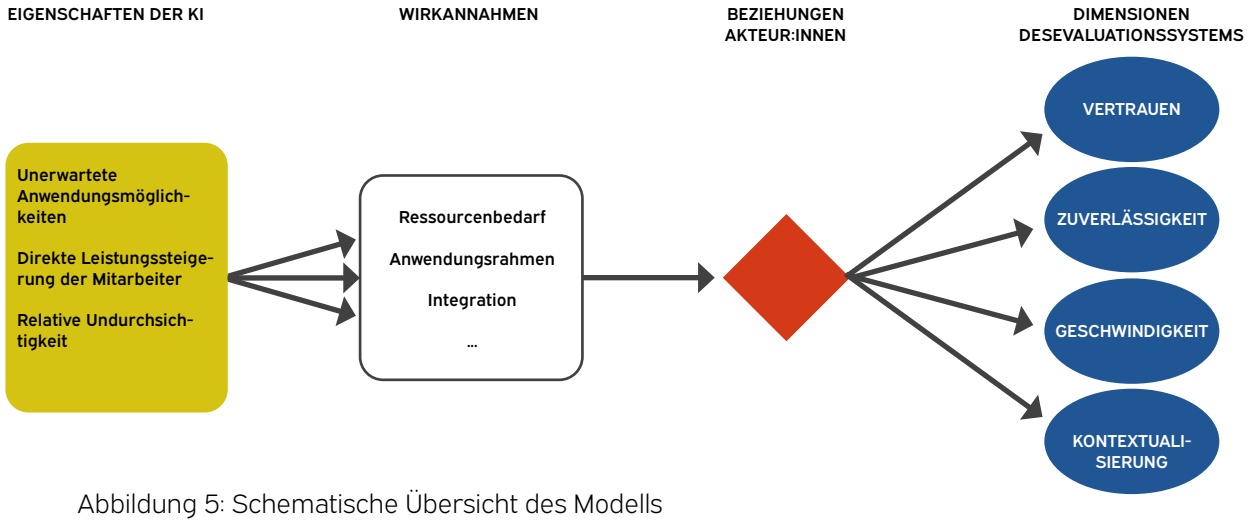


Abbildung 5: Schematische Übersicht des Modells

Dabei kann man die Komplexität des Modells nach Bedarf beliebig variieren bzw. Stück für Stück ausbauen. Man kann sich wie in Abbildung 3 eine einzelne Eigenschaft von KI herausziehen und hier Wirkannahmen aufstellen, die man dann für einzelne Akteur:innen und eine ausgewählte Dimension des Evaluationssystems anwendet. Alternativ kann man aber auch wie in Abbildung 3 mehrere Eigenschaften von KI, Akteur:innen und Dimensionen in Zusammenhang zueinander ausspielen, oder ausgehend von einfacheren Überlegungen dazu übergehen, diese schrittweise komplexer werden zu lassen.

Durch das schrittweise Befüllen unseres Modells soll eine strukturierte Auseinandersetzung vor dem Hintergrund begrenzender Parameter ermöglicht werden, um sich trotz der Komplexität mit den vielseitigen Möglichkeiten strukturiert auseinandersetzen zu können. Das Modell eignet sich dabei für zwei Aufgabenstellungen: es kann dabei unterstützen relevante Fragestellungen aufzuwerfen, oder aber bei der Formulierung von strategischen Überlegungen (hierbei sowohl für einzelne Akteur:innen, als auch für Systeme mehrerer Akteur:innen) helfen. Beides kann die Handlungsfähigkeit der Akteur:innen im Evaluationssystem gegenüber neuen Entwicklungen im Bereich KI erhöhen bzw. dabei helfen koordinierte Entscheidungen zugunsten des Systems zu treffen.

ÜBERLEGUNGEN ZU AUSWIRKUNG VON KI AUF EVALUATIONSSYSTEM

Eigenschaft 1: Hoher Anspruch an Vertrauen in die Prozesse, die Evidenz erzeugen. Hohe Erwartungen an den vertrauenswürdigen Umgang mit Daten und Evidenz

1) **Unerwartete Anwendungsmöglichkeiten**: Diese Eigenschaft von KI fordert dieses Vertrauen heraus, weil im Sinne der „jagged frontier" ein ständiger Auslotungsprozess im Gange ist, welche Aufgaben wie von KI sinnvoll übernommen werden können und ob und wo genau eine funktionale Arbeitsteilung zwischen Mensch und KI bei einer Evaluation auch tatsächlich umgesetzt wird.

2) **Direkte Leistungssteigerung der Mitarbeiter:innen**: Diese Eigenschaft von KI verschiebt die Handlungsfähigkeit und damit potentiell die Verantwortlichkeit von einer Institution auf die Ebene der Mitarbeiter:innen. Reputation einer Unternehmung könnte verstärkt von vermuteter Vertrauenswürdigkeit einzelner Personen abgelöst werden.

3) **Relative Undurchsichtigkeit**: Diese Eigenschaft fordert Vertrauen ebenfalls heraus. Die „Blackbox" KI birgt zu jedem Zeitpunkt die Gefahr, dass plausible, aber erfundene oder inkorrekte Behauptungen nicht als solche erkannt werden.

Eigenschaft 2: Zuverlässigkeit der Wissensproduktion und Reproduzierbarkeit

1) **Unerwartete Anwendungsmöglichkeiten**: Diese Eigenschaft könnte die Zuverlässigkeit der Wissensproduktion insofern erhöhen, als dass mehr

Werkzeuge für die Erhebung und Auswertung von Daten, insbesondere großer Mengen qualitativer Daten, verfügbar sind und relativ einfach angewendet werden können.

2) **Direkte Leistungssteigerung der Mitarbeiter:innen**: Diese Eigenschaft könnte die Transparenz der Wissensproduktion senken, weil individuelle Mitarbeiter:innen mitunter weniger auf den methodischen Austausch mit Kolleg:innen angewiesen sind. Das war aber auch bisher schon der Fall, wenn es Mitarbeitende mit Expertise und Fähigkeiten gibt, die an einer Institution nicht redundant vorhanden sind. Es ist auch denkbar, dass der Einsatz von KI-Methoden zur Erledigung repetitiver Datenüberarbeitungen den Effekt haben, dass man durch den Einsatz Zeit gewinnt, sich methodisch und inhaltlich auszutauschen.

3) **Relative Undurchsichtigkeit**: Diese Eigenschaft steht im grundsätzlichen Widerspruch zur Zuverlässigkeit. So sind beispielsweise GPTs dazu programmiert, nie die gleichen Antworten zu liefern und daher ist eine Reproduzierbarkeit im bisherigen Sinne gar nicht möglich. Außerdem entwickeln sich die Angebote und damit auch die generierbaren Ergebnisse ständig weiter und die frei verfügbaren Zugänge zu GPTs können durch private Betreibende jederzeit entzogen oder gedrosselt werden, weshalb die individuellen Anwenderinnen hinsichtlich der Reproduzierbarkeit ebenfalls vor Problemen stehen könnten.

Eigenschaft 3: Geschwindigkeit und (Kosten-)Effizienz der Evidenzproduktion für rasche Entscheidungsfindung

1) **Unerwartete Anwendungsmöglichkeiten**: Eine sehr positive Eigenschaft mit dem Potenzial, die Geschwindigkeit und Effizienz rasch zu steigern.

2) **Direkte Leistungssteigerung der Mitarbeiter:innen** (**MA**): Schlüssel dazu, die Effizienzsteigerung zu verwirklichen. Die MA müssen die neuen Anwendungsmöglichkeiten erkennen und rasch(er) umsetzen können (als die Konkurrenz im System).

3) **Relative Undurchsichtigkeit:** Die tatsächlichen Effizienzgewinne könnten schwer fassbar sein, weil mitunter hohe Prüfungskosten der Reliabilität der Daten oder hohe Kosten für die technische Infrastruktur oder technische Unterstützung entstehen könnten.

Eigenschaft 4: Gleichzeitig werden schwer objektivierbare Kontextualisierung und Branchenwissen benötigt, das seit je her eine Art „Blackbox" im Evaluationssystem begriffen werden kann

1) **Unerwartete Anwendungsmöglichkeiten**: Sowohl Auftraggebende als auch Auftragnehmende können GPTs zur Reflexion branchenspezifischer Fragestellungen nutzen, etwa zum Brainstorming. Es könnten sich ebenso eine Vielzahl an Möglichkeiten des Monitorings und bis zu einem gewissen Grade auch automatisierte Analysen mit niederschwellig verfügbaren KI-Methoden entwickeln. Eine weitere spekulative Aussicht könnte sein, dass sich das interne Monitoring selbst ändern wird und durch nicht-perfekte, aber akzeptable Analysen einer lokalen KI ersetzt werden wird.

2) **Direkte Leistungssteigerung der Mitarbeiter:innen**: Individuen haben nun die theoretische Möglichkeit, auf das schriftlich verfügbare und somit **explizite Branchenwissen** zuzugreifen und können sich damit möglicherweise Teile von Branchenwissen in einem Maße aneignen oder simulieren, was bisher mehrere Jahre Arbeitserfahrung benötigt hätte. Gleichzeitig ist es sowohl für die KI als auch für branchenferne Evaluierende oder Politikgestaltende unmöglich **implizites Branchenwissen**, Umgangsformen, Gruppendynamiken oder historische Verläufe zu verstehen und zu interpretieren. Diese Arten von kontextualisiertem Wissen werden nach wie vor nur von erfahrenen Akteur:innen eingebracht werden können. Die Herausforderung für Institutionen ist es zu verstehen, wann was gebraucht wird und ggf. Kontrollpunkte in neuen Abläufen einzubauen. Es gibt aber auch Hinweise, dass die Bedeutung von Branchenwissen abnimmt, wie es am Beispiel von online Suchmaschinen bereits erforscht wird (vgl. Fisher et al. 2022).

3) **Relative Undurchsichtigkeit**: Auch wenn es irgendwie möglich ist, solche Kontrollpunkte effektiv in den Evaluationsprozess einzubauen, wird es trotz entsprechender Dokumentation sehr schwierig sein zu erkennen, welche Aspekte einer Arbeit aus der Feder einer KI und welche von Menschen stammen und zu beurteilen, was davon legitime Aussagen sind, und was nicht.

ÜBERLEGUNGEN ZU AUSWIRKUNG VON KI AUF BEZIEHUNGEN UND PROZESSE

Die Integration von KI-basierten Tools in die Evaluationslandschaft der FTI-Politik in Österreich wird nicht nur bestehende Prozesse beeinflussen, sondern auch neue Akteursdynamiken und Interaktionsmuster schaffen. Hier versuchen wir basierend auf unseren Grundannahmen zu antizipieren, welche Auswirkungen der Eintritt von KI ganz grundsätzlich auf die Beziehungen zwischen den Akteur:innen haben könnten. Diese Überlegungen sollen den Diskurs rund um die Technologie weiter anregen und sind in keiner weise abschließend.

Auftraggebende-Auftragnehmende:
Traditionell haben Auftraggebende die Evaluierenden beauftragt und erhielten von ihnen Berichte. Auftragnehmende sind daher auch rechenschaftspflichtig über die Methoden und Technologien, die zur Datenverarbeitung eingesetzt werden. Mit den neuen Technologien könnten Auftraggebende aber auch direkter in den Evaluierungsprozess eingreifen, indem sie bestimmte Aspekte der KI-Modelle mitgestalten.

Eine verstärkte Zusammenarbeit zwischen Auftraggebenden und -nehmenden während der Konzeption und Entwicklung von KI-Modellen könnte zu maßgeschneiderten Monitoring und Evaluationssystemen führen, die besser auf die Bedürfnisse der Auftraggebenden zugeschnitten sind.

Auftragnehmende-Evaluierende:
Auftragnehmende sind für die Durchführung der Evaluierung verantwortlich, während es relativ viele Freiheiten für die evaluierenden Individuen gibt. Aufgrund der Eigenschaft von KI, dass sich unerwartete Anwendungsmöglichkeiten ergeben und sie direkt von Mitarbeiter:innen eingesetzt werden, sind Arbeitgebende mitunter im Unklaren, ob und in welcher Form die Evaluator:innen mit KI arbeiten. Bei intransparenter Handhabung könnte die Konkurrenz unter den Mitarbeiter:innen steigen, weil sich der Leistungsdruck verstärkt. Kolleg:innen, die KI nicht rechtzeitig oder ausreichend einsetzen, könnten im kollegialen Wettbeweb abgehängt werden.

Evaluierende müssen ihre Fähigkeiten erweitern, um nicht nur traditionelle Evaluierungskompetenzen, sondern auch Kenntnisse im Umgang mit KI-Systemen zu entwickeln. Auftragnehmende müssen dafür sinnvolle Rahmenbedingungen und die entsprechenden Unterstützungsmaßnahmen schaffen.

Evaluierender-Evaluierungsobjekt:
Evaluierte könnten sich weigern, Daten zur Verfügung zu stellen, wenn es Zweifel über die transparente Bearbeitung gibt. Deshalb müssen Evaluierende transparent und offen kommunizieren, wie KI in den Evaluierungsprozess integriert wird. Dies könnte eine stärkere Einbindung von Evaluierungsobjekten in den Bewertungsprozess erfordern.

Partizipative Ansätze, bei denen Evaluierungsobjekte Einblick in die Funktionsweise der eingesetzten KI erhalten und Feedback dazu geben können, könnten zu fundierteren und akzeptierteren Evaluierungsergebnissen führen.

Evaluierungsobjekt-Auftraggebende:
Die Ansprüche von Geldgeber:innen in Bezug auf Monitoring und Impact Assessment könnten mit den besseren Möglichkeiten zur Automatisierung steigen. Agenturen und Programme könnten in einen noch höheren Leistungsdruck in Bezug auf Rechenschaftslegung fallen, weshalb sich die Evaluierten gegen die neuen Möglichkeiten des Monitorings wehren könnten.

fteval JOURNAL
for Research and
Technology Policy Evaluation

December 2024, Vol. 55, pp. e6, 1-29
DOI: 10.22163/fteval.2024.663

NACHHALTIGER, SCHNELLER, TIEFER, BREITER: FTI-BEITRÄGE ZU NACHHALTIGEN TRANSFORMATIONSPROZESSEN ERFASSEN

AUTOR:INNEN: JAKOB KOFLER, ISABELLA WAGNER UND HARALD WIESER

BEITRAGENDE: ALFRIED BRAUMANN, LISA BOHUNOVSKY, INES HINTERLEITNER, HARALD HOCHREITER, MARLENE LANG, LISA KOCH, INA MATT, SABINE MAYER, SUSANNE MEYER, SASCHA SARDADVAR, KLAUS SCHUCH

DOI: 10.22163/FTEVAL.2024.663

ABSTRACT

Dieser Artikel fasst die Ergebnisse einer fteval-Arbeitsgruppe zur Wirkungsmessung von FTI-Beiträgen zu nachhaltigen Transformationsprozessen zusammen. Es werden bestehende Zugänge in Österreich vorgestellt, zusätzliche Ansätze aus der Literatur diskutiert und darauf basierend Möglichkeiten aufgezeigt, wie Indikatorik und Monitoring weiterentwickelt werden können. Damit sollen erste Grundlagen geschaffen werden, wie Monitoring und Evaluation eine nachhaltige Transformation in Österreich umfassender abbilden und somit besser unterstützen können.

1. NACHHALTIGKEIT UND TRANSFORMATION IN MONITORING, EVALUATION UND LERNEN (MEL)

Der Forschungs-, Technologie- und Innovationspolitik (FTI-Politik) kommt eine zentrale Rolle in der Gestaltung und Begleitung von nachhaltigen Transformationsprozessen zu. Mit der steigenden Dringlichkeit und Bedeutung von tiefgreifenden Veränderungen in Wirtschaft und Gesellschaft wird von der FTI-Politik zunehmend erwartet, dass sie konkrete Beiträge zu gesellschaftlichen Herausforderungen wie der Bekämpfung von Klimawandel, Biodiversitätsverlust, Pandemien oder Armut leistet (Schot & Steinmueller, 2018). Die österreichische Bundesregierung hat darauf reagiert, indem Teile der bisherigen Forschungsförderung im Sinne einer 'transformativen Innovationspolitik' umgestaltet wurden: Neue FTI-Initiativen wurden ins Leben gerufen, neue Förderinstrumente wurden ermöglicht und neue Formen der Zusammenarbeit in der öffentlichen Verwaltung wurden etabliert (Ecker et al., 2023).

Mit den neuen Ambitionen einer transformativen Innovationspolitik sind viele Herausforderungen für die Gestaltung und Durchführung von Monitoringprozessen und Evaluierungen verbunden. Die effektive Unterstützung von Transformationsprozessen erfordert ein fundiertes Verständnis des Anwendungskontextes und der möglichen Wirkungsbeiträge der vorhandenen FTI-politischen Instrumente, sowie ein zeitnahes Feedback zu den bereits erzielten und noch nicht erzielten Wirkungen. Damit dies bereitgestellt werden kann, müssen die bestehenden Prozesse der Beschaffung und Bearbeitung von Informationen zum Teil grundlegend neu gedacht werden. Davon betroffen ist auch der ‚harte' Kern von Monitoring und Evaluierung: die Indikatorik bzw. die Messverfahren, die angewandt werden, um konkrete Leistungen, Aktivitäten oder Wirkungen zu erfassen.

Dieser Artikel setzt sich mit bestehenden FTI-Indikatoren zur Erfassung von Wirkungen für eine nachhaltige Transformation auseinander. Obwohl der Ansatz einer transformativen Innovationspolitik bereits vor einigen Jahren Eingang in die nationale und internationale FTI-Politik gefunden hat, besteht noch erheblicher Nachholbedarf in Hinblick auf die Wirkungsmessung (Biggeri & Ferrannini, 2020; Brodnik & Dinges, 2022). Brodnik und Dinges stellen zu bestehenden Indikatoren fest:

"*The existing sets of indicators […] embrace the concept of transformative innovation policy only to a very limited extent. While there are examples and initiatives of indicators that aim to systematically measure the influence of R&I activities on the realisation of overarching societal goals (such as the SDGs or Agenda 2030[1]), they are currently not well established. There is either a conceptual ambiguity or the data is currently neither available nor systematically collected*" (Brodnik & Dinges, 2022: 36f).

