# UK Evaluation Society 2023 Annual Conference

## (London, 2. – 4. Oktober 2023)

# Konferenzbericht

Alkuin Kölliker, Fachbereich Evaluationen
Eidgenössische Finanzkontrolle (EFK)

*Ziel des vorliegenden Konferenzberichtes ist es, auf Grundlage der besuchten Veranstaltungen, des Konferenzprogramms sowie informeller Gespräche am Rande der Konferenz einen Eindruck zu geben, welche Themen die Evaluationsgemeinschaft im Vereinigten Königreich sowie ihre Gäste aus anderen Ländern aktuell beschäftigen. Dabei sollen insbesondere Aspekte berücksichtigt werden, welche auch für die Evaluationsgemeinschaft in der Schweiz von Interesse sein können. Da die maximal mögliche Teilnahme aufgrund der verschiedenen parallel geführten Sessions auf rund einen Drittel der Veranstaltungen beschränkt war, können die im vorliegenden Bericht wiedergegebenen Informationen und Eindrücke nur einen Teil des Konferenzgeschehens abbilden.*

**Inhalt**

# 1 Einleitung – Konferenzprogramm und Teilnehmende

Die dreitägige Jahreskonferenz 2023 der UK Evaluation Society (UKES) fand am 2. – 4. Oktober 2023 statt. Die Konferenz trug den Titel «Rising to Challenges: How does Evaluation rise to the challenge of competing issues, crisis and uncertainty?»

Der erste Konferenztag wurde ausschliesslich in virtueller Form durchgeführt (online). Die beiden weiteren Konferenztage wurden im «Friends House» an der Euston Road in London abgehalten (hybride Veranstaltung mit persönlicher Teilnahme sowie Möglichkeit zur Online-Teilnahme). Gemäss der Teilnehmerliste nahmen an der Konferenz mehr als 430 Personen teil, darunter 183 als Vortragende und 251 als (ausschliesslich) Zuhörende.

Nach Auskunft der Organisatoren stieg die Anzahl der Mitglieder der UKES in den vergangenen Jahren stark an, von rund 400 auf gegenwärtig etwa 1100. Die UKES entwickelte in dieser Zeit zusätzlich zu den herkömmlichen Jahreskongressen weitere Aktivitäten und Veranstaltungen zwischen diesen Kongressen. Im Vergleich zur SEVAL in der Schweiz scheint in der UKES der akademische Sektor stärker vertreten zu sein, während staatliche Auftraggeber schwächer vertreten sind.

Gemäss der Programmübersicht beinhaltete die Konferenz in insgesamt 17 Sessions (mit jeweils zwei oder fünf parallelen Veranstaltungen) insgesamt 64 Veranstaltungen mit 111 Beiträgen. Zusätzlich waren fünf Keynote-Präsentationen im Plenum vorgesehen. Einzelne Teilnehmende konnten aufgrund der parallel durchgeführten Sessions somit an maximal 22 Veranstaltungen (also rund einem Drittel aller Veranstaltungen) teilnehmen. Die Veranstaltungen waren in vier Themenbereiche (*Tracks*) unterteilt:

- Transdisciplinary evaluation approaches
- Ethics for evaluation, and EDI (equality, diversity and inclusion) in evaluative practice
- Innovative evaluation approaches
- Teaching and learning of evaluative practice

Zusammenfassende Informationen zu einzelnen Beiträgen wurden im detaillierten Konferenzprogramm veröffentlicht. Die dazugehörigen Folienpräsentationen wurden aber weder während noch nach der Konferenz verfügbar gemacht.

# 2 Einblicke in einzelne besuchte Veranstaltungen

*In diesem Abschnitt werden besonders relevant erscheinende Inhalte aus ausgewählten besuchten Veranstaltungen wiedergegeben. Nicht alle besuchten Veranstaltungen können dabei berücksichtigt werden. Der Fokus der besuchten Veranstaltungen lag bei den beiden Themenbereichen innovative bzw. transdisziplinäre Evaluationsansätze.*

### Improving Evaluation in Government

*(Montag 2. Oktober 2023, Session 1, Track Innovation)*

Die Evaluation Task Force ist eine seit 2020 bestehende gemeinsame Einheit des britischen *Cabinet Office* und des Finanzministeriums, welche 2020 etabliert wurde. Sie soll mittels Fachunterstützung für die zuständigen Verwaltungseinheiten dazu beitragen, dass Evidenz und Evaluationen im Mittelpunkt von Ausgabenentscheidungen stehen.