Fragen der Wirkungsmessung sind institutionenübergreifend von Interesse und betreffen sowohl interne Monitoringprozesse als auch externe Analysen durch Evaluator:innen oder Begleitorgane. Die *fteval – österreichische Plattform für Forschungs- und Technologiepolitikevaluierung* hat vor diesem Hintergrund eine Arbeitsgruppe ins Leben gerufen, um den Status quo und Handlungsbedarf in Bezug auf die Messung von FTI-Beiträgen zu nachhaltigen Transformationsprozessen zu erheben. An der Arbeitsgruppe beteiligten sich Expert:innen aus Verwaltung (*Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie - BMK*), Förderagenturen (*Forschungsförderungsgesellschaft - FFG, Forschungs- und Wissenschaftsfonds – FWF, Wirtschaftsagentur Wien*), Universitäten (*Allianz nachhaltige Universitäten, Universität für Bodenkultur*) und der Evaluation (*KMU Forschung Austria, WPZ Research, Zentrum für soziale Innovation, fteval*).

Der inhaltliche Austausch in der Arbeitsgruppe zeigte auf, wie unterschiedlich die Begriffe der Nachhaltigkeit und Transformation bisher aufgegriffen wurden. Die gemeinsame Auseinandersetzung mit den Anforderungen der jeweiligen Organisationen legte aber auch dar, wie schwierig es ist, bestehende Konzepte und Ansätze aus der Transformationsforschung für die Verwendung zu operationalisieren und damit ‚auf den Boden zu holen'. Bestehende Ansätze, die in der Evaluierungspraxis bereits Früchte getragen haben (z.B. Dinges et al., 2022; Janssen, 2019)narrowly targeted policy is back on the table. Recent advances in the fields of new industrial policy (NIP, stießen unter den beteiligten Personen auf wenig Resonanz. Viele der bisher in der Literatur diskutierten Indikatoren werden von konkreten Transformationstheorien (z.B. der Ansatz der transformative Outcomes, siehe Ghosh et al., 2021) abgeleitet, die auf einer für viele Anwendungskontexte zu hohen Flughöhe ansetzen und nicht ausreichend flexibel an die jeweiligen Anforderungen angepasst werden können.

Im vorliegenden Artikel verfolgen wir einen anderen Ansatz, indem wir konkrete Indikatoren und Monitoringsysteme zur Wirkungsmessung in den

1 https://sdgs.un.org/goals

Vordergrund stellen. Als Orientierung dient uns eine einfache Untergliederung von vier Dimensionen nachhaltiger Transformation: der *Nachhaltigkeit*, *Tiefe*, *Breite* und *Geschwindigkeit* von Veränderungsprozessen. Anhand dieses Schemas lässt sich verdeutlichen, welche Schwerpunkte in Monitoring und Evaluierung bisher in Bezug auf die Wirkungsmessung gesetzt wurden und welche Aspekte von nachhaltiger Transformation auf Basis bestehender Indikatoren nur unzureichend erfasst werden können. Es wird insbesondere deutlich, dass bereits einige Schritte in der Messung der ökologischen, sozialen und ökonomischen Wirkungen im Sinne eines umfassenden Nachhaltigkeitsverständnisses getätigt wurden, die bestehenden Indikatoren aber insbesondere in Bezug auf die drei anderen Qualitäten transformativen Wandels noch wenig Anhaltspunkte bieten. *Es bleibt somit weitgehend unklar, ob und inwieweit FTI-politische Maßnahmen zu einem tiefgreifenden, breiten und beschleunigten Wandel beitragen*.

Um ein Lagebild der Wirkungsmessung in der österreichischen FTI-Politik aus einer Perspektive nachhaltiger Transformation zu schaffen, gehen wir in den nächsten Abschnitten auf die folgenden Fragen ein:

Wie werden FTI-Beiträge zu Nachhaltigkeit und Transformationsprozessen in Österreich erfasst? Welche Ansätze der Wirkungsmessung werden jeweils verfolgt?

Welche spezifischen Lücken in den bestehenden Analyserastern erfordern den Einsatz neuer oder zusätzlicher Indikatoren, um den vielfältigen Wirkungen von FTI-Initiativen in der nachhaltigen Transformation gerecht zu werden?

2. TRANSFORMATIVE INNOVATIONSPOLITIK IN ÖSTERREICH

Die Implementierung einer transformativen Innovationspolitik baut in Österreich auf der langjährigen Tradition thematischer FTI-Programme auf, im Rahmen derer seit mehr als zwei Jahrzehnten Forschungsprojekte, Netzwerkaktivitäten und Kompetenzaufbau in Bereichen der nachhaltigen Entwicklung unterstützt werden (Weber & Kubeczko, 2023). Insbesondere in Bezug auf die ökologische Nachhaltigkeit wurden viele Programme auf die Erzielung von positiven Beiträgen zu Umwelt- und Klimaschutz ausgerichtet (Wieser et al., 2021). Beispiele, in denen Forschung und Innovation eine Richtung gegeben

wird, waren etwa die Programme „Mobilität der Zukunft“[2] oder „Vorzeigeregion Energie“[3] (Weber & Kubezcko, 2023).

Transformation involviert jedoch mehr als eine Richtungsvorgabe oder -änderung. Im Gegensatz zu Nachhaltigkeit (als Prinzip) beschreibt Transformation nicht, *warum* eine Veränderung passieren soll, sondern *wie* diese passieren kann (ein Prozess). Transformativer Wandel involviert häufig Veränderungen in den folgenden drei Dimensionen (z.B. Andersen et al., 2023):

- **Tiefe**: Diese Dimension bezieht sich auf systemische und strukturelle Veränderungen von Institutionen, Regeln und Infrastrukturen, die den Status quo aufrechterhalten. Hier wird nicht nur Innovation, sondern auch Exnovation berücksichtigt.
- **Breite**: Diese Dimension berücksichtigt die Vielfalt von Lösungen für unterschiedliche (lokale) Probleme sowie die diversen Zielgruppen, die erreicht und involviert werden. Dies beinhaltet auch Interdependenzen zwischen unterschiedlichen Systemen (beispielsweise dem Energiesystem und dem Mobilitätssystem), und betrachtet sowohl technologische als auch soziale Innovationen.
- **Geschwindigkeit**: In dieser Dimension geht es um die zeitliche Dynamik von Veränderungsprozessen, insbesondere in Bezug auf Möglichkeiten der Beschleunigung.

Um transformativen Wandel anzustoßen, sind somit neben der Bestimmung der Entwicklungsrichtung weitere Maßnahmen erforderlich, die strukturell wirken, kontextspezifische Anforderungen berücksichtigen und Impulse zur Beschleunigung von Veränderungsprozessen setzen. In den vergangenen Jahren wurden in diesem Zusammenhang einige neue Instrumente in der österreichischen Innovationspolitik eingeführt, wie Innovationslabore, regulatorische Freiräume, oder öffentlich-öffentliche Kooperationen (siehe hierzu der Abschnitt zu transformativer Innovationspolitik im Österreichischen *Forschungs- und Technologiebericht 2023*, Ecker et al., 2023). Darüber hinaus sind auch flankierende Maßnahmen – wie ein gemeinsames Lernen (d. h. der Wissens- und Erfahrungsaustausch zwischen Wissenschaft, Wirtschaft und weiteren Akteuren) sowie die gezielte Einbindung bisher wenig eingebundener Akteursgruppen (z. B. Bürgerinnen und NGOs) – von zentraler Bedeutung. Dabei bedarf es möglichst kurzer Lernschleifen, etwa durch regelmäßiges

2 https://www.ffg.at/mobilitaetderzukunft

3 https://www.ffg.at/vorzeigeregionenergie

Monitoring oder begleitende Evaluierungen, da Feedback aus ex-post-Evaluationen in dynamischen Situationen oft zu spät kommt.

In Österreich gibt es mehrere Institutionen, die transformative Innovationspolitik umsetzen. Ihre Initiativen und Programme sind darauf ausgerichtet, systemische Veränderungen in Wirtschaft und Gesellschaft, wie die Energiewende oder die Mobilitätswende zu unterstützen, um zentrale gesellschaftliche Herausforderungen wie Klimawandel und Nachhaltigkeit zu bewältigen. Im Folgenden werden die wichtigsten Akteure der österreichischen FTI-Politik und ihre Initiativen und Programme beschrieben.

- *Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK)*: Das BMK ist federführend bei der Entwicklung und Umsetzung einer transformativen Innovationspolitik in Österreich. In den vergangenen Jahren wurden vier FTI-Initiativen mit explizit transformativer Ausrichtung definiert: „Mobilitätswende", „Energiewende", „Kreislaufwirtschaft und Produktionstechnologien" und die „Mission Klimaneutrale Stadt".

- *Klima- und Energiefonds*: Der Klima- und Energiefonds unterstützt mehrere FTI-Initiativen, die eine nachhaltige Transformation fördern. Ein Beispiel hierfür war das Programm „Vorzeigeregion Energie", indem Modellregionen unterstützt wurden, die innovative Technologien und Lösungen im Bereich der erneuerbaren Energien und Energieeffizienz entwickeln und umsetzen. Ziel war es, durch Pilotprojekte und Best Practices den Weg für eine nachhaltige Energiezukunft zu ebnen. Die neue „FTI-Initiative für die Transformation der Industrie" im Rahmen der Klima- und Transformationsoffensive des Bundes soll seit 2023 die Industrie bei der Klimaneutralität unterstützen. Hier werden FTI und Investitionsförderungen gebündelt. Des Weiteren entsteht mit dem Programm Leuchttürme der Wärmewende 2024 ein transformatives Programm, welches FTI, Demonstration, Umsetzung, aber auch Qualifikation, Diversität und Chancengleichheit abdeckt.

- *Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF):* Das BMBWF spielt eine zentrale Rolle bei der Koordination der EU-Missionen in Österreich. Diese Missionen sind Teil des 9. EU-Forschungs- und Innovationsprogramms „Horizon Europe" und zielen darauf ab, durch konkrete, lösungsorientierte Maßnahmen große gesellschaftliche Herausforderungen wie die Anpassung an den Klimawandel und den Übergang zu einer nachhaltigen Landwirtschaft und Bodennutzung zu bewältigen. Das BMBWF unterstützt die Umsetzung der Mis-

sionen durch Forschungsförderung, Koordination und Vernetzung der beteiligten Akteure.

- *Agenturen:* Die Umsetzung dieser Politiken wird über die nationalen Forschungsförderungsagenturen in Österreich abgewickelt. Diese sind aws, CDG, FFG, FWF, LBG, und ÖAW[4]. Neben der Zusammenarbeit mit den erwähnten Ressorts und dem Klima- und Energiefonds können dafür auch Mittel der Nationalstiftung „Fonds Zukunft Österreich"[5] abgerufen werden. In kompetitivem Verfahren vergibt die Nationalstiftung Mittel des Fonds auf Basis der Österreichischen FTI-Strategie und den Empfehlungen des Rates für Forschung, Wissenschaft, Innovation und Technologieentwicklung[6] (FORWIT).

Nachfolgend wird dargestellt, wie verschiedene Akteure der österreichischen FTI-Politik Nachhaltigkeit und Transformation in ihren jeweiligen Monitoringansätzen bisher zu erfassen versuchen.

3. STATUS QUO DER IMPLEMENTIERUNG

Im Folgenden werden die bisher verfolgten Ansätze der an der fteval-Arbeitsgruppe vertretenen Institutionen vorgestellt. Die Darstellung erhebt keinen Anspruch auf Vollständigkeit. Sie gibt die in der Arbeitsgruppe diskutierten Ansätze und die Beiträge der Teilnehmenden wieder.

3.1 BMK

SEKTION INNOVATION & TECHNOLOGIE: GESAMTHAFTES MONITORINGSYSTEM

Die Sektion Innovation und Technologie des BMK hat die Aufgabe, technologische Innovationen und Forschung zu fördern, um die Wettbewerbsfähigkeit Österreichs zu stärken und die grüne und digitale Transformation voranzutreiben. Sie entwickelt derzeit (Stand 2024) das Monitoringsystem

4 Austria Wirtschaftsservice (https://www.aws.at/), Christian Doppler Forschungsgesellschaft (https://www.cdg.ac.at/), Öst. Forschungsförderungsgesellschaft (https://www.ffg.at/), FWF Der Wissenschaftsfonds (https://www.fwf.ac.at/de/), Ludwig Boltzmann Gesellschaft (http://www.lbg.ac.at/de), Öst. Akademie der Wissenschaften (http://www.oeaw.ac.at/).

5 https://www.stiftung-fte.at/

6 https://forwit.at/

für die Budgets und Aktivitäten der angewandten Forschung in ihrem Zuständigkeitsbereich weiter. Diese Weiterentwicklung basiert auf einer explizit dafür durchgeführten Vorstudie (Warta et al., 2023).

Das Monitoringsystem bezieht sich auf die drei Wirkungsziele für die Sektion Innovation und Technologie (BVA UG 34):

1. Steigerung der Forschungs-, Technologie- und Innovations-Intensität (FTI-Intensität) des österreichischen Unternehmenssektors

2. Entwicklung von modernen, effizienten, leistungsfähigen und sicheren Technologien und Innovationen zur Bewältigung der großen gesellschaftlichen Zukunftsherausforderungen, wie Klimawandel und Ressourcenknappheit (societal challenges)

3. Steigerung der Beschäftigung im Bereich Forschung, Technologie und Innovation mit besonderem Augenmerk auf Erhöhung des Anteils der Frauen

Für jedes Ziel gibt es Indikatoren entlang einer Interventionslogik von Aktivitäten, Outputs, Outcomes und Impacts. Die Weiterentwicklung des Monitoringsystems sieht vor, dass Indikatoren sich nicht auf einzelne Programme oder Initiativen beziehen sondern auf die Gesamtheit der Budgets für die Finanzierung von Förderungen und die Basisfinanzierung außeruniversitärer Forschungseinrichtungen. Aktivitäten zur nachhaltigen Transformation entstehen nicht nur aus einzelnen top-down Förderprogrammen, sondern auch aus themenoffenen bottom-up Förderungen. Das Monitoringsystem soll einen gesamthaften Blick auf die durch Technologie und Innovation angestoßene nachhaltige Transformation bringen.

Im Folgenden werden beispielhaft die Indikatoren für Ziel 2 zur Bewältigung des Klimawandels und der Ressourcenknappheit (Tabelle 1) dargestellt.

Tabelle 1 Indikatoren für das Ziel 2 Bewältigung des Klimawandels und der Ressourcenknappheit

Level	Indikator	Quelle	Zeitpunkt
Input	Anteil des geplanten UG34 Budgets mit intendierter Klimawirkung (Score 2) oder Klimawirkung als Nebeneffekt (Score 1)	Green Budgeting für UG34[7]	Jährlich, Vorab-Planung der Budgets
Output	Anteil der Nennungen aus FFG-geförderten Projekte, einen Beitrag zu den SDGs 7, 11, 12 und/oder 13 zu leisten	FFG und AWS Datenauswertung	Jährlich, Bekanntgabe durch Förderungsnehmende bei Antragstellung
Output	Systeminnovation (Indikator noch ein Entwicklung)		
Output	Anteil der Projekte bei denen eine Einbindung von Umsetzungspartner:innen, Anwender:innen oder sonstigen Bedarfsträger:innen erfolgt	FFG Datenauswertung	jährlich
Outcome	Anteil der geförderten Projekte, deren Projektergebnisse ein hohes Potenzial für Beiträge in den Umweltdimensionen der EU Taxonomie aufweisen bzw. einen Beitrag realisiert haben	FFG-Wirkungsmonitoring	Jährlich, 4 Jahre nach Projektende
Impact	Patentpublikationen für Umwelttechnologien	Österreichisches Patentamt (ÖPA)	Jährlich
Impact	Anteil des Sektors F&E am umweltbezogenen Produktionswert	Environmental Goods and Service Sector (EGSS; Statistik Austria)	Jährlich

7 Erstellt auf Basis der Green Budgeting Methode des Bundes Green Budgeting Methode des Bundes (bmf.gv.at)

In der Sektion werden derzeit verschiedene Ansätze erprobt und teilweise bereits umgesetzt. Dabei ist zu berücksichtigen, dass diese auf unterschiedlichen Ebenen (Sektion, Programm, Innovation) angesiedelt und an unterschiedliche Kontexte angepasst sind. Bereits umgesetzt ist die Erfassung von Beiträgen zu EU-Nachhaltigkeitszielen, die von Förderungsnehmenden geleistet wurden, anhand der positiv umformulierten und leicht angepassten „Do No Significant Harm" (DNSH)-Kriterien der EU-Taxonomie. Die Erfassung erfolgt im Rahmen des Wirkungsmonitorings der FFG, über die ein Großteil der F&E-Förderungen der Sektion abgewickelt wird. Das Wirkungsmonitoring der FFG basiert auf einer Befragung der Fördernehmer:innen auf Projektbeteiligungsebene, d.h. bei Kooperationsprojekten erhält jede teilnehmende Organisation vier Jahre nach Projektabschluss einen Fragebogen. Dieser zeitliche Aspekt hat den Vorteil, dass Wirkungen, die insbesondere bei F&E-Projekten mit einer zeitlichen Verzögerung von mehreren Jahren auftreten können, zu dem Zeitpunkt erfasst werden, zu dem sie bereits eingetreten sind. Somit ist dieser Ansatz nicht auf eine ex-ante Abschätzung der Wirkungspotenziale angewiesen. Es wird jedoch nur erfasst, *ob* nach Einschätzung der Projektdurchführenden ein Beitrag zu den sechs Aspekte der EU-Taxonomie-Verordnung geleistet wurde, nicht *wie und in welchem Umfang* dieser erfolgt ist.

Die Sektion nutzt darüber hinaus Informationen über die Bewertung der Nachhaltigkeitskriterien (im Zuge des Auswahlprozesses) für das Monitoring. Hier wurde in den letzten Jahren die Art der Bewertung angepasst, weshalb 2024 gemeinsam mit dem BMAW eine Evaluierung des Nachhaltigkeitskriteriums durchgeführt wurde (Seus et al., 2025).

Des Weiteren ist die Sektion dabei, eine Indikatorik auszuarbeiten, die Projekte erfasst, die sich mit **Systeminnovationen** beschäftigen und diese vorantreiben. Dazu baut die Sektion auf zwei unterschiedlichen Ansätzen auf, die für die FTI-Schwerpunkte „Klimaneutrale Stadt" und „Mobilitätswende" entwickelt wurden.

ANSATZ DER „SYSTEM READINESS" AUS DEM FTI-SCHWERPUNKT „KLIMANEUTRALE STADT"

Für den FTI-Schwerpunkt „Klimaneutrale Stadt" wurde ein Modell zur Bewertung des Reifegrads von Systeminnovationen im urbanen Kontext entwickelt, das verschiedene Dimensionen der Reifegradmessung berücksichtigt.[8] Der

8 https://smartcities.at/projects/system-readiness/

Ansatz basiert auf dem etablierten Konzept der Technology Readiness Levels, das in neun Stufen die technologische Reife einer Innovation beschreibt, und erweitert dieses Konzept um weitere Dimensionen. Dazu gehört beispielsweise, inwieweit die regulatorischen und rechtlichen Rahmenbedingungen bereits der Innovation entsprechen oder inwieweit die notwendige Infrastruktur für die Innovation bereitsteht. Das Modell kann in Monitoringprozesse integriert werden, um Fortschritte in der Entwicklung von Systeminnovationen zu verfolgen. Das Monitoringsystem für Systeminnovationen wird 2025 pilotiert und es wird dann entschieden, in welcher Form es weitergeführt wird.

ANSATZ EINES NAVIGATIONSSYSTEMS AUS DEM FTI-SCHWERPUNKT „MOBILITÄTSWENDE“

Im FTI-Schwerpunkt „Mobilitätswende“ wurde der Ansatz eines Navigationssystems entwickelt, um Beiträge der geförderten Projekte zur Weiterentwicklung und Umsetzung von Systeminnovationen zu verfolgen. Dieser Monitoringansatz basiert auf einem Prozessverständnis von Transformation, das auf dem etablierten Konzept der Multilevel-Perspektive (MLP) (Geels, 2002) aufbaut und Wirkungsbeiträge entlang von 17 Handlungsfeldern bzw. Indikatoren berücksichtigt (siehe unten). Eine weitere Besonderheit ist, dass der Fortschritt nicht auf Projektebene, sondern in Bezug auf eine konkrete Systeminnovation erfasst wird, d.h. mehrere Projekte können zu einer Systeminnovation beitragen, die Informationen dazu werden nach Systeminnovationen gebündelt und aufbereitet. Der Monitoringansatz sieht darüber hinaus einen strukturierten Lernprozess zwischen Fördernehmer:innen und den für die Umsetzung der jeweiligen Systeminnovation relevanten Akteuren vor. Der Ansatz wird derzeit pilotiert (Stand 2025).

Das Monitoring basiert vorrangig auf Basis von qualitativen Informationen – daher die Betonung von Handlungsfeldern anstelle von Indikatoren. Die Handlungsfelder, in denen Wirkungen erfasst werden, gliedern sich in drei Ebenen (Kofler & Wieser, 2023), der hier dargestellte Stand basiert auf der Studie und entspricht nicht dem Stand der aktuellen Pilotierung:

KOMPETENZAUFBAU UND LÖSUNGSENTWICKLUNG

- Es liegt eine klare und verständliche Problemdefinition vor.
- Zu einem oder mehreren Erklärungs- und Lösungsansätzen konnte zumindest in einzelnen Fachkreisen ein Konsens gebildet werden.
- Die positiven und negativen Wirkungen des Lösungsbausteins lassen sich gut abschätzen.

- Der Lösungsbaustein konnte sich in einer kontrollierten Testumgebung beweisen.
- Es konnte eine ausreichend große Community gebildet werden.
- Es konnten gemeinsame Visionen und Narrative gebildet werden.

VERBREITUNG UND VERBREITERUNG

- Es konnten Praxiserfahrungen aus dem Umfeld der Anwender:innen gesammelt werden.
- Erwartungen der Anwender:innen wurden in der Entwicklung und Gestaltung des Angebots berücksichtigt.
- Eine Institution (oder mehrere) widmet sich der Verbreitung und/oder Verbreiterung des Lösungsbausteins.
- Es konnte ein Diskurs zu den erforderlichen regulativen Rahmenbedingungen angestoßen werden.
- Es konnten neue Anwendungskontexte erschlossen werden.
- Der Lösungsbaustein erlangte öffentliche Bekanntheit.
- Entscheidungstragenden stehen die erforderlichen Evidenzen zur Verfügung.

UMSETZUNG UND DURCHDRINGUNG

- Die für den Lösungsbaustein erforderlichen Infrastrukturen sind vorhanden.
- Die erforderlichen Gesetzgebungen und Normierungen wurden umgesetzt.
- Der Lösungsbaustein erfährt eine hohe soziale und kulturelle Akzeptanz.
- Es konnte eine nachhaltige Finanzierung für die Skalierung und Bereitstellung des Lösungsbausteins gesichert werden.

3.2 FFG

Als konzeptionelle Grundlage für die Erfassung der Nachhaltigkeitswirkungen verwendet die FFG die Sustainable Development Goals (SDGs) der Vereinten Nationen. Diese bieten 17 breit legitimierte Ziele mit 169 Unterzielen. Die SDGs wurden in das bestehende Programmmonitoring und -controlling integriert.

Die Wirkungen werden auf Projektebene erfasst, indem bei der Antragstellung eine Einschätzung des Zielbeitrags abgefragt wird. Diese kann durch qualitative Angaben ergänzt werden. Die Erfassung erfolgt anhand einer F&E-relevanten Auswahl von Unterzielen. Im Projektendbericht wird der Zielbeitrag erneut abgefragt und ermöglicht damit eine erneute Messung zu einem Zeitpunkt, zu dem die Wirkungen bereits eingetreten sind bzw. die Umsetzung und die erwarteten Wirkungen deutlich weiter fortgeschritten und damit besser abschätzbar sind. Darüber hinaus ermöglicht der Vergleich über die Zeit eine Aussage darüber, wie gut die Projekte ihre Beiträge ex ante einschätzen können.

In der FFG gibt es noch keine expliziten Indikatoren zu Transformation, aber bereits interne Überlegungen und Initiativen dazu. Auf Ebene der transformativen Programme (z.B. „Expedition Zukunft" oder „Impact Innovation") sowie in einzelnen Teams (z.B.: Business Development) wird derzeit ein verstärkter Fokus auf das Monitoring und Evaluierung von Transformation gelegt.

3.3 FWF

Der Wissenschaftsfonds FWF gibt in seinem Werte-Statement[9] an, dass in allen Bereichen seiner Arbeit auf klimafreundliche, ökologische und soziale Nachhaltigkeit geachtet wird. Darüber hinaus setzt sich der FWF für Rahmenbedingungen ein, die es Forschenden ermöglichen, Forschungsprojekte nachhaltig und klimaschonend durchzuführen. Außerdem bestehen neben den themenoffenen Förderungen, in denen Nachhaltigkeit in all ihren Dimensionen erforscht werden kann, auch spezielle thematische Ausschreibungen (zB Zero Emissions Award der alpha+ Stiftung). Der FWF versucht auch im Betrieb der Geschäftsstelle möglichst nachhaltig zu agieren [10]. Mit (inter-)nationalen Förderorganisationen findet ein intensiver Austausch zum Thema statt. Seit Ende 2024 wird vom FWF eine umfassende Nachhaltigkeitsstrategie für die Geschäftsstelle und Förderhandeln - in enger Absprache mit den relevanten Stakeholdern - entwickelt, die sich an den gesetzlich definierten Aufgaben laut Forschungs- und Technologieförderungsgesetz orientiert. Im Monitoring werden insbesondere Diversitätsaspekte hinsichtlich Gender berücksichtigt. Der FWF vergibt Förderungen für Grundlagenforschung bottom-up und themenoffen. Die Ausrichtung auf „gesellschaftliche Transformation" stellt daher keinen thematischen Schwerpunkt dar.

9 https://www.fwf.ac.at/ueber-uns/werte

10 https://www.fwf.ac.at/ueber-uns/aufgaben-und-aktivitaeten/nachhaltigkeit

3.4 UNIVERSITÄTEN

Es gibt diverse universitäre Netzwerke, die sich mit der Messung von Nachhaltigkeit befassen und einschlägige Arbeitsgruppen betreiben, wie die Allianz Nachhaltige Universitäten in Österreich[11] oder das Bündnis Nachhaltige Hochschulen[12]. Grundsätzlich wird in diesen Netzwerken ein „Whole Institution Approach"[13] verfolgt, der versucht verschiedene Nachhaltigkeitsebenen und Aspekte einer Institution unter einem kohärenten gemeinsamen Prinzip zu organisieren. Die Idee ist dabei, dass die gesamte Institution „Universität/ Hochschule" unter diesem Kohärenzprinzip ihre Multiplikator:innenwirkung wahrnimmt.

Ein Forschungsprojekt der BOKU und Universität Graz zu „Wissensbilanzierung und Nachhaltigkeit"[14] beschäftigte sich 2022/23 mit Bewertungssystemen von Nachhaltigkeitsforschung und lehre sowie Ansätzen zur Bewertung und Darstellung des Nachhaltigkeitsimpacts von Forschung und Lehre.

In diesem Projekt wurden auch Überlegungen angestellt, was transformative Forschung ausmacht. Außerdem wurden in diesem Projekt 1) internationale Rankings erhoben, die mitunter Nachhaltigkeitsindikatoren und -impact berücksichtigen, 2) Beispiele zur Nachhaltigkeits-bewertung und 3) Ansätze zur Wirkungsmessung der Nachhaltigkeitswissenschaften.

Basierend auf diesen Zugängen wurden in dem Projekt additionale Kriterien zur Nachhaltigkeitsbewertung von Lehre und Forschung erarbeitet, die auf institutioneller Ebene wirken und eine differenzierte Betrachtung ermöglichen sollen. Ziel dieser Kriterien ist es, (1) die Rahmenbedingungen und Strukturen darzustellen, die die Universitäten in ihrem eigenen Wirkungsraum gestalten und anbieten, um Forschung und Lehre für Nachhaltigkeit zu ermöglichen und zu fördern, sowie (2) eine erste Basis für vergleichende universitäre Berichterstattung/Wissensbilanzierung und Bewertung im Kontext mit Nachhaltigkeitsaktivitäten und -wirkungen bilden. Zusätzlich wurde je ein Reflexionsschema für Nachhaltigkeit Forschung und Lehre entwickelt, mit dessen Hilfe die Forschungs- und Lehrtätigkeiten zur Förderung einer nachhaltigen Entwicklung der Universitäten differenziert in ihrer Nachhaltigkeitstiefe aufgezeigt werden können.

11 https://nachhaltigeuniversitaeten.at/

12 https://www.nachhaltige-hochschulen.at/

13 https://www.unesco.de/node/6100

14 https://boku.ac.at/nachhaltigkeit/nachhaltigkeit-in-der-forschung/wissensbilanzierung-und-nachhaltigkeit-projektergebnisse

3.5 WIRTSCHAFTSAGENTUR WIEN

In der Wirtschaftsagentur wurde im Zuge einer Neustrukturierung des Portfolios ein systematisches Monitoring von Nachhaltigkeitskriterien eingeführt. Die Maßnahmen zur Erfassung der Nachhaltigkeitswirkungen sind vorerst auf die Projektauswahl fokussiert. Grundsätzlich darf in keinem Projekt gegen die „Do no significant harm"-Kriterien (DNSH) verstoßen werden. Hierfür wurde die Definition von DNSH in Anlehnung an die Kriterien der Europäischen Union für den spezifischen Kontext angepasst.

Derzeit gibt es zwei Programme speziell zum Thema Klimaschutz und ein Programm zu den SDGs. Die Antragstellenden müssen die Relevanz ihres Projektes für den Klimaschutz und die SDGs darlegen. Dieses Kriterium wird in der Bewertung als K.O.-Kriterium behandelt. Es wird sowohl auf Unternehmens- als auch auf Projektebene abgefragt, wobei nur die Projektebene in die Bewertung einfließt.

3.6 ZWISCHENFAZIT

Der Überblick zeigt, dass im österreichischen Innovationssystem bereits von mehreren Akteur:innen Ansätze implementiert wurden, um Beiträge zu ökologischer Nachhaltigkeit zu messen. Ansätze, die explizit den Beitrag zu einer nachhaltigen Transformation erfassen, sind noch auf einzelne Förderprogramme beschränkt und befinden sich erst in der Pilotphase. Während die Ziele der FTI-Politik im Hinblick auf das Erreichen transformativer Veränderungen ambitionierter geworden sind, besteht im Monitoring somit noch Aufholbedarf. Diese Lücke zwischen der Umsetzung transformativer Innovationspolitik und den Prozessen, wie das für die Umsetzung notwendige Wissen produziert und angewendet wird, ist international zu beobachten und es wurde noch keine zufriedenstellende Lösung gefunden (Kattel & Mazzucato, 2023). Die in Österreich verwendeten Ansätze konzentrieren sich primär darauf, bei Fördernehmer:innen festzustellen, ob sie einen nachhaltigen Beitrag leisten und um welche Art von Beitrag es sich handelt – dies erfolgt unter Orientierung an den Nachhaltigkeitskonzeptionen der Europäischen Kommission bzw. der Vereinten Nationen. Um den Prozess hin zu einer ökologisch nachhaltigen Wirtschaft und Gesellschaft, die Transformation, zu erfassen, ist dies jedoch nicht ausreichend. Im Folgenden Abschnitt wird auf einige Ansätze aus der Literatur eingegangen, bevor wir mit einem Ausblick für die Weiterentwicklung abschließen.

4. ANSÄTZE ZUR ERFASSUNG VON TRANSFORMATIVEN FTI-WIRKUNGEN AUS DER LITERATUR

In der Literatur wurden in den vergangenen Jahren einige analytische Grundlagen entwickelt, die zur Bestimmung von transformativen FTI-Wirkungen herangezogen werden könnten. Zum Teil wurden die entwickelten Analyseraster bereits in Evaluierungen angewandt. Nachfolgend stellen wir drei Analyseraster kurz vor:

- Der Ansatz der „**technologischen Innovationssysteme**“ (zB Janssen 2019) unterscheidet sieben Funktionen, die Innovationssysteme erfüllen sollten, damit sie funktionieren: Ausrichtung der Forschung, Entwicklung von Wissen, Austausch von Wissen, Unternehmerisches Experimentieren, Mobilisierung von Ressourcen, Legitimierung / Widerstand entgegenwirken, und Marktbildung.

- Der Ansatz der „**transformative Outcomes**“ (zB Boni et al. 2022) wurde aus dem Verständnis abgeleitet, dass sich Transformationen aus der Interaktion unterschiedlicher Ebenen ergeben: Innovationen bilden sich in Nischen heraus (die es zu schützen gilt) und müssen sich gegen dominante Strukturen und Regeln (die es zu destabilisieren gilt) durchsetzen. Für die Messung von „transformative Outcomes“ haben Brodnik und Dinges (2022) eine Liste an möglichen Indikatoren abgeleitet (siehe Tabelle 3 im Anhang).

- Der Ansatz der „**Interventionspunkte**“ (zB Kanger 2020) basiert auf denselben Theorien von soziotechnischen Transitionen, leitet daraus aber sechs Interventionspunkte ab, die einige Hinweise geben, in welchen Bereichen die Politik intervenieren soll, um transformative Wirkungen zu erzielen. Tabelle 1 verschafft einen Überblick der identifizierten Interventionspunkte.

1. STIMULIERUNG VON UNTERSCHIEDLICHEN NISCHENINNOVATIONEN
▪ F&E-Projekte fördern ▪ Experimentierräume schaffen ▪ Informationsaustausch und Wissenstransfer unterstützen
2. SKALIERUNG VON NISCHENINNOVATIONEN
▪ Regulatorische Anreize schaffen ▪ Innovative öffentliche Beschaffung ▪ Finanzierung sichern ▪ Netzwerke und Plattformen zur Diffusion unterstützen
3. DESTABILISIERUNG VON DOMINANTEN REGELN UND STRUKTUREN
▪ Regulatorische Interventionen für Phase Out von Systemen und Praktiken ▪ Anreizstrukturen anpassen ▪ Divestment
4. DIE SOZIALEN UND WIRTSCHAFTLICHEN FOLGEN VON TRANSFORMATIONSPROZESSEN ABFEDERN
▪ Regionale Entwicklungspolitik ▪ Qualifikationsmaßnahmen ▪ Finanzielle Kompensation
5. UNTERSCHIEDLICHE SYSTEME KOORDINIEREN UND ABSTIMMEN
▪ Nationale Strategien und Visionen formulieren ▪ Cross-sektorale Programme etablieren ▪ Plattformen für Koordination und Datenaustausch etablieren
6. DIE GESELLSCHAFTLICHEN RAHMENBEDINGUNGEN ANPASSEN
▪ Internationale Abkommen ▪ Politische Ziele

Tabelle 2 Politische Interventionspunkte zur Forcierung von soziotechnischen TransitionenQuellen: Kanger et al. (2020), Kivimaa et al. (2023)

5. ÜBERLEGUNGEN FÜR EINE INDIKATORIK VON TRANSFORMATIVEN FTI-WIRKUNGEN

In der Arbeitsgruppe haben wir die bestehenden analytischen Grundlagen aus der Literatur diskutiert und festgestellt, dass sie am besten für Bewertung von Maßnahmenbündel bzw. „Policy Mixes" geeignet sind (z.B. im Rahmen einer Evaluierung der nationalen FTI-Politik), aber insbesondere auf der Ebene einzelner Initiativen noch viel Übersetzungsarbeit erforderlich ist. Aus der Perspektive konkreter FTI-Initiativen besteht zum Teil eine große Lücke zwischen den erwartbaren Wirkungen und den Wirkungen, die aus Sicht der bestehenden Analyseraster, die auf der Ebene der Systemtransformation ansetzen, erforderlich wären.

Brodnik & Dinges (2022) machten bereits einen Schritt in Richtung einer besseren Anschlussfähigkeit der Analyseraster für das Monitoring und die Evaluierung von FTI-Initiativen, indem sie Verknüpfungen zwischen den Kategorien der Analyseraster und konkreten Indikatoren aus anderen Bereichen wie den SDGs und Responsible Research & Innovation (RRI) herstellten. Vielversprechender erscheint uns jedoch, Indikatoren auf Grundlage des Konzepts transformativen Wandels anstelle von vordefinierten Theorien von Transformationsprozessen zu entwickeln. Je nach Kontext werden unterschiedliche Transformationsprozesse angestoßen und speziell auf der Ebene der FTI-Initiativen bestehen große Unterschiede in den jeweiligen Interventionslogiken.