Die Präsentation konzentrierte sich auf das Monitoring von Evaluationen bei rund 250 Grossprojekten mit einem Gesamtvolumen von rund 800 Mrd. Pfund (ca. 880 Mrd. Franken). Während 2019 nur 8 % dieser Ausgaben von «robuster» Evaluation begleitet wurden, konnte dieser Anteil inzwischen auf – immer noch relativ bescheidene – 16 % verdoppelt werden. «Robuste» Evaluation wird durch die Anwendung angemessener Methoden und durch weitere Qualitätsmerkmale charakterisiert.

In der Präsentation wurde anhand von Beispielen auf spezifische Herausforderungen bei der Evaluation von Grossprojekten eingegangen, insbesondere auf deren hohe Komplexität, lange Dauer und grossen Umfang. Als besonders problematisch wird der Mangel an Know-how zur Evaluation von solchen Projekten eingeschätzt (*skills gap*).

Für die Evaluationsgemeinschaft in der Schweiz von Interesse könnten die methodischen Leitlinien für Evaluationen sein, welche 2020 vom britischen Finanzministerium im sogenannten Magenta Book veröffentlicht wurden. Diese Publikation wurde auch in anderen Veranstaltungen der Konferenz wiederholt als Referenzdokument thematisiert.

Im Rahmen der Präsentation wurde auf einen Bericht des britischen Rechnungshofs (National Audit Office) zur Rolle von Evaluationen bei der Arbeit der Regierung hingewiesen.[1]

### Theory Based Evaluation Meets Econometrics

*(Montag 2. Oktober 2023, Session 3, Track Transdisciplinary)*

Nach dem Austritt aus der EU wurden in Grossbritannien insgesamt zwölf neue Zollfreigebiete (*free ports*) errichtet. In diesen meist in Küstennähe gelegenen Zonen gelten besondere Regelungen insbesondere zur Verzollung von Waren und zur Besteuerung von Unternehmen.

In einer vom Beratungsunternehmen Arup durchgeführten umfassenden Untersuchung sollen insbesondere die wirtschaftlichen Auswirkungen der Zollfreigebiete begleitend evaluiert werden. Bemerkenswert bei dieser noch nicht abgeschlossenen Evaluation ist der Umfang in verschiedenen Dimensionen. Es werden im Rahmen von 5 Haupt- und 34 Unterfragen Analysen auf der Mikro-, Meso- und Makroebene durchgeführt (einzelne Unternehmen / einzelne Zollfreigebiete / Summe aller Zollfreigebiete).

Die Evaluation kombiniert ökonometrische und theoriebasierte Methoden.[2] Dabei waren zunächst zwei bis drei Jahre für Theoriebildung zu den konkreten Wirkungsmechanismen im Zusammenhang mit den Zollfreigebieten vorgesehen. Danach sind vier bis fünf Jahre eingeplant für das Testen der resultierenden Theorien mit Hilfe ökonometrischer und weiterer Methoden, wie etwa *propensity score matching* und *difference-in-differences* (DiD). Die Resultate der Ergebnisse aufgrund verschiedener Methoden werden danach durch Triangulation zusammengeführt.

Speziell bei dieser Evaluation ist die Komplexität der Wirkungsmechanismen, welche einen theoriebasierten Ansatz begünstigt. Es brauchte viel Zeit, die notwendigen vertieften Kenntnisse zum Programm selbst sowie zur Situation in den einzelnen Zollfreigebieten zu erlangen. Ein Vorteil war die Verfügbarkeit umfangreicher relevanter Daten.

Offen blieb in der Präsentation und der anschliessenden Diskussion der genaue Prozess, in welchem über einen längeren Zeitraum die Hypothesen zu den kausalen Verbindungen im Wirkungsmodell für diese theoriegeleitete Evaluation entwickelt wurden. Damit blieb der Aspekt der Theorie innerhalb der «theory based evaluation» schwer nachvollziehbar.

### Large Language Model Applications for Evaluation

*(Montag 2. Oktober 2023, Session 4, Track Innovation)*

Im Bereich der Künstlichen Intelligenz (KI) bringen *Large Language Models* (LLMs) für Evaluierende vielversprechende Möglichkeiten, aber auch substantielle Risiken mit sich.

LLMs sind die Technologie hinter Anwendungen wie ChatGPT. Mit ihrer Hilfe könnten bei Evaluationen zeitaufwändige Aufgaben der Textanalyse und Texterstellung automatisiert und beschleunigt werden. Zu den Anwendungsbeispielen gehören die Zusammenfassung von Textdaten, die Extraktion von Schlüsselinformationen aus Texten, die Analyse und Klassifizierung von Textinhalten, das Schreiben von Texten und die Übersetzung. In der Präsentation wurde

[1] Vgl. National Audit Office (2021, 2022).