Transformative Veränderungen können hingegen auf unterschiedlichen Ebenen auftreten und unterschiedliche Formen (tief, breit, schnell) annehmen (siehe Abschnitt 2). Indikatoren von transformativen Wirkungen können somit flexibler auf diverse Kontexte angewandt und neben der Tiefe, die in den bestehenden Analyserastern am stärksten betont wird, auch andere Dimensionen wie die Breite und Geschwindigkeit berücksichtigen. Sowohl die Breite als auch die Geschwindigkeit sind in vielen Bereichen von entscheidender Bedeutung. Diese Ausweitung des Begriffs der Transformation entspricht der Idee, dass Monitoringkonzepte so gestaltet sein sollten, dass sie nicht bestimmte Wege einschränken, sondern eher lösungsoffen bleiben und Debatten anregen (Rafols & Stirling, 2020).

6. AUSBLICK

Mit diesem Bericht möchte die Arbeitsgruppe einen ersten Diskussionsanstoß zur Weiterentwicklung und Konsolidierung transformativer Indikatorik im österreichischen System erzielen. Im Forschungsprojekt „TIME: Transformatives Innovationsmonitoring- und Evaluierungsframework“ (10/2024 – 06/2025) werden die Ergebnisse der Arbeitsgruppe gemeinsam mit interessierten Stakeholdern aufgegriffen und in Richtung eines umfassenden Indikatorensets weiterentwickelt. Das Forschungsprojekt wird vom Klima- und Energiefonds gefördert. Als weitere Entwicklungsschritte sind folgende Maßnahmen vorgesehen: Zunächst wird der konzeptionelle Rahmen des Indikatorensets weiterentwickelt, bestehende Indikatoren aus wissenschaftlicher und grauer Literatur werden gescreent und Qualitätskriterien definiert. Zusätzlich werden im Rahmen eines Workshops auf der REvaluation-Konferenz Indikatoren mit internationalen Expert:innen gesammelt und diskutiert. Anschließend wählt das Projektteam geeignete Indikatoren aus und entwickelt diese weiter. In einem darauffolgenden Workshop werden diese mit Verantwortliche für FTI-politische Maßnahmen aus dem BMK und dem KLIEN hinsichtlich ihrer Praxistauglichkeit und Generalisierbarkeit überprüft. In einem zweiten Workshop validieren Evaluator:innen abschließend die Anwendbarkeit, bevor das finale Indikatorenset fertiggestellt und ein Leitfaden für dessen Nutzung erstellt wird.

REFERENZEN

Andersen, A. D., Geels, F. W., Coenen, L., Hanson, J., Korsnes, M., Linnerud, K., Makitie, T., Nordholm, A., Ryghaug, M., Skjolsvold, T., Steen, M., & Wiebe, K. (2023). Faster, broader, and deeper! Suggested directions for research on net-zero transitions. *Oxford Open Energy*, *2*, oiad007. https://doi.org/10.1093/ooenergy/oiad007

Boni, A., Molas-Gallart, J., Velasco, D., Fernández-Méndez, P., Terrazas, P., & Schot, J. (2022). *Evaluating transformative innovation policy. Insights from two experimental policies*. https://www.semanticscholar.org/paper/Evaluating-transformative-innovation-policy.-from-Boni-Molas-Gallart/f9e50932861c31be658356fcd88d8dce44dc21c9

Brodnik, C., & Dinges, M. (2022). The Evolution of Research and Innovation Policy Paradigms and Associated Evaluation and Indicator Frameworks. In G. Vélez-Cuartas & O. Y. Romero-Goyeneche (Hrsg.), *Transformative Metrics: Contributions to the Studies for Monitoring and Evaluating How Science, Technology, and Innovation Can Address Social and Environmental Challenges*. Fondo Editorial FCSH. https://doi.org/10.17533/978-628-7592-15-5_1

Brunnhuber, Nadine, Feldbaumer, Marcus, ranzer-Sudra, Karin, Steffl, Thomas, Warmuth, Hannes, & Windsperger, Bernhard (2022). *Tech4Green. Disruptive Technologien für eine nachhaltige Produktion* (S. 214). Österreichische Gesellschaft für Umwelt und Technik (ÖGUT), Institut für Industrielle Ökologie, IIÖ, scenario editor. https://produktionderzukunft.at/resources/pdf/Tech4Green_Ergebnisbericht.pdf

Dinges, M., Kerlen, C., Kaufmann, P., Wang, A., Toepel, K., Kofler, J., Meyer, S., & Wieser, H. (2022). Theories of change for transformation-oriented R&I polices: the case of the 7th Energy Research Programme in Germany. *fteval Journal for Research and Technology Policy Evaluation*, *53*, 57–68. https://doi.org/10.22163/fteval.2022.542

Ecker, B., Brunner, P., Dudenbostel, T., Grabner, S. M., Hartmann, A. E., Heckenberg, D., Heckl, E., Kasneci, G., Kofler, J., Philipp, S., Régent, V., Sardadvar, S., Schuch, K., Sturn, D., Wagner, V., Warta, K., & Wieser, H. (2023). *Forschungs- und Technologiebericht 2023*. Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF), Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK), Bundesministerium für Arbeit und Wirtschaft (BMAW).

European Commission. Directorate General for Research and Innovation (2020). *Framing R&I for transformative change towards sustainable development in the European Union.* Publications Office. https://data.europa.eu/doi/10.2777/503127

Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study. *Research Policy, 31*(8-9), 1257–1274. https://doi.org/10.1016/S0048-7333(02)00062-8

Ghosh, B., Kivimaa, P., Ramirez, M., Schot, J., & Torrens, J. (2021). Transformative outcomes: assessing and reorienting experimentation with transformative innovation policy. *Science and Public Policy*, *48*(5), 739–756. https://doi.org/10.1093/scipol/scab045

Haddad, C. R., Nakić, V., Bergek, A., & Hellsmark, H. (2022). Transformative innovation policy: A systematic review. *Environmental Innovation and Societal Transitions*, *43*, 14–40. https://doi.org/10.1016/j.eist.2022.03.002

Janger, Jürgen, & Slickers, Tim (2023). Wissensproduktion und Wissensverwertung in Österreich im internationalen Vergleich. *WIFO*, *10/2023*, 699–714. https://www.wifo.ac.at/wp-content/uploads/upload-8011/mb_2023_10_04_wissensproduktion_.pdf

Janssen, M. J. (2019). What bangs for your buck? Assessing the design and impact of Dutch transformative policy. *Technological Forecasting and Social Change*, *138*, 78–94. https://doi.org/10.1016/j.techfore.2018.08.011

Kanger, L., Sovacool, B. K., & Noorkõiv, M. (2020). Six policy intervention points for sustainability transitions: A conceptual framework and a systematic literature review. *Research Policy*, *49*(7), 104072. https://doi.org/10.1016/j.respol.2020.104072

Kofler, J. & Wieser, H. (2023). Navigating transitions: a 'reflexive navigation system' for monitoring and learning in mission-oriented innovation policies. Präsentation auf der EU SPRI Konferenz, 14-16 Juni, Brighton.

Lazarevic, D., Salo, H., & Kautto, P. (2022). Circular economy policies and their transformative outcomes: The transformative intent of Finland's strategic policy programme. *Journal of Cleaner Production*, *379*, 134892. https://doi.org/10.1016/j.jclepro.2022.134892

Penna, C. C. R., Romero Goyeneche, O. Y., & Matti, C. (2023). Exploring indicators for monitoring sociotechnical system transitions through portfolio networks. *Science and Public Policy*, *50*(4), 719–741. https://doi.org/10.1093/scipol/scad015

Rafols, I., & Stirling, A. (2020). *Designing indicators for opening up evaluation. Insights from research assessment*. https://doi.org/10.31235/osf.io/h2fxp

Ravn, T., Nielsen, M. W., & Mejlgaard, N. (2015). *Metrics and indicators of Responsible Research and Innovation_Progress report D3.2_Monitoring the Evolution and Benefits of Responsible Research and Innovation (MoRRI)*. https://doi.org/10.13140/RG.2.2.12773.40165

Santos, A. M., & Coad, A. (2023). Monitoring and evaluation of transformative innovation policy: Suggestions for Improvement. *Socio-Economic Planning Sciences*, *90*, 101714. https://doi.org/10.1016/j.seps.2023.101714

Seus, S., Geyer, A., Langkau, S., & Hansmeier, H. (2025). *Studie zur Evaluierung der FFG Nachhaltigkeitskriterien: Endbericht.* Fraunhofer ISI & inspire research. https://repository.fteval.at/id/eprint/737/

Schot, J., & Steinmueller, W. E. (2018). Three frames for innovation policy: R&D, systems of innovation and transformative change. *Research Policy*, *47*(9), 1554–1567. https://doi.org/10.1016/j.respol.2018.08.011

Warta, K., Kaufmann, P., Kofler, J., & Wieser, H. (2023). *Konzeptvorschlag für ein wirkungsorientiertes FTI-Monitoring*. Technopolis Group | Austria. https://doi.org/10.22163/fteval.2023.621

Weber, K. M., & Rohracher, H. (2012). Legitimizing research, technology and innovation policies for transformative change: Combining insights from innovation systems and multi-level perspective in a comprehensive 'failures' framework. *Research Policy*, *41*(6), 1037–1047. https://doi.org/10.1016/j.respol.2011.10.015

Weber, M., & Kubeczko, K. (2023). Kapitel 13. Innovationssystem und -politik. In C. Görg, V. Madner, A. Muhar, A. Novy, A. Posch, K. W. Steininger, & E. Aigner (Hrsg.), *APCC Special Report: Strukturen für ein klimafreundliches Leben* (S. 403–412). Springer. https://doi.org/10.1007/978-3-662-66497-1_17

AUTOR:INNEN

JAKOB KOFLER
KMU Forschung Austria
Gußhausstraße 8, 1040 Wien
Email: j.kofler@kmuforschung.ac.at
ORCID: 0000-0002-6059-9989

ISABELLA WAGNER
Österreichische Plattform für Forschungs- und Technologiepolitikevaluierung
Linke Wienzeile 246, c/o ZSI, 1150 Wien
Email: wagner@fteval.at
ORCID: 0000-0002-2772-6771

HARALD WIESER
KMU Forschung Austria
Gußhausstraße 8, 1040 Wien
Email: h.wieser@kmuforschung.ac.at
ORCID: 0000-0002-5066-4214

MITGLIEDER DER ARBEITSGRUPPE:
Alfried Braumann (Wirtschaftsagentur Wien), Lisa Bohunovsky und Ines Hinterleitner (BOKU) Harald Hochreiter und Marlene Lang (FFG), Lisa Koch (BMIMI), Jakob Kofler (KMU Forschung Austria), Ina Matt (FWF), Sabine Mayer (FFG), Susanne Meyer (BMIMI), Sascha Sardadvar (WPZ-Research), Klaus Schuch (ZSI), Harald Wieser (KMU Forschung Austria), Isabella Wagner (Plattform fteval)

ANHANG

Tabelle 3 Vorschläge für Indikator-Kategorien für „Transformative Outcomes“ („Table 1. Indicator Categories for Transformative Outcomes“ aus Brodnik & Dinges, 2022)

Process	Transformative Outcome	Indicator Categories	Example of Indicators for Some of the Categories [unit]
Building and nurturing niches	Shielding: protecting new and more sustainable practices from external influences and helping them grow.	▪ R&D budget and subsidies for niche innovation. ▪ Fiscal support for niches (e.g., taxation). ▪ Public/Collective purchasing and procurement of niche innovations. ▪ Voluntary agreements with niche actors. ▪ Supportive regulation for niches. ▪ Experiments aimed at changing framework conditions (e.g., regulatory sandboxes).	▪ Business and government expenditures in R&D (euros). ▪ Subsidies and tax credits (euros). ▪ Procurement contracts (euros). ▪ The stringency of the regulation (qual.). ▪ The number of experiments (count). ▪ The number of agreements (count).
	Learning providing regular opportunities for discussing experiences, obstacles, and needs related to a new practice as well as challenging related values and assumptions that people might have.	Analytical descriptive knowledge about the current system and associated sustainability problems: ▪ Different types of system maps (e.g., policy landscape, project portfolios, etc.). ▪ Scientific publications (including conference papers or discussion papers). ▪ Grey literature. ▪ Datasets and databases of environmental or problem-related data. Normative knowledge about sustainability goals and desirable system states: ▪ Visions. ▪ Problem framings. ▪ Scenarios (qualitative, quantitative, or mixed). ▪ Capacities to develop effective sustainability interventions: ▪ Stakeholder track record in deploying sustainability initiatives. ▪ Existence of spin-offs/follow-up projects.	Network maps (qual./visual). ▪ The number of publications per year. ▪ Types of framing technologies, publication venues. ▪ The number of projects (count). ▪ The number of routines and strategies (count). ▪ The number of coalitions (count).

		Practical skills and knowledge that incorporate sustainability in routine actions: ▪ Evidence that sustainability has been anchored in routines beyond intervention. ▪ Evidence that sustainability has been anchored in strategies beyond intervention ▪ Interpersonal skills for developing coalitions and alliances. ▪ New networks and coalitions that are maintained beyond the project/intervention.	
	Networking: protecting and progressing new practises by gaining the interest of more people and creating connections between them.	Champions / Individuals: ▪ The number of champions. ▪ Type of champions (individual, organisational, etc.). ▪ Position/embeddedness of champions in a network. Actors' networks: ▪ Degree of formalisation of networks (from loosely ▪ connected individuals to formal networks). ▪ Autonomy and resources of networks. ▪ Heterogeneity of network. ▪ Inclusiveness of network. Intermediaries: ▪ Presence and number of intermediaries. ▪ Changes in the type of intermediary (individual, organisation, etc.). ▪ Roles of intermediaries (niche-, regime-, process-, systemic intermediary). ▪ Position/embeddedness of intermediaries in a network. ▪ System aggregation level at which intermediaries operate (local, regional, national, international).	▪ The number of champions (count.). ▪ Network metrics (indexes/quant.). ▪ Number [count.] and type of intermediaries (qual.).

	Navigating expectations: navigating and converging expectations of different actors the legitimacy of new practises is developed, and their potential explored.	Narratives: ▪ Presence of a new narrative or signs of an emerging narrative in different outlets (e.g., media, scientific, political, industry publications). ▪ Framing of solutions to sustainability issues widens (from a narrow problem-solution to a wider meaning). ▪ Changes to advocating narrative/counter-narrative. ▪ Coalitions around particular framings and narratives. Visions: ▪ Directionality of existing visions/new visions. ▪ Increase in reach/buy-in of visions. ▪ 'Quality' of vision (e.g., co-developed, widely shared, transformational aspirations, etc.)	▪ The number of newspaper articles (count.). ▪ The number of parliamentary discussions (count.). ▪ Opinion polls (qual.-quant.). ▪ Semantic metrics for narratives (qual.-quant.) ▪ The number of different coalitions (qual.).
Process	**Transformative Outcome**	**Indicator Categories**	**Example of Indicators for Some of the Categories [unit]**
Expanding and mainstreaming niches	Upscaling: conducting deliberate action to get more users involved in new and more sustainable practises.	Scaling: ▪ The number of stakeholders/stakeholder groups that engage with new practise. ▪ Changes in the number of practises adopted in a specific area/sector and at a certain level (local, national, transnational). ▪ Changes in the speed of adoption of practise in a specific area/sector and at a certain level (local, national, transnational). Scalable potential: ▪ Cost for an additional application of practise. ▪ Valorisation of practise by stakeholders	▪ Demand size for a niche (euros). ▪ Cost estimate for niche practise adoption (euros).

	Replicating: transferring the new and more sustainable practises to another location.	Replicating: ▪ Practise is applied in different settings/circumstances. Replication potential: ▪ Independence of practise from cultural (e.g., user preferences) or structural (e.g., governance arrangements) particularities.	Number of different geographical markets for niches (count.).
	Circulating: exchanging knowledge, ideas, and resources between multiple related alternative practises.	▪ Knowledge and experience collection and synthesis. ▪ External knowledge and experience accessibility. ▪ Knowledge and experience sharing among stakeholders.	▪ The number of accesses to a website (count.). ▪ The number of attendees in a workshop (count.). ▪ The number of recipients of newsletters (count.).
	Institutionalising: turning new and more sustainable practises into more permanent and more widely available ones.	▪ Guidelines for best practises are developed. ▪ New standards are developed. ▪ Existing standards are adapted. ▪ New laws are developed. ▪ Existing laws are adapted. ▪ Practise features in emerging/dominant discourse.	The number of guidelines, standards, laws, etc. (count.).
Process	**Transformative Outcome**	**Indicator Categories**	**Example of Indicators for Some of the Categories [unit]**
Opening-up and unlocking regimes	De-aligning and destabilising regimes: disrupting and weakening dominant practises. This can be done by changing one of the dominant dimensions for example through the introduction of new policies.	Top-down: ▪ Phase-out policies. ▪ Bans on entrenched practises. ▪ Removal of subsidies of entrenched practises. ▪ Targeted financial incentives for alternative practises. Bottom-up: ▪ Public demonstrations, rallies, or marches. ▪ Boycotts. ▪ Petitions. ▪ Media campaigns. ▪ Public debates. ▪ Emerging discourses and metaphors.	▪ Number and stringency of policies (count. and qualitative). ▪ The number of grassroot events (count.). ▪ Opinion polls (qual.-quant.).

	Unlearning and deep learning in regimes: dominant actors question their assumptions and change their view on the potential of new and more sustainable practises and the ability of the dominant practise to respond to threats and opportunities, such as climate change and digitalisation.	▪ Evidence that new problem framings are being adopted by regime actors, e.g., in regime publications and advertisement campaigns. ▪ Evidence of changes in the direction of routine (r&i) search processes (i.e., moving into previously unexplored areas of knowledge); ▪ Existence of re-skilling, retrofitting, and repurposing programmes.	▪ Types of media and marketing ▪ campaigns (qual.). ▪ The number of patents (beyond ▪ regimes core area) (count.). ▪ The number of programmes ▪ (count.).
	Strengthening regime-niche interactions: frequency and quality of interactions between empowered actors from the niche and the regime on a non-competitive basis.	▪ Establishment of partnerships and collaborations between regimes and niches. ▪ Corporate venture capital initiatives for niche innovations. ▪ Merges and acquisitions (m&a) between the regime and niche actors (e.g., firms).	▪ The number of partnerships (count.) size of venture capital funds (euros). ▪ Number and size of m&a (count. / euros).
	Changing perceptions of landscape pressures: dominant actors to reach the point of view that immediate action is warranted, and new emerging more sustainable narratives need to be promoted.	▪ New regime discourses and narratives (framing) around a landscape trend (e.g., climate change). ▪ Announcement of new strategies, products, or services that seek to address a pressure or benefit from an opportunity at the landscape level.	▪ Semantic metrics (qual.-quant.). ▪ The number of announcements ▪ (count.).

fteval JOURNAL
for Research and
Technology Policy Evaluation

November 2024, Vol. 55, pp. e7, 1-20
DOI: 10.22163/fteval.2024.662

THEORETICAL REFLECTIONS ON THE ROLE OF SOCIAL INNOVATION IN CHALLENGE AND GOAL-ORIENTED R&I AGENDA[1]

KLAUS SCHUCH
DOI: 10.22163/FTEVAL.2024.662

1. INTRODUCTION

To overcome or at least contain major societal challenges such as the climate crisis, biodiversity crisis, resource crisis, democracy and trust crisis, as well as the energy and mobility transition (to recall just a few of the most important ones), the idea that technology alone will not be enough has become mainstream. There is growing consensus that complementary, sometimes alternative and transformative social innovations are needed (Weber et al., 2024; Edler 2022, Köhler et al. 2019; Wunder et al. 2019; Schot and Steinmueller 2018; Schartinger et al. 2017). The importance of social innovations for supporting transformation of our economy and society towards sustainability issues is also featured by the socio-technic and sustainability-transition view on system-transforming processes (Havas et al., 2023).

This paper endeavours to show why social innovations should also be included in challenge- and goal-oriented research agendas and how the latter need to

1 This paper builds on a policy brief of the author on "Die Rolle der Geistes-, Sozial- und Kulturwissenschaften und die Bedeutung von sozialer Innovation bei der Umsetzung der Missionsziele (The role of the humanities, social and cultural sciences and the importance of social innovation in achieving the Mission objectives)" published at the Austrian ERA Portal in March 2024. It takes up some ideas mentioned there and takes them further. (https://era.gv.at/public/documents/5030/Policy_Brief_GSK_soziale_Innovation_Missionen.pdf)

be conceptually expanded to ensure that this inclusion is successful. To derive and substantiate this approach, a definition of social innovation is contrasted with the traditional techno-economic understanding of innovation in Section 2 of this paper. An attempt is made to show that the analytical differences between the two definitory approaches are not insurmountable when it comes to innovation. Innovation is understood as the introduction of something new on the market or a social sphere that has value and impact and which is based on the translation of an idea or a concept into tangible outputs or practices. If such an innovation is successfully utilised (by businesses, consumers, civil society organisation, citizens etc.), it could lead to outcomes that improve efficiency and effectiveness, increase the capacity to act, and meets needs better than with existing solutions or even address so far unmet needs.

The third Section aims to explain that what is commonly understood or can be understood by social innovation is closely linked to the understanding of the term "social" in "social innovation" and that this understanding can by no means be taken for granted. The section attempts to show that a certain understanding of social innovation, which is linked to practice theory, provides a bridge for integrating social innovation into challenge and goal-oriented R&I programmes. This understanding emphasises the intentional change of social practices to achieve better solutions to certain social problems (*"social in a broad sense" – see Section 3 and 4)* than is possible with conventional practices. We will argue that such an approach to social innovation makes it applicable for addressing the subject of major societal challenges because it does not remain stuck in a socio-political understanding.

In the fourth Section, we retrace to anchor social innovation into Geel's multi-level perspective framework on socio-technical transitions (Geels, 2020 and 2002) by referring to the exceptional work by Weber et al. (2024). In this way, we show how to position social innovation within the theoretical framework of socio-technical system changes and thus make it more compatible with innovation policy.

The fifth and final Section concludes on how challenge- and goal-oriented R&I programmes should be changed to make them receptive to social innovation research. If technology alone will probably not be enough, as postulated in the opening sentence of this paper, than R&I programmes need to be composed in a holistic manner to achieve the societal goals addressed by transformative research agendas.

2. THE RELATIONSHIP BETWEEN SOCIAL AND TECHNICAL-ECONOMIC INNOVATION

In this Section, we aim to provide a solution to the question of what social innovation is and how it relates to technical and economic innovation. Although there are various definitions in scholarly discourse, coupled with conceptual ambiguity (van der Have and Rubalcaba, 2016; Cuna and Benneworth, 2013; Pol and Ville 2009), an understanding seems to be gaining ground, at least in German-speaking academic schools of thought, that defines social innovation *"as a new combination or figuration of practices in social fields of action that is initiated by certain actors or constellations of actors with the aim of overcoming needs and problems better than is possible with existing practices. An innovation is therefore social insofar as it changes social action and is socially accepted and disseminated in society"* (Howaldt and Schwarz, 2016, p. 6). The advantage of this somewhat unwieldy definition lies in its analytical and epistemological orientation power when compared with central definitory elements of the established techno-economic definition of innovation (see Schuch and Šalamon, 2021).

1. The object of social innovation in this definition is a changed social practice. However, as this school of thought is strongly rooted in practice theory, practice is not limited to the social practice of individual actions and actors but also considers structural and institutional changes, which in turn have an impact on collective practices. The "classical" object of techno-economic innovation is a new technology (technique) in form of a new product or production technology. The OECD subsumes product or production technology under the term technological innovations, but generally speaks of business innovation, which also includes organisational, managerial, marketing and business model innovations (OECD and Eurostat, 2018).

2. Novelty is an important characteristic of both social and techno-economic innovation, whereby new does not necessarily mean "radically new", but can also mean new in a specific context or for a specific actor. Especially when it comes to changes in social practices, it is actually less about the one 'heroic' first-mover and more about scaling and mainstreaming the new social practice in order to generate impact on a large scale, which possibly even contributes to social change.

3. While an economic purpose is decisive for technical-economic innovations, a social purpose is decisive for social innovations (Weber et al. 2024) to be able to overcome social problems better than with traditional practices. However, "social" is by no means limited to the socio-political context, but can include other social spheres such as a society's handling of natural resources or the maintenance of a basic trust in civil security.

4. What both forms of innovation have in common is an intentionality and solution orientation initiated by actors. While the actors for technical-economic innovations are more or less clearly defined (mainly business and R&D organisations), social innovators can be found in all segments of society. If actors, intention and solution orientation are no longer clearly identifiable, fuzzy and overlapping, for example due to multiple subsequent uses and adaptations, and become – so to say – "generalised" or "mainstream", then one will speak of the phenomenon of social change and no longer of a concrete social innovation.

5. Both forms of innovation require practical application. In the case of business innovations, this is economic practice; in the case of social innovation, it is social practice. If there is no application, then we speak of an invention or idea. With social innovations, however, sometimes the 'prophetic' idea, expressed through *"the definition and articulation of new social facts"* (McGowan et al. 2021; p. 21), can lead to certain implementations being initiated.[2]

2 This is where the discourse on whether advocacy is already a social innovation comes in. In the context of the definition of SI used here, advocacy per se is not a social innovation because the immediate solution orientation (expressed by a new product, a new service or a changed practice etc.) can be missing. Advocacy usually draws attention to a problem (e.g. the effects of climate change). Sometimes, however, advocacy also propagates solutions (e.g. 100 km/h on motorways), although these are often not implemented (and therefore not used).

Table 1 summarises the similarities and differences between techno-economic innovation and social innovation.

Tab. 1: Similarities and differences between techno-economic innovation and social innovation in a nutshell

Category	Techno-economic innovation	Social innovation
Object of innovation	a changed technology/product	a changed social practice
Novelty	new in a specific context or for an actor (mostly companies, but can be other actors too)	new in a specific context or for an actor (no ex-ante restriction to a specific actor)
Purpose	economic purpose to cope with problems/needs/wants better than with conventional technologies/products (solution-orientation!)	social purpose to cope with problems better than with conventional practices (solution-orientation!)
Intentionality	prompted by certain actors	prompted by certain actors
Application	in practice	in practice

Source: adapted from Schuch and Šalamon, 2021.

It is also important to recognise that social innovations are not a dichotomous contrast to techno-economic innovations. According to Weber et al. (2024) innovations can be purely technological, non-technological, or a combination of the two, that is, socio-technical. Social innovations can be triggered by technology and technology can also be actively utilised for social innovations. An illustrative example for this combination are digitally supported social innovations (Bria et al, 2016), which were triggered for instance by the use of smartphones for the organisation of work, leisure, care-taking, education, etc. Thus, the differentiation between technological and social innovation turns out to be rather artificial, as most innovations rely to some degree both on new technologies and non-technological changes (Weber et al., 2024). The history of ideas on social innovation is rich in this respect (see Godin and Schubert, 2021; and Godin 2015 for a detailed reflection). However, Edwards-Schachter and Wallace (2017) found that the discourse about social innovation became increasingly neglected during 1975 to 1995, when the focus on techno-economic innovation became dominant both in academic and policy circles making social innova-

tion a peripheral issue, although social innovation in relation to socio-technical innovation did not entirely disappear (Schuch, 2023; Streicher and Schuch, 2022).

There are also social innovations in which technology plays no significant role. The early days of veganism, still considered "esoteric" by many in Europe only ten to twenty years ago, could be an example of this, although the vegan industry has made enormous technological progress with significant investments in the meantime. In 2023, the global vegan food market was valued at more than USD 37 billion, growing at a compound annual growth rate of 13.5% during the forecast period to 2032[3]. However, veganism started primarily as a social movement, which quickly responded with no- or low-tech supported social innovations such as vegan cooking classes to meet its needs and, over the course of a few decades, revolutionised a dominant social practice and a long-standing business model from a niche[4]. As an increasingly "mainstream or generalised" social practice, it reduces animal suffering, which for many was and is a central motive for an intended effect, and it reduces greenhouse gas emissions from animals (which is probably more of an indirect and unintended effect).

3. THE MEANING OF "SOCIAL" IN SOCIAL INNOVATION?

The understanding of the term "social innovation" presented in the section above postulates the importance of social innovation for innovation and change, but does not yet clarify its function or its substance in relation to transformation or social change, which is why we attempt an approximation in this section by reflecting on the different meanings of "social" in the use of the term "social innovation", without plunging into the depths of sociological subtleties and paradigm battles.[5]

3 https://www.fortunebusinessinsights.com/de/markt-f-r-vegane-lebensmittel-106421; accessed on 11 June 2024.

4 Very recently, Austria established a job profile and apprenticeship for vegan and vegetarian cookery as of 2025, which signals a transition from a niche to a mainstreamed practice with new institutional characteristics.

5 We are explicitly looking forward to feedback on the simplifications made here, so that we can either develop the concept further or discontinue it.

Table 2 attempts to differentiate the meaning of "social" in the term "social innovation" with reference to current discourses, whereby a simple scheme is deliberately presented to make the different attributions of meaning to the term "social" more clearly comprehensible.

Tab. 2: Dimensions of the term "social" in social innovation discourses

Dimensions of the term "social" in SI discourses

Social = 'Help for the socially disadvantaged' = socio-political meaning	**Social** = 'doing good' = normative meaning	**Social** = ,changing social relations and status' = socio-structural perspective	**Social** = ,changing social practices' = behavioural perspective

Convergence in SI practice, although the socio-political approach is more narrow negating ecological concerns in their substance as an object of SI, which, however, occupy an important space in the 'doing good' approach.

Sometimes combined in SI practice.

Potential Key aspects for transformative social innovation

The majority of articles in relevant social innovation journals such as "International Journal of Social Entrepreneurship and Innovation", the "Stanford Social Innovation Review" or "The European Public & Social Innovation Review" as well as contributions at major international social innovation research conferences (such as ISIRC) pursue a socio-politically connoted understanding of the term social innovation. "Social" here gets the notion of "charitable, helpful, merciful"; and relates to measures to improve the economic and/or social situation of disadvantaged social groups in particular. In this context, social innovations often refer to the improvement of the living conditions of socially disadvantaged groups, which in turn can be defined very diversely, depending on the object of investigation and research question (e.g. minorities of all kinds). This rather dominant and, of course, very important approach for society is outlined in the first pillar in Table 2.

This understanding of social innovation is naturally subject to some criticism. This includes, for example, the limited target group focus on socially disadvantaged groups only and the restricted socio-politically focus, whereby the distinction between "social needs" and "social wants" often remains unclear, as does the question of who actually has the power of definition and agenda

setting[6]. And, although the meaningfulness of socio-politically motivated social innovation should of course not be disputed per se, the question also arises as to what extent social innovation in this understanding is actually just a substitute for the failures of (regular) social policy and policies. Social innovations that compensate for social grievances to prevent bad things from happening and take the pressure off an overheated cooking pot (metaphorically speaking) also tend to stabilise the established socio-political regime more than to reform or even transform it.

This second understanding of social innovation is linked to the first approach, which also understands the term "social" in social innovation very normatively in the sense of "doing something good". This second understanding of "social", however, transcends the thematic socio-political agenda of the first approach by taking a broader view on "social challenges or problems" and, for example, also addresses social problems in dealing with nature. This perspective opens up the inclusion of several different target groups, not just marginalised groups. In essence, the difference between this second and the first understanding of the term is that "doing good" cannot be understood exclusively in socio-political terms, but also includes other aspects, such as social innovations that strive for other improvements in the real world, e.g. ecological improvements, or common good issues such as in the field of mobility or energy provision. Sometimes both approaches can be found in combination (e.g. social innovations dealing with energy poverty or climate resilience of marginalised groups).

A fundamental point of criticism of the "doing good" approach in general is the normative dilemma of deciding what is actually "good" and who has the power to define it in view of different interests and perceptions, not least among target groups that come from and have been socialised in certain coercive milieus. In this context, there is also the accusation of "teleological naivety". This refers to the often naïve self-image of social innovators that they assume to only do good, but sometimes neglect or even negate power relations, social structures, conflicts of interest and rebound effects triggered by their interventions.

The third conceptual understanding of social innovation outlined in Table 2 focuses on the purpose of changing social interactions, relationships and the social status of those affected. It is a socio-structural perspective that is often characterised by keywords such as "empowerment", "ownership" and "participation" (*have a say*). It refers to the capacity to act. Here, too, the distinction

6 This also points to the question of the role of a certain marginalised group as passive object or as an active subject.

between this and the other conceptual understandings is by no means clear-cut, as there are also cases in practice where a connection between the "doing good" approach and the socio-structural approach is aspired. The frequently quoted EU terminology of social innovation used by the European Social Fund in the sense of *"social in its ends and in its means"* indicates this connection, in which the process design of social innovation is understood to be just as central as the actual goal (e.g. better care; easier access etc.) (European Commission, 2021). When designing a socially innovative process, particular emphasis is placed on aspects such as inclusion, democratic participation and empowerment, i.e. not just better provision of help for a "care case". However, cases of substantial participation (beyond pure contribution) tend to be rare or abstract because there are intersectional restrictions that often cannot be substantially overcome by individual social innovations. The socio-structural perspective can also be a bridge between the first two, very normative understandings of the term "social" and the fourth understanding discussed here, because the changes in social relations can affect both the micro level and the macro level. An example of the changes in social relationships at the micro level are street newspapers, which have turned (sometimes begging) homeless people into vendors and thus improved their social status. Examples on the macro level were the abolition of slavery or serfdom, or nowadays the recognition of a gender beyond the male-female dichotomy. The latter examples not only changed social relationships but also institutional structures, which in turn changed (or are in the process of changing) collective and individual practices.

The fourth prototypical understanding of the term "social" in social innovation shown in Table 2 focuses on changing social practice(s) and can be regarded as behaviourally oriented. Practice means more than an activity; practice also refers to networks of practice, institutions that promote or constrain certain practices, cognitive frameworks that give meaning to certain practices, etc. This understanding, which also forms the basis of the definition of social innovation presented in Section 2, is of interest because it refers to practices in a wide variety of spheres of social life and is not limited to socio-political or philanthropic areas.

Social innovations, understood as intentional social practice changes, prompted by certain actors, with the aim of overcoming needs and problems better than is possible with existing practices, can take place in a wide variety of areas such as the way

- how we consume (e.g. veganism; shared economy; packaging-free shopping);

- how we live and organise ourselves (e.g. digitalisation of social relationships; new working modes such as working from home; shared flats for elderly people);
- how we produce (e.g. circular economy; bioeconomy; community-based agriculture; energy communities);
- how we create and reproduce meaning (e.g. global/local; digital/non-digital; both formal and informal);
- and the way we interact with our environment (e.g. changing mobility behaviour to help tackling the climate crisis; urban gardening as an approach to prevent biodegradation; socio-ecological use and partnership models to reduce waste of resources).

Central points of criticism to this understanding of social innovation refer to the difficult but in fact promising distinction between social innovation and social change (and thus between agency and behavioural change), or the difficult operationalisation of this understanding in real-world applications. Moreover, if the focus on practice changes lacks the dimension of a "social purpose" as a demarcation to the economic profit-making purpose, then also changed social practices in business such as the introduction of Taylorism at the beginning of the 20th century or other organisational changes in business operations could be regarded as social innovation too. However, such an unbounded understanding of social practices in the context of social innovations would undermine the identity core of a social innovation, namely its social purpose. Such a delimited understanding, which ignores the difference between "economic" and "social" purpose also endangers the definitional and epistemological content and power of social innovation to provide hermeneutic orientation (Schuch and Šalamon, 2021).

The practice-oriented focus of social innovation appears to be particularly interesting for the transformation discourse, not only because it opens up a broad space for social innovation action, but also because this approach can be grounded in practice theory (Howaldt and Schwarz, 2017), whereby elements of "behaviour" and "behavioural change" as well as "intentionality" and "agency" can both be considered and analysed. In practice theory, social practices can be scrutinised on different levels between action and structure, which, in addition to the structural context that expresses cultural knowledge, norms or values, also takes the subjective perspective into account. In real life, social practices consist of more or less organised bundles of activities. They are understood as social because they are shared by people and help us to unders-

tand the social world around us. Social practices, however, are subject to social change and social innovations can lead to new social practices and to social change as discussed in the next Section.

4. INTEGRATING SOCIAL INNOVATION IN CHALLENGE- AND GOAL-ORIENTED R&I AGENDAS

While we are aware of the differences between challenge- and goal-oriented R&I, we believe that these are at an abstract level not so much of primary importance in relation to the logic presented. Both, challenge- and goal-oriented innovation agendas often, but not always, point to the same need, namely, to support the transformation of our societies (including the way how we produce and consume) in the direction of ecological, economic and social sustainability. That this claim may be unattainable or even could have dangerous implications for a free social order is not something we want to go into here. However, the necessity of institutional and practice change is considered as a crucial factor in achieving broader or more narrow goals stipulated by several challenge and goal-oriented R&I agendas including the EU Missions. Havas et al. (2023) define goal-oriented transformative change as a closely interrelated set of fundamental changes at the level of socio-technical or an entire socio-economic system, with changes simultaneously affecting its underlying technologies, business models, cognitive frames, institutions, networks as well as practices, initiated by various types of actors to achieve a major overarching goal.

The emphasis of new social practices also makes it easier to understand social innovations as part of new social practices that can be either system-stabilising (e.g. to take the pressure out of an acute problem situation) or system-transforming, which manifests itself in institutional pressure, the breaking of path dependencies and – as a result – in institutional change (Haxeltine et al., 2017). Avelino et al. (2019; p. 196) understand transformative social innovation *"as social innovation that challenges, alters or replaces dominant institutions in the social context"* and as *"irreversible, persistent adjustment in societal value, outlooks and behaviours".* Bundles of new social practices can lead to a reconfiguration of dominant practices and institutions and can contribute via such processes to social change. Social change, in turn, is understood in terms of co-evolutionary changes in structures, policies, institutions, practices and behaviours. These are often supported by technology. The recursive relationship between social

innovation and the institutionalisation of new social practices is similar to evolutionary and complex systems thinking about technological innovation (Weber et al., 2024).

The understanding of transformative social innovation opens a bridge to the multi-level perspective framework on transition (Geels, 2002), which is a process-oriented dynamic heuristic of a system in flux, that argues that sustainability transitions cannot be achieved as merely bottom-up or top-down processes but require an interplay of micro-level niche developments and meso-level regime changes. In this heuristic, the niche level provides space for experimentation, including experimentation for social innovations. It is the locus where it is possible to deviate from an existing practice and obtain knowledge about user behaviour, collaboration needs, new practices and rules (Smith and Raven, 2012). The regime level, on the other hand, provides stability through its dominant institutions, infrastructures, and a common understanding of problems and possible solutions. Weber et al. (2024) mention in their Multi-Level Perspective on Social Innovation, which builds on the heuristics of Geels, that the regime level is governed or influenced by the following three social forces: institutions, social network and cognitive frames.

Moving novel socially innovative solutions that meet social needs or wants better than traditional ones from the niche to the regime level is not just a matter of diffusion and upscaling, which also results in a lot of trial and error but requires the embedding of these new solutions in institutional environments that may themselves equally require major changes (Weber et al., 2024). However, incumbent regime actors might be tempted to hinder social innovation initiatives if their status is challenged by them. Such a conflict can either lead to the failure and decline of the social innovation or to its continuous improvement and persistence. In the latter case of success, Weber et al. (2024, p. 58) explain that *"a process of reframing and adapting institutional frameworks, policies, and practices begins that opens up institutional space in the regime for the new Social innovation to inhabit and leads to the circulation and anchoring of the new knowledge associated with it."* They further argue that regime changes caused by successful transformative social innovation initiatives (or bundles of social innovations) are usually characterised by low speed and broad scope, which means that they could need decades to unfold (low speed) but can lead to a broad scope of changes too. Examples for such a system transition are the energy or the mobility transition (see Weber et al., 2024).

The third and most abstract (highest) level in this multi-level perspective on transition heuristics is that of the socio-technical landscape (Geels 2002),

which provides even stronger structuration and guidance through cultural norms, cognitive beliefs, or existential external pressures such as climate crisis, but is beyond the influence of single actors and considered as a slowly changing context both for (social) innovation and socio-technical transformation. In this evolutionary thinking, the concept of social exnovation becomes important too, which is understood as the purposive and explicit phasing out or modification of unsustainable social practices and institutions.

Although social innovation has been addressed in the multi-level perspective framework on transition only in recent years, it seems that this framework is sufficiently open and flexible to accommodate social innovations and how they evolve from niches to widespread solutions and practices as shown by Weber et al. (2024). Moreover, the sustainability transition literature, which has been stressing environmental aspects also opened up to social issues recently (e.g. in relation to just transitions).

Edler et al. (2022) have analysed the policy approaches for promoting social innovation in Germany with regard to mission-oriented innovation policy (MOIP) and transformative innovation policy (TIP). In transformative innovation policy (TIP), socially desirable transformation dynamics are taken up and strengthened (Diercks et al. 2019; Schot and Steinmueller 2018; Steward 2012). This involves identifying social innovations in their niche, creating scope for bottom-up dynamics and improving the conditions for further development through scaling-up or system-wide adaptation (scaling-out). In contrast, the MOIP (Larrue 2021; Mazzucato 2018) defines very specific goals (missions) in the political process, which are then to be achieved through the mobilisation of innovation. In terms of social innovation, the MOIP goes beyond the emergence of social innovations and deliberately poses the question what new social practices could be initiated to achieve mission goals. Experimental spaces that are set up and supported in a transdisciplinary scientific manner are essential for this. However, Edler et al. (2022) also state that neither the initiation nor mobilisation of social innovations play a major role in EU Missions. This is also due to the conceptual deficit to consider social innovation as an adequate lever. Without appropriate political support, e.g. through funding measures, the potential of social innovation, expressed in initiated new practices and the willingness of citizens to change their behaviour, will not be exploited.

5. CONCLUSIONS

To understand and utilise the potential of social innovation for challenge- and goal-oriented transformative R&I approaches, it is firstly important to move social innovation away from the purely socio-political perspective that has dominated social innovation research to date. The success of transformation towards sustainability will not depend so much on convincing marginalised groups to adopt sustainable social practices, but on whether it is possible to convince the majority (i.e. usually the broad middle class) to do so. The latter also has significantly greater power to bring about change than marginalised groups.

Looking at the SDGs also makes clear that many of the goals postulated there, some of which also provide guidance for the European Framework Programme for Research and Innovation (Horizon Europe) (Mayer and Schuch, 2019), cannot be achieved through technology (alone) (e.g. *"Gender Equality"; "No Poverty"; "Good Health and Wellbeing"; "Reduced Inequalities"*; to name just a few; see also Wilsdon et al., 2023, on the strong contribution of SSH[7] in SDG-related research). In the first five defined EU Missions for instance *("Cancer", "Cities", "Climate", "Soil", "Waters"),* which are part of Horizon Europe, the technological framing is still dominant, but by no means exclusive[8]. Even if the EU Missions provide a rather technological and, in some cases, even limited transformative perspective, social innovations, understood as new social practices in the definition presented in Section 3, could find their place in them[9]. R&I policy, however, needs to recognise that to achieve the objectives of challenge- or goal-oriented R&I programmes fairly and integratively, it needs to enable

7 SSH= Social Sciences and Humanities

8 The Horizon Europe Regulation (2020) defines "missions" as follows: ‚mission' means a portfolio of excellence-based and impact-driven R&I actions across disciplines and sectors intended to:
- achieve, within a set timeframe, a measurable goal that could not be achieved through individual actions,
- have impact on society and policy-making through science and technology, and
- be relevant for a significant part of the European population and a wide range of European citizens.

9 Examples of this could be new public health processes for the early detection of cancer (especially in men); patient-inclusive monitoring of diseases/health; changes in eating habits and lifestyles; new business models in the circular economy through the involvement of citizens and consumers; climate-conscious and resilient urban planning, development and usage models; alternatives to dominant mobility behaviour; self-organised energy communities and much more. The social and economic sciences can also contribute to a better understanding of social developments in cross-mission issues too (such as topics relating to fairness, social resistance, the risk of poverty through missions, environmental economics or legal research on property/common property).

appropriate strategic and operational measures and instruments for the synergetic integration of social innovation and social innovation research into such programmes. Furthermore, to strengthen the socio-ecological resilience, EU Mission agendas should be set up inclusively from the outset, i.e. with the involvement of citizens and civil society organisations, which can be important incubators of social innovations too.

It is also important to move away from the heroic and equally naïve idea that social innovations only create something good. We need to accept that every social innovation inevitably inheres interests and power relations and creates different degrees of *"empowerment"* and *"(dis)empowerment"* (Avelino et al., 2019). Veganism makes few friends with milk producers and pig farmers. However, the more farmers keep their animals in unworthy conditions, i.e. do not change their traditional capitalistic farming practices, the more veganism will grow and take away the guild's dominant business model.

Transformation tasks will also meet with resistance, not only from dominant companies, capitalistic practices and regulations, but also from ordinary people who fear that their freedoms will somehow be restricted (e.g. the freedom to drive a combustion car into the city centre). To understand the resistance to transformational endeavours, to shed light on it and to develop social exnovation from unsustainable practices, and – probably most important - to translate this topic into a rational social discourse, comprehensive SSH research is needed. On the other hand, to take such topics on board, SSH has to be prepared to break out of its academic, too often self-referential comfort zone. Thus, the SSH are also called upon to redefine their relationship and function to society and the grand challenges (see König, Nowotny and Schuch, 2019).

Finally, since the contribution of (transformative) social innovation to challenge- or goal-oriented R&I programmes is little known, which is partly due to the predominately marginal involvement and participation of SSH in general and social innovation research in particular in such programmes, inspiring national and international examples should be collected, analysed and prepared for the scientific discourse as well as for an integrative, mutually open dialogue between science and society through various formats.

Learning experiences from social innovation pilots within (future) challenge- or goal-oriented R&I programmes should then also be incorporated into the R&I policy discourse. Innovation policy should politically support and promote the diversity of social innovation in such a way that it can unfold its transformation potential constructively (see Eder et al., 2022). Challenge- and goal-oriented research funding would do well to grant space and support to social innovation

undertakings and to find the right balance between intervention and emergence. In the future, evaluations of challenge- and goal-oriented R&I programmes should then also determine whether these programmes provide sufficient space for social innovations, whether they support them adequately, and whether the programme regulations are conducive for social innovation and social innovation research or not. In this sense, design evaluations will be needed first, followed by process evaluations, which will have to be supplemented by impact evaluations in the future.

6. LITERATURE

Avelino, F., Wittmayer, J. M., Pel, B., Weaver, P., Dumitru, A., Haxeltine, A., Kemp, R., Jorgensen, M. S., Bauler, T., Ruijsink, S. and T. O'Riordan (2019). Transformative social innovation and (dis)empowerment. *Technological Forecasting & Social Change* 145 (2019), pp. 195-206.

Bria, F., Gascó, M., Baeck, P., Halpin, H. Almirall, E. and Kresin, F. (2015). Growing a Digital Social Innovation Ecosystem for Europe. Final Report of DSI project. Edited by the European Commission, https://data.europa.eu/doi/10.2759/448169.

Cunha, J. and P. Benneworth (2013). Universities' contributions to social innovation: Towards a theoretical framework, paper presented at EURA Conference 2013, 3–6 July, Enschede.

Diercks, G., Larsen, H. and Steward, F. (2019). Transformative innovation policy: Addressing variety in an emerging policy paradigm, Research Policy, 48/4, pp. 880–894.

Edler, J., Ostertag, K. und Schuler, J. (2022). Die Rolle sozialer Innovationen im Rahmen staatlicher missionsorientierter und transformativer Innovationspolitik. In Howaldt, J., Kreibich, M., Streicher, J. und Thiem, C. (Hg): *Zukunft gestalten mit sozialen Innovationen. Neue Herausforderungen für Politik, Gesellschaft und Wirtschaft.* Campus-Verlag; S. 39-56.

Edwards-Schachter, M. and Wallace, M. L. (2017). 'Shaken, but not stirred': Sixty years of defining social innovation. Technological Forecasting and Social Change 119(4):64-79; DOI:10.1016/j.techfore.2017.03.012.

European Commission (2021). Regulation (EU) 2021/1057 of the European Parliament and of the Council of June 24, 2021 establishing the European Social

Fund Plus (ESF+) and repealing Regulation (EU) No 1296/2013. https://eur-lex.europa.eu/eli/reg/2021/1057/oj.

Geels, F. W. (2020). Transformative innovation and socio-technical transitions to address grand challenges. Luxembourg: Publication Office of the Europe Union (R&I Paper Series Policy Brief, 2020/02).

Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study. *Research Policy* 31 (8-9), pp. 1257-1274.

Godin, B. and C. Schubert (2021). Research on the history of innovation: from the spiritual to the social. In J. Howaldt, Kaletka, C. and Schröder, A (eds*): A Research Agenda for Social Innovation*, Edward Elgar Publishing, pp. 21-38.

Godin, B. (2015). *Innovation Contested. The Idea of Innovation Over the Centuries*. London: Routledge.

Havas, A., Schartinger, D. and Weber, M. (2023). Innovation studies, social innovation, and sustainability research: From mutual ignorance towards an integrative perspective? Environmental Innovation and Societal Transitions, Vol. 48, 2023, DOI:10.1016/j.eist.2023.100754.

Haxeltine, A., Pel, B., Wittmayer, J., Dimitru, A., Kemp, R., Avelino, F. (2017). Building a middle-range theory of Transformative Social Innovation; theoretical pitfalls and methodological responses. *European Public & Social Innovation Review* 2 (1), pp 59-77.

Howaldt, J. and Schwarz, M. (2017). Social Innovation and Human Development—How the Capabilities Approach and Social Innovation Theory Mutually Support Each Other. Journal of Human Development and Capabilities 18(2):1-18. DOI:10.1080/19452829.2016.1251401.

Howaldt, J. and Schwarz, M. (2016). Social innovation and its relationship to social change. Verifying existing social theories in reference to social innovation and its relationship to social change, D1.3, a deliverable of the project Social Innovation: Driving Force of Social Change (SI-DRIVE), Dortmund: Sozialforschungsstelle.

Köhler, J., Geels, F. W., Kern, F., Markard, J., Onsongo, E., Wieczorek, A., Alkemade, F., Avelino, F., Bergek, A., Boons, F., Fünfschilling, L., Hess, D., Holtz, G., Hyysalo, S., Jenkins, K., Kivimaa, P., Martiskainen, M., McMeekin, A., Mühlemeier, M. S., Nykvist, B., Pel, B., Raven, R., Rohracher, H., Sandén, B., Schot, J., Sovacool, B.,

Turnheim, B., Welch, D. and Wells, P. (2019). An agenda for sustainability transitions research: State of the art and future directions Environmental Innovation and Societal Transitions, Volume 31, 2019, pp. 1-32. https://doi.org/10.1016/j.eist.2019.01.004.

König, T., H. Nowotny and K. Schuch (2019). Impact Re-Loaded, *fteval Journal for Research and Technology Policy Evaluation,* 48/July 2019, pp. 8-9.

Larrue, P. (2021). The design and implementation of mission-oriented innovation policies. A new systemic policy approach to address societal challenges. OECD Science, Technology and Industry Papers No. 100, February 2021, Paris: OECD.

Mayer K, Schuch K. (2019). Fostering the Sustainable Development Goals in Horizon Europe. doi: 10.22163/fteval.2019.416.

McGowan, K., Westley, F., Moore, M. L., Alexiuk, E., Antadze, N., Geobey S. and O. Tjornbo (2021). The importance of systems thinking and transformation for social innovation research: the evolution of an approach to social innovation. In Howaldt, J., Kaletka, C. and Schröder, A. (eds): *A research agenda for social innovation.* Edward Elgar Publishing, pp. 59-79.

Mazzucato, M. (2018). *Mission-Oriented Research & Innovation in the European Union. A problem-solving approach to fuel innovation-led growth*, Luxembourg: Publications Office of the European Union.

OECD and Eurostat (2018). Oslo Manual 2018. Available online at https://www.oecd.org/en/publications/oslo-manual-2018_9789264304604-en.html

Pol, E and S. Ville (2009). Social Innovation: Buzz Word or Enduring Term?'*Journal of Socio-Economics*, 38(6), pp. 878-885, DOI: 10.1016/j.socec.2009.02.011.

Schartinger, D., Wepner, B., Andersson, T., Abbas, Q., Asenova, D., Damianova, Z., Dimona, A., Ariton, V., Hannum, C., Eker, S., Schräder, A. and Zirngibl, M. (2017). SI-DRIVE: Social Innovation: Driving Force of Social Change, https://www.si-drive.eu/wp-content/uploads/2017/03/SI-DRIVE-Deliverable-D6_3-Environment-final.pdf.

Schot, J. and W. E. Steinmueller (2018). Three frames for innovation policy: R&D, systems of innovation and transformative change, *Research Policy,* 47(9), pp. 1554-1567.

Schuch, K. (2024). Die Rolle der Geistes-, Sozial- und Kulturwissenschaften und die Bedeutung von sozialer Innovation bei der Umsetzung der Missionsziele.

Austrian ERA Portal; https://era.gv.at/public/documents/5030/Policy_Brief_GSK_soziale_Innovation_Missionen.pdf.

Schuch, K. (2023) Social innovation and social sciences. In Howaldt, J. and Kaletka, C. (eds). Encyclopaedia of Social Innovation. Edward Elgar Publishing, pp. 376-381.

Schuch, K. and Šalamon, N. (2021). Social innovation and social sciences: reflections on a difficult relationship. In Howaldt, J., Kaletka, C. and Schröder, A. (eds): A research agenda for social innovation. Edward Elgar Publishing, pp. 245-262.

Smith, A. and Raven, R. (2012). What is protective space? Reconsidering niches in transitions to sustainability. *Research Policy,* 2012, vol. 41, issue 6, pp. 1025-1036.

Steward, F. (2012). Transformative innovation policy to meet the challenge of climate change: sociotechnical networks aligned with consumption and end-use as new transition arenas for a low-carbon society or green economy, *Technology Analysis & Strategic Management*, 24/4, pp. 331–343.

Streicher, J. und Schuch, K. (2022). Soziale Innovationen in Österreich: Vision gesucht. In Howaldt, J., Kreibich, M., Streicher, J. und Thiem, C. (Hg): *Zukunft gestalten mit sozialen Innovationen. Neue Herausforderungen für Politik, Gesellschaft und Wirtschaft.* Campus-Verlag; S. 71-86.

Van der Have, R. P. and L. Rubalcaba (2016). Social innovation research: An emerging area of innovation studies? *Research Policy*, 45/9, pp. 1923-1935.

Weber, M., Giesecke, S., Havas, A., Schartinger, D., Albiez, A., Horak, S., Blind, K., Bodenheimer, M., Daimer, S., Shi, L., Stadler, M. and Schmitz, D. (2024). Social innovation: (accompanying) instrument for addressing societal challenges?, Studien zum deutschen Innovationssystem, No. 10-2024, Expertenkommission Forschung und Innovation (EFI), Berlin.

Wilsdon, J., Weber-Boer, K., Wastl, J. and Bridges, E. (2023) *Reimagining the Recipe for Research and Innovation: the secret sauce of social science*, London: Sage / Academy of Social Sciences.

Wunder, S., Albrecht, S., Porsch, L. und Öhler, L. (2019). Kriterien zur Bewertung des Transformationspotentials von Nachhaltigkeitsinitiativen, https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2019-03-26_texte_33-2019_transformationspotenzial.pdf.

AUTHOR

KLAUS SCHUCH
ZSI – Centre for Social Innovation
Linke Wienzeile 246, 1150 Wien
E: schuch@zsi.at
ORCID: 0000-0002-7603-4384

fteval JOURNAL
for Research and
Technology Policy Evaluation

December 2024, Vol. 55, pp. e8, 1-18
DOI: 10.22163/fteval.2024.664

AI GOVERNANCE IN AUSTRIA AND AUSTRALIA: LEARNING FROM THE ANTIPODEANS

PETER BIEGELBAUER
DOI: 10.22163/FTEVAL.2024.664

ABSTRACT

In all OECD countries, discussions abate around the question of how to best regulate artificial intelligence (AI) to exploit the economic and social potential of this technology, while minimising risks. The article asks the question of how to govern AI and what role the state might take, with respect to questions of regulation and utilization of AI in its own ranks. Case studies for comparison are Austria and Australia, two democratic high-income countries with differences in geographic location, political system, and economic structure. While both countries aim for a balance of innovation, ethics, and regulation in their AI strategy, the chosen elements for obtaining this balance vary remarkably regarding structures, processes, and tools.

Keywords: AI, AI governance, Austria, Australia, policy learning

The economic prospects of artificial intelligence (AI) are substantial, with projections indicating its impact on productivity, GDP growth, and employment patterns worldwide. By 2030, the AI market is expected to drive a significant portion of the global economy, potentially contributing $19.9 trillion in economic value, and steering around 3.5% of global GDP (IDC, 2024). In the U.S., generative AI alone could contribute between $2.6 and $4.4 trillion annually to GDP, driven by productivity gains and labour automation (McKinsey, 2023). Even when the figures may be debatable (Acemoglu, 2024), one thing is quite clear: the economic impact of AI is sizable. However, the economic impact of AI adoption is not uniform globally. According to the International Monetary Fund (2024), economies with strong digital infrastructure and skilled labor,

such as those in advanced markets, are better positioned to capture AI's benefits.

Therefore, and despite frequent warnings that the "AI hype" is coming to an end anytime soon (e.g., by the inventor of the much-hyped tech hype-cycle: Gartner, 2017), the global race for AI experts, technologies, corporations, and infrastructure is in full effect. All OECD countries feature AI policies, governance structures, and increasingly, also regulations (Galindo et al, 2021; Huw et al, 2023; Walter, 2024). Moreover, policymakers all over the world are thinking about how to foster their own AI eco-systems, how to support these, how their efforts compare to others - and how they could evaluate their accomplishments in this matter (OECD, 2024a). Therefore, AI policy papers, structures, and measures will become subject to evaluations before long.

When thinking about this worldwide competition, it is also important to keep in mind that AI, similar to other technologies, comes with risks attached (Biegelbauer et al, 2022; 2024). AI models learn from vast datasets that often inadvertently embed societal biases, leading to discriminatory outcomes in fields like hiring, lending, and law enforcement (Zou and Schiebinger, 2018). This risk is particularly concerning for algorithms used in decision-making for sensitive areas, such as criminal justice or healthcare - the prime reason for the EU to classify AI systems focusing on these societal areas as high-risk applications in its AI Act (EU, 2024).

From an innovation perspective, interesting questions to answer are: how can we govern artificial intelligence and exploit the economic and social potential of this technology while minimising its risks? What might be the role of the state, especially with respect to questions of regulation and utilization of AI in its own ranks?

In Europe, we are used to compare ourselves with each other, most often looking at the innovation leader countries in the North of the continent (EC, 2024). Yet, perhaps we could also learn from looking a bit farther, and once we have encouraged ourselves to do so, why not look as far as you can - to Australia.

I have a specific angle to this question, since I stayed in Australia from March to June 2024 as a distinguished visiting scientist with the national science agency Commonwealth Science and Industry Research Organization (CSIRO) upon invitation of Justine Lacey, the leader of CSIRO's Future Science Platform. This provided me with the opportunity to talk to AI and AI ethics experts in Brisbane, Sydney, and Canberra. My discussion partners came primarily from CSIRO, where I could work together with people from locations over all

of Australia, but also the University of Queensland, the Technical University of Queensland, as well as the Department of Industry, Science and Resources (DISR), the Digital Transformation Agency (DTA), and the Ministry for Agriculture, Fisheries and Forestry (MAFF), where I had a presentation on policy coordination - as it turns out, a topic quite relevant to AI governance.

Moreover, I am leading the AIT AI Ethics Lab of the AIT Austrian Institute of Technology, where a team of social and computer scientists develops trustworthy AI and works on ethical and legal questions pertaining to AI. Amongst other things, we have been tasked by the Federal Ministry for Arts, Culture, Civil Service and Sport to create practical guidelines for AI in the public administration. These should literally guide civil servants in their daily work and provide a framework for reflection of digitalization and AI, technology assessment, ethics, and training standards (BMKÖS, 2023, 11). The guidelines were co-developed with the BMKÖS Department for Strategic Performance Management and Public Service Innovation, with a second and enlarged version coming out a year after the first one (BMKÖS 2024b).

A comparison between Austria and Australia can provide some interesting results regarding AI policies, governance structures, and regulations. To obtain a better understanding of differences and similarities regarding these, a comparison should start with a short look at economic and political framework conditions influencing AI governance, before turning to the topic of AI governance itself.

AUSTRALIA AND AUSTRIA: ECONOMY AND POLITICS

The two countries have a number of things in common, for example in terms of GDP per capita, but also their role in global digitalization with both countries arguably not being home to a leading AI eco-system. Yet, there are also some differences to be reckoned with. Here, size matters - in terms of population, country, economy (i.e., key factors of production), as do differences in the government structures.

Australia now has 26 million inhabitants, with a GDP of USD 71,930 per capita in 2022 (OECD, 2024b), living on 7.7 million km² and an economy highly dependent on mining - close to 10% of GDP - and agriculture. Exports of minerals, particularly coal, iron ore, and gold, are important. However, the Australian government is increasingly prioritizing diversification through services and

knowledge-based industries, including finance, education, and healthcare. The Australian economy's robust trade relationships with Asian markets, particularly China, have also influenced its economic landscape (Australian Bureau of Statistics, 2024).

Austria, on the other hand, has 9 million inhabitants, with a GDP of USD 71,014 per capita in 2022 (OECD, 2024b), living on close to 84,000 km² and a more industrialized and diversified economy, rooted in manufacturing, engineering, and financial services. Similar to Germany, industrial production in mid-tech sectors is important, specifically machinery, automotive parts, and chemical industries. Situated at the heart of Europe (to cite a slogan of the Austrian tourism industry, an important sector for both countries), Austria benefits from EU trade policies and a strong relationship with neighboring Germany (Statistics Austria, 2024).

The Australian political system is half-jokingly called "Wash-minster" system, pointing at similarities to both the US ("Washington") and the UK ("Westminster"), with a strong second chamber in parliament, an active state level (both similar to the US) and a prime minister heading government (similar to the UK). Its policies are shaped by a two-party system dominated by the Liberal-National Coalition and the Australian Labor Party, fostering stability. Australia's foreign policy has traditionally aligned with Western allies, although its proximity to Asia has led to greater engagement with the Asia-Pacific region.

Austria is a federal republic featuring an important state level and a two-chamber parliament, one of which (the Bundesrat supposed to represent regional interests) is weak, however. Austria's political landscape is multiparty, with a coalition government structure, fostering a collaborative but - with the two formerly largest parties, Social Democrats and Conservatives, shrinking - increasingly complex governance model. Austria's EU membership places it firmly within the framework of European policies, particularly in areas such as trade, innovation, and environment. Austria also maintains a longstanding policy of neutrality, which influences its foreign policy and contributes to a unique position within the EU.

Differences between the two countries also pertain to the Anglo-Saxon pragmatic way of approaching (case) law, business, and government in Australia ready to address problems by creating surprising combinations of factors - the "Wash-minster system" itself being an example, on the one hand. On the other hand, in Austria the Germanic principle-led approach is dominant, to (Roman) law, business, and government, ready to sometimes go for the abstract axiomatic argument.

A short side remark on AI ethics: epitomizing the differences between Germanic and Anglo-Saxon thinking is the difference between Immanuel Kant and Jeremy Bentham. Both are important 18th century European Enlightenment thinkers with a sizable impact on philosophy and specifically the applied philosophy subfield AI ethics, featuring different approaches to moral philosophy, but also politics and law. Kant's categorial imperative ("act only according to that maxim whereby you can at the same time will that it should become a universal law") and Bentham' fundamental axiom ("it is the greatest happiness of the greatest number that is the measure of right and wrong") leads to, amongst other things, also different angles on how to think about the ethical imperatives regarding AI (Stahl, 2021), with ramifications for, e.g, policy imperatives.

Regarding AI, both countries are trying to mitigate the impact of US and Chinese hyperscalers by working together to strengthen their AI ecosystems. The US corporations Alphabet, Amazon, Meta, Apple, and Microsoft mostly are seen as cooperation partners, e.g., in initiatives such as the Australian National AI Centre NAIC. At the same time the danger of being left out has driven both Australian and Austrian governments to engage in funding programmes such as AI Mission Austria (with the three largest national research funders joining forces), and the invitation of inbound investments, such as the Australian federal government, together with the Queensland government, investing AUD 1 bn into quantum computing firm PsiQuantum in 2024.

Table 1 offers an overview over some of the key characteristics of the economic and political framework conditions for AI governance Australia and Austria.

Table 1. Economic and Political Framework Conditions Australia and Austria

	Australia	Austria
Country size	7,700,000 km2	84,000 km2
GDP per capita	USD 71,930 (2022)	USD 71,014 (2022)
Structure political system	"Wash-minster" with medium strong executive and strong state-level representation	Continental European with strong executive and weak state-level representation
Supranational context	Anglo-Saxon states cooperation, Commonwealth of Nations	European Union
Reaction to global AI competition	Cooperate with US AI leaders	Cooperate within EU structures

Approach to ethics	Anglo-Saxon European Enlightenment based (pragmatist)	German European Enlightenment based (principle based)

AI GOVERNANCE: ACTORS AND POLICIES

After having described some of the key elements of the AI governance framework conditions, we will take a comparative look on actors and policies of AI governance, especially ministries, agencies, policy support, and coordination structures of the field.

AUSTRALIA

Regarding AI governance the main government players in Australia are the Department of Industry, Science and Resources (DISR) and the Digital Transformation Agency (DTA). Other ministries have smaller portfolios, such as the Department of Infrastructure (catchy slogan: Connecting Australians!), which oversees the Australian Communications and Media Authority ACMA, dealing with disinformation, among other things.

DISR is therefore also responsible for the key documents of Australian AI governance, e.g., the national AI Ethics Framework (2019), 'Applying the AI Ethics Principles' (2024b), and 'Safe and responsible AI in Australia' (2023), as a result of which a series of participatory events took place (partly interministerial, transdisciplinary, encompassing various economic sectors, online and offline), with the preliminary result of an 'Australian Government's interim response' (2024a) and the prospect of various further governance activities, including an AI Act.

Interestingly, the AI Ethics Framework rests on a set of principles mostly identical with their EU counterparts, specifically the European Commission's High-Level Experts Group Ethics Guidelines (2019). However, subsequent documents have a more industry-friendly approach, which in comparison to their European counterparts have less of an emphasis on human rights.

In Australia, agencies play a key role: in addition to DTA and ACMA, e.g., the Office of the Australian Information Commission (OAIC) on data protection and freedom of information, the eSafety Commissioner on online safety, and the Privacy Commissioner.

In fact, one of the key differences between Australian and Austrian (or most other European countries) policies regarding US Tech corporations is the

Australian position to stand their ground in cases of rights infringement. This seems to be supported by the independence of regulators to, amongst things, interpret basic rights such as privacy. Independent regulators interpret their role more political than their EU counterparts, which often act in a more technocratic fashion. An example is the conflict of the Australian eSafety Commissioner Julie Inman Grant with X-owner Elon Musk's understanding of freedom of speech, which frequently entails ignoring (or even actively engaging in) hate speech and disinformation (The Guardian, 2024).

When it comes to AI, DTA has the central operational role among the agencies and is responsible for working with AI within the administration, including a corresponding DTA/DISR task force. The DTA is responsible for the Australian Government Architecture AGA, a public database containing all government documents relevant to digitisation, as well as for the Digital Review of Digital Competency, which covers all government agencies (although this has only been carried out once so far and has led to considerable controversy) and the Investment Oversight Framework for all government ICT investments. Under the auspices of the DTA Australian Public Service (APS) Trials were held for the first time in 2024, with more than 50 agencies (including the national science agency, the Commonwealth Science and Industry Research Organization CSIRO) trialling the MS Copilot software, resulting in an impressive evaluation report from which there is a lot to learn (DTA, 2024).

At state level, Queensland and New South Wales are innovating and early on have developed frameworks, guidelines, and education programmes for AI, with other states following more reluctantly. Data and Digital Minister's Meetings between federal and state governments have been introduced more recently.

Science policy support is coordinated, with the science agency CSIRO as a hub. The current chair of the National Science and Technology Council formerly was CSIRO Chief Scientist. CSIRO organises also the National AI Center NAIC and the Responsible AI Network RAIN, both of which are funded by DISR, and was also commissioned to create the national AI Ethics Framework with its 8 principles (comparable to the EC HLEG principles, see EC, 2019). Side note: in summer 2024 NAIC has become integrated into DISR, with CSIRO organizational support ongoing.

In 2024, for several months, a temporary Artificial Intelligence (AI) Expert Group was established by DISR to advise government on technical, and regulatory expertise on AI.

The key traumatic incident for Australian AI governance is the Robodebt scandal involving a governmental automated debt recovery system targeting welfare recipients for alleged overpayments (cp. Royal Commission into the Robodebt Scheme 2023). The algorithmic decision-support system was initiated in 2015 by the Department of Human Services under the pretext of increasing efficiency in identifying and recovering debts from overpaid social welfare benefits.

The programme faced widespread criticism and legal challenges for placing the burden of proof on welfare recipients to disprove these automated debt claims—a reversal of the standard legal responsibility typically held by government. This practice was found to be legally unsound by the Federal Court of Australia in 2019, which led to a class-action settlement of AUD $1.2 billion to affected individuals.

In 2022, a Royal Commission was established to investigate the origins and failures of the Robodebt scheme. It identified significant governance issues, lack of oversight, and ethical failings within the departments involved (cp. Royal Commission into the Robodebt Scheme 2023).

AUSTRIA

Austrian AI governance represents a multi-layered and collaborative approach. Key actors in the federal government are the Federal Chancellery (BKA) and the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation, and Technology (BMK), central to shaping AI policy, and jointly coordinating the AI Policy Forum, which integrates the federal ministries in an effort to ensure the alignment of the Austrian AI governance. Mind you, government coordination is no easy task (Dinges et al 2018)!

The policy forum was established in 2021 and its trick is to be a thematic working group to support the federal ministries in applying the AI strategy AIM AT 2030 (BMK and BMDW, 2021) - and developing it further. It promotes the exchange of experiences and approaches to the use of AI in the federal ministries and their discussion of current issues relating to artificial intelligence. A key element of the AI Policy Forum is the establishment of ad hoc working groups on various AI-related topics. For example, a working group has been established on the implementation of the AI Act of the EU. The forum tries to stay connected to the Austrian AI eco system and, e.g., invites experts from research, business, social partners, and NGOs to showcase their perspectives.

In this way a number of governance elements have already been created, from an AI eco-system map to a list of AI projects within the federal government, and a webpage for the national AI strategy (www.ki-strategie.at) hosting policy measures and good practice examples for AI governance from ministries, research institutions, and international organizations.

Austria's overarching framework for AI governance is laid out in the AIM 2030 strategy, which was launched in 2019 and updated in 2024. AIM AT 2030 emphasizes AI's role in achieving national goals in innovation, sustainability, and public welfare, aligning with the broader EU objectives.

Austria has established an institutional framework to support its AI agenda, with prominent roles for both National and European Digital Innovation Hubs. These hubs, supported by the Austrian Research Promotion Agency (FFG) and coordinated under programs such as Horizon Europe and Digital Europe, are to foster AI innovation and provide support for digital transformation across sectors.

At the policy advisory level, FORWIT (Austrian Council for Sciences, Technology, and Innovation) provides strategic advice to top-level policymakers. FORWIT brings together industry leaders, academic experts, and representatives from public institutions to address emerging issues also in AI and recommend policy adjustments.

The bulk of scientific policy support is organized through a community of policy advice organizations such as the AIT Austrian Institute of Technology, which is partially owned by the state, represented by BMK. In addition, AIT researchers not only develop AI, but also work in AI ethics and support the government, e.g., through partaking in debates on research programs and in co-developing policy documents such as the practical guidelines for AI in the public administration (BMKÖS, 2023; 2024b), and the AI Guidelines (BMKÖS, 2024a).

Austria's Service Desk for AI located at the Austrian Regulatory Authority for Broadcasting and Telecommunications (RTR) is an effort to make AI accessible and beneficial for both companies and the public. The service desk is to support SMEs, guide the public on AI-related issues, and build trust by providing transparent and practical information.

Both AI Advisory Board and AI Stakeholder Forum were created in 2024. The Advisory Board is an additional body within RTR consisting of 11 members from ethics, research, economics, law, and technical sciences advising government on AI. The AI Stakeholder Forum is an initiative organised by BKA and BMK

bringing together actors from federal government and various stakeholders outside government who are involved in or affected by the topic of AI in different ways.

Regulatory discussions and documents on AI governance in Austria are deeply embedded in European regulatory processes. Being part of the EU, Austrian civil servants, and experts have been part of European regulatory activities early on, most importantly in the AI Act, the first comprehensive binding regulation on AI in the democratic world.

Accordingly, documents such as the Austrian AI Strategy AIM AT 2030 (BMK and BMDW, 2021), the AI Guidelines (BMKÖS, 2024a), the Strategy Digital Competencies Austria (BMF, BMKOES, BMAW, BMBWF, 2023) or the Practical Guidelines on AI in the Public Administration (BMKÖS, 2023; 2024b) are all heavily influenced by EU discussions.

Regional actors also play a role in Austrian AI governance, Vienna being the leading example, having established its own guidelines, training, and research initiatives already some time ago. Other states such as Styria, Carinthia, and Upper Austria more recently have started initiatives driven by state and industry players.

There is a place for niches in the global AI race, exemplified by an Austrian strength in quantum computing, with several strong research groups and the legacy of Anton Zeilinger, the 2022 physics Nobel Prize winner - the second Nobel Prize in quantum physics for an Austrian, after Erwin Schrödinger in 1933 (indeed, the guy with the, depending on your point of view, more or less dead cat). An AIT working group is successfully translating basic science results into quantum encryption products, several firms are working on encryption technologies.

With regards to AI in government there have been no major scandals in Austria. However, the topic became part of public discussions already several times, e.g., when there was a debate on a chat bot of the Public Employment Service AMS based on Open AI's ChatGPT, which was found to be biased against gender and ethnicity (Der Standard, 2024a).

Table 2. Actors and Policies of AI Governance in Australia and Austria

	Australia	Austria
Main government actors: ministries and agencies	DSIR (industry and broad public), DTA (public administration)	BKA (coordination), BMK (industry, applied research), BMKÖS (public administration)
Main government actors: independent regulators	ACMA, OAIC, eSafety Commissioner, Privacy Commissioner	RTR (incl. AI Service Desk), DSB Data Protection Authority
Independent regulators' stance	Political interpretation of legal framework	Technocratic interpretation of legal framework
State-level activities	Large states (NSW, QLD) almost faster than federal level	Large city-state Vienna on par with federal level
Policy coordination	DISR/DTA federal level, DDMM federal-state level	AI Policy Forum (12 federal ministries), more informal federal-state meeting formats
Scientific policy support	National science agency CSIRO as hub	FORWIT (policy related), e.g., AIT (science related), AI Advisory Board (transdisciplinary)
Public debates on policies	High key: Robodebt scandal politicized discussions	Low key: ChatGPT-based AMS Infomat output biased

NOW THAT WE HAVE COME TO KNOW EACH OTHER...

By way of drawing conclusions, several observations can be made. First, the location determines the perspective. If you are located in the midst of an ocean, you will have to look farther abroad than when your location is in the midst of a continent consisting of dozens of small states. Whilst Australia is acutely aware of what Europe, the US, and China are doing, Austria is concentrated on the EU. More concretely, whilst Austria focuses very much on the EU, Australia has oriented itself towards the international British-led AI Safety process, which started with the Bletchley Park Declaration in November 2023 (PMO, FCDO and DSIT, 2023), however taking regularly notice in their policy documents and

reacting accordingly to developments in the EU and the US alike (e.g., DISR, 2023; 2024a). There are pros and cons to each, provided the limited policy intelligence resources of small countries. But certainly, Austria could learn something from looking a bit farther than just the EU.

Then again, the EU AI Act from 2024, with all its gaps and flaws, is promising in terms of trustworthy AI. It has been charged by NGOs for being overly permissive, and by industry for stifling innovation. Both actor groups have a point, but there comes the time, when - after years of negotiation - you have to draw a line and issue a regulation. The US have followed a different path and, for the time being, ended up with literally dozens of bills on federal and state levels, which never managed to be passed, because lawmakers feeling the pressure from big tech companies could not come to a decision. Similarly, until now the Australian debates whether there should be a national AI law or not, have not come to the point of actually working on a possible regulation. And although it is much more likely that Australia will pass such a law than the USA under President Trump (and the increasingly libertarian IT industry in Silicon Valley), valuable time will have passed before the governance of AI can become more tangible.

Indeed, Australia has chosen a particularly innovation- and business-oriented approach to AI regulation with a focus on international competitiveness. The regulatory approach chosen by the EU (with Austrian policy-makers partaking in the processes) stresses ethical thinking, data, and civil rights protection. The question remains if an emphasis on ethics and civil rights automatically translates into innovation hostility (as the argument sometimes goes) or if "AI made in Europe" in the future might stand for "trustworthy AI", which might well make a difference for the consumers and thus become a strength instead of a weakness.

The Australian governance system has been proliferating agencies, regulators, and policy documents. Independent regulators have proven their strengths by taking on international actors such as US tech giants - with governments and NGOs all over the world closely watching (and sometimes acutely envying) Down Under. Austrian AI governance is more attuned to a coordination and/or centralization of actors. Moreover, there is little political will to confront US or Chinese hyperscalers. Yet, the coordination attempted by the AI Policy Forum on the ministerial shop floor level is remarkable and innovative, as it leads to actual interorganizational exchange and learning.

Regarding the nuts and bolts of the civil service, i.e., the utilization of AI applications for the daily business of government, the handouts of the DTA are outstanding in their good accessibility and practicality. Moreover, the AGA is

an excellent place to find AI related documentation. However, and I readily confess to be anything but impartial on this matter (being one of its authors), the guidelines on AI in the administration commissioned by the BMKÖS (2023, 2024b) may have no equivalent, combining technical, legal, ethical, and civil service perspectives on AI in a single document. The guidelines have been developed and tested with civil servants from several ministries to include their domain knowledge and viewpoints.

The civil service in both countries have engaged on working with AI solutions, certainly on an individual, but also increasingly on a collective organizational level. The Australian DTA has carried out its Australian Public Service (APS) Trials with more than 50 agencies, resulting in a very interesting process of learning. The ensuing evaluation report is something also actors abroad can draw from (DTA, 2024).

In spring 2024 the Austrian Ministry of Education, Science and Research (BMBWF) has started a test of AI applications such as OpenAI's ChatGPT via Microsoft's cloud computing platform Azure for 250 civil servants with positive results and plans open access to these AI applications for all 1.200 staff members (Der Standard, 2024b). According to the implementation plan of the Austrian AI strategy AIM AT 2030 from fall 2024 other federal ministries plan to follow swiftly (BKA and BMK, 2024). An evaluation of these activities similar to the Australian APS Trials would be helpful for learning from experience.

The Australian National AI Center NAIC has developed a multitude of engaging activities primarily for the good of society and SMEs, including conferences, workshops, training videos, as well as informational material of all sorts and for different target groups. It cleverly includes the Responsible AI Network RAIN, which like NAIC is a cooperation of business, research, and civil society organizations. The Austrian Service Desk for AI has focused on different set of activities, which are similarly important. However, a perspective such as offered by RAIN currently is missing in Austria.

What has been recognized in both countries is that a balanced combination of innovation, ethics, and regulation is crucial for a sustainable AI strategy. Yet chosen solutions display remarkable variety. To summarise: similar problems, different solutions - and clear potential for both sides to learn from each other at the level of structures, processes, and tools.

ACKNOWLEDGEMENTS

The author is grateful to Michael Wiesmüller and Daniela Muhrhammer-Sas from BMK for debating some of the above issues in July 2024 , to Thomas König for the possibility to discuss a more encompassing set of arguments at the Austrian FORWIT in September 2024, as well as to two anonymous reviewers for helpful comments in November 2024. AI-tools have been used for ideation and identifying literature, with the sole responsibility being with the author.

REFERENCES

Acemoglu, Daron (2024). "The Simple Macroeconomics of AI." SSRN Electronic Journal. https://doi.org/10.3386/w32487 (retr. 13 December 2024)

Australian Bureau of Statistics (2024). Australian Industry: Annual estimates of key economic and financial performance of industries in Australia, including income, expenses, profit and capital expenditure, release 31 May 2024. https://www.abs.gov.au/statistics/industry/industry-overview/australian-industry/latest-release (retr. 1 November 2024)

Biegelbauer Peter, Alexander Schindler, Rodrigo Conde-Jimenez, and Pia Weinlinger (2024). "Exciting Opportunities and Necessary Safeguards for Large Language Models in the Public Sector", ERCIM News, Vol. 136, pp.25-26, https://ercim-news.ercim.eu/en136 (retr. 1 November 2024)

Biegelbauer Peter, Anahid Jalali, Sven Schlarb, Michela Vignoli (2022). „Ethical AI: Why and How?", October 2022, ERCIM News, Vol. 131, pp.8–9, https://ercim-news.ercim.eu/en131/special/ethical-ai-why-and-how (retr. 1 November 2024)

BKA and BMK (2024), Strategie der Bundesregierung für Künstliche Intelligenz Umsetzungsplan 2024, https://www.ki-strategie.at/umsetzungsplan-2024/ (retr. 13 December 2024)

BMF, BMKOES, BMAW, BMBWF (2023), Strategie Digitale Kompetenzen Österreich, https://www.digitalaustria.gv.at/dam/jcr:e84a42c3-f2e7-4642-9ca0-76d7e8c61216/Strategie-Digitale-Kompetenzen-Oesterreich-PDF-UA-1.pdf (retr. 13 December 2024)

BMK, BMDW (2021) Strategie der Bundesregierung für Künstliche Intelligenz: Artificial Intelligence Mission Austria 2030 (AIM AT 2030), https://www.digital-

austria.gv.at/dam/jcr:6dacb3c5-ca2b-4751-9653-45ed8765cacd/AIM_AT_2030_UAbf.pdf (retr. 13 December 2024)

BMKÖS (2024a), KI-Guidelines: Empfehlungen zur Nutzung digitaler Informations- und Gestaltungsangebote im Arbeitsprozess, https://oeffentlicherdienst.gv.at/verwaltungsinnovation/public-management-und-governance/digitale-verwaltung/ki-guidelines/ (retr. 13 December 2024)

Bundesministerium für Kunst, Kultur, öffentlicher Dienst und Sport (2023). Leitfaden Digitale Verwaltung und Ethik: Praxisleitfaden für KI in der Verwaltung, Version 1.0. https://oeffentlicherdienst.gv.at/wp-content/uploads/2023/11/Leitfaden-Digitale-Verwaltung-Ethik.pdf (retr. 1 November 2024)

Bundesministerium für Kunst, Kultur, öffentlicher Dienst und Sport (2024b). Leitfaden Digitale Verwaltung: KI, Ethik und Recht: Praxisleitfaden für die Verwaltung, Version 2.0.

Department of Industry, Science and Resources (2019). AI Ethics Framework. Canberra.

Department of Industry, Science and Resources (2023). Safe and responsible AI in Australia discussion paper. https://consult.industry.gov.au/supporting-responsible-ai (retr. 1 November 2024)

Department of Industry, Science and Resources (2024a). Australian Government's interim response to a safe and responsible AI consultation https://www.industry.gov.au/news/australian-governments-interim-response-safe-and-responsible-ai-consultation (retr. 1 November 2024)

Department of Industry, Science and Resources (2024b). Applying the AI Ethics Principles https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/australias-ai-ethics-principles (retr. 1 November 2024)

Der Standard (2024a). „Blamage um Chatbot des AMS wirft weitere Fragen auf". 19 January 2024. https://www.derstandard.at/story/3000000203652/chatbot-ams (retr. 4 November 2024)

Der Standard (2024b), „Bildungsministerium testet als erstes Ressort KI-Anwendungen", 31 July 2024, https://www.derstandard.at/story/3000000230589/bildungsministerium-testet-als-erstes-ressort-ki-anwendungen (retr. 13 December 2024)

Digital Transformation Agency (2024). Evaluation of the whole-of-government trial of Microsoft 365 Copilot. https://www.digital.gov.au/initiatives/copilot-trial (retr. 29 October 2024)

Dinges, Michael, Peter Biegelbauer and Doris Wilhelmer (2018). "The Tower of Babylon in the Governance of Research, Technology and Innovation: Participatory Foresight as a Method of Policy Coordination." Futures 100: 34-44.

European Commission (2024). European innovation scoreboard. https://research-and-innovation.ec.europa.eu/statistics/performance-indicators/european-innovation-scoreboard_en (retr. 1 November 2024)

European Commission High-level expert group on artificial intelligence (2019). Ethics Guidelines for Trustworthy Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (retr. 29 October 2024)

European Union (2024). Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj/eng (retr. 29 October 2024)

Galindo Laura, Karine Perset and Francesca Sheeka (2021), "An overview of national AI strategies and policies", OECD Going Digital Toolkit Notes, No. 14, OECD Publishing, Paris, https://doi.org/10.1787/c05140d9-en (retr. 1 November 2024)

Gartner (2017). Top Trends in the Gartner Hype Cycle for Emerging Technologies. https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017 (retr. 1 November 2024)

Huw Roberts, Josh Cowls, Emmie Hine, Jessica Morley, Vincent Wang, Mariarosaria Taddeo and Luciano Floridi (2023). "Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes". The Information Society, 39:2, 79-97, https://doi.org/10.1080/01972243.2022.2124565

International Data Corporation (2024). "Artificial Intelligence Will Contribute $19.9 Trillion to the Global Economy through 2030." https://www.idc.com/getdoc.jsp?containerId=prUS52600524 (retr. 29 October 2024)

International Monetary Fund (2024). "AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity." https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity (retr. 29 October 2024)

McKinsey (2023). “The State of AI: Productivity Gains and Economic Impact Predictions.” (retr. 29 October 2024) https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier (retr. 29 October 2024)

OECD Nominal GDP (2024b), https://www.oecd.org/en/data/indicators/nominal-gross-domestic-product-gdp.html? (retr. 13 December 2024)

OECD.AI Policy Observatory (2024a), https://oecd.ai/en/ (retr. 13 December 2024)

PMO, FCDO, and DSIT (2023), The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 (retr. 13 December 2024)

Royal Commission into the Robodebt Scheme (2023). Report Vol 1. https://robodebt.royalcommission.gov.au/ (retr. 29 October 2024)

Stahl Bernd (2021). Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies. Springer.

Statistics Austria (2024). Gross domestic product and main aggregates. https://www.statistik.at/en/statistics/national-economy-and-public-finance/national-accounts/gross-domestic-product-and-main-aggregates (retr. 29 October 2024)

The Guardian (2024). “eSafety commissioner ends heated fight with Elon Musk’s X over Sydney church stabbing posts” on 11 October 2024. https://www.theguardian.com/technology/2024/oct/11/esafety-commissioner-ends-heated-fight-with-elon-musk-x-over-sydney-church-stabbing-posts-ntwnfb (retr. 1 November 2024)

Walter Yoshija (2024). “Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation - A contemporary overview and an analysis of socioeconomic consequences.” Discov. Artif. Intell. 4. https://doi.org/10.1007/s44163-024-00109-4

Zou James and Londa Schiebinger (2018). “AI can be sexist and racist — it’s time to make it fair.” Nature 559: 324 - 326. https://doi.org/10.1038/d41586-018-05707-8 (retr. 13 December 2024)

AUTHOR

PETER BIEGELBAUER
AIT Austrian Institute of Technology
Giefinggasse 4
1210 Wien
peter.biegelbauer@ait.ac.at
ORCID 0000-0002-4113-6524

EDITORIAL BOARD

Published by the Austrian Platform for Research and Technology Policy Evaluation (fteval)
Linke Wienzeile 246, c/ZSI, 1150 Vienna, Austria
T: +43 1 495 04 42 - 33, **E:** office@fteval.at, **W**: www.fteval.at

The fteval Journal for Research and Technology Policy Evaluation is a forum to discuss current evaluation practices in the field of research, technology and innovation policy.

AUSTRIAN PLATFORM FOR RESEARCH AND TECHNOLOGY POLICY EVALUATION MEMBERS:

Austrian Federal Ministry of Education, Science and Research (BMBWF), Austrian Federal Ministry of Labour and Economy (BMAW), Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), Austrian Cooperative Research (ACR), Austrian Council for Research and Technology Development, Austrian Institute of Technology (AIT), AQ Austria – Agency for Quality Assurance and Accreditation Austria, Austria Wirtschaftsservice (AWS), Christian Doppler Research Association (CDG), convelop Cooperative Knowledge Design GmbH, FH Campus Wien , FORWIT - Austrian Council for Sciences, Technology and Innovation, Austrian Research Promotion Agency (FFG), Austrian Science Fund (FWF), Institute for Advanced Studies (IHS), Industriewissenschaftliches Institut (IWI), Joanneum Research Forschungsgesellschaft mbH, Austrian Institute for SME Research (KMU Forschung Austria), Ludwig Boltzmann Society (LBG), Agentur für Bildung und Internationalisierung (OeAD), Austrian Academy of Sciences (OEAW), Technopolis Austria, Vienna Business Agency – A service offered by the City of Vienna, Austrian Institute of Economic Research (WIFO), WPZ Research GmbH, Vienna Science and Technology Fund (WWTF), Centre for Social Innovation (ZSI)

ZVR-Zahl: 937261837, **ISSN-Nr**.: 1726-6629, **DOI:** 10.22163/fteval.2024.633

AUSTRIAN PLATFORM
for Research and Technology Policy Evaluation
fteval

Linke Wienzeile 246, c/o ZSI
1150 Vienna, Austria

+43 1 495 04 42 - 33
office@fteval.at
www.fteval.at

The fteval Journal for Research and Technology Policy Evaluation is a forum to discuss current evaluation practices in the field of research, technology and innovation policy.

ZVR-Zahl: 937261837
ISSN-Nr. 1726-6629
DOI: 10.22163/fteval.633