[2] Vgl. dazu Unterabschnitt «Anwendung theoriebasierter Evaluationsansätze» in Abschnitt 3.

davon ausgegangen, dass mehr als zwei Drittel der Evaluationsaufgaben in den nächsten fünf Jahren von LLMs betroffen sein werden. Evaluationen könnten kostengünstiger, schneller und somit für Auftraggeber attraktiver werden. Es wird damit gerechnet, dass in Zukunft LLM-Funktionen vermehrt auch in Standard-Anwendungen wie Word oder Excel eingebaut werden.

Die Verwendung von LLMs beinhaltet jedoch auch wesentliche Risiken. Im Rahmen der Präsentation thematisiert wurden Probleme beim Datenschutz, der teilweise hohe Preis einzelner Anwendungen und der (beispielsweise sprachbedingte) Ausschluss bestimmter Arten von Quellen. Weitere schwerwiegende Risiken, welche in anderen Veranstaltungen der Konferenz identifizierte wurden (etwa die Produktion falscher und methodisch intransparenter Resultate durch die «Black Box» der KI) wurden in dieser Veranstaltung nicht vertieft thematisiert.[3]

Interessant war ein Hinweis auf die von MERL Tech organisierte Natural Language Processing Community of Practice. Diese erlabt es Evaluierenden, sich über die Chancen und Risiken der Anwendung von LLMs bei Evaluationen auszutauschen. Das Forum MERL Tech wurde gemäss den (spärlichen) Angaben auf der Internetseite offenbar im Zusammenhang mit Aktivitäten der Rockefeller-Stiftung im Jahr 2014 etabliert. Die Abkürzung «MERL» steht dabei für «monitoring, evaluation, research, and learning».

Die Präsentation erfolgte durch einen Mitarbeiter von NIRAS, einer in Dänemark beheimateten Beratungsfirma im Bereich der Entwicklungszusammenarbeit mit weltweit 2600 Beschäftigten. NIRAS ist auch in der Schweiz vertreten und hat in den letzten Jahren verschiedene Beratungsmandate von Bundesstellen (SECO, DEZA) erhalten.

### Five Ways for Evaluation to Develop Value Amidst Complexity

*(Montag 2. Oktober 2023, Keynote 1)*

In ihrer Keynote-Präsentation stellte Dr. Emily Gates (Assistant Professor, Boston College) das herkömmliche Verständnis von Evaluationen in Frage, indem sie neuere Entwicklungen in verschiedenen Dimensionen vorstellte – und gleichzeitig auch einforderte.

Sie stellte diese Entwicklungstrends in einen Zusammenhang mit aktuellen gesellschaftlichen und politischen Entwicklungen (multiple Krisen vor dem Hintergrund sozialer Ungerechtigkeit und ökologischer Probleme), welche in der Evaluationsgemeinschaft Selbstkritik und Innovationen ausgelöst haben. Konkret thematisierte sie die Entwicklung der Evaluationspraxis in den folgenden fünf Dimensionen:

- **Gesellschaftliche Rolle der Evaluation:** Von der Lösung (einzelner) sozialer Probleme zu (umfassenderem) Verantwortungsbewusstsein und Gerechtigkeitsstreben
- **Verhältnis zwischen Auftraggebenden und Auftragnehmenden:** Von der Zufriedenheit der Auftraggebenden zu gegenseitiger Rechenschaftspflicht sowie gemeinsamen Lern- und Anpassungsprozessen
- **Analyserahmen für Veränderungen:** Von im Vorfeld entwickelten logischen Modellen zur gemeinsamen Entdeckung und Darstellung von Veränderungen
- **Beurteilung und Einsatz von Methoden:** Von Hierarchien zwischen einzelnen Methoden (von *Randomized Control Trials* ganz oben bis zu anekdotischer Evidenz ganz unten) zu kreativem Methodenmix
- **Funktion der Evaluation:** Von der Beurteilung der Leistungserbringung zur gemeinsamen Entwicklung von Mehrwert in einem komplexen Kontext

In der Diskussion wurde unter anderem festgestellt, dass grössere (darunter vermutlich insbesondere staatliche) Auftraggebende aktuell eher noch resistent gegenüber den skizzierten Trends sind. Ebenfalls wurde die Sorge thematisiert, dass der Methodenreichtum in der Evaluation durch eine allzu starke Richtungsvorgabe im Sinne dieser Trends beeinträchtigt werden könnte.

[3] Vgl. dazu Unterabschnitt «Zukünftige Rolle von Künstlicher Intelligenz bei Evaluationen» in Abschnitt 3.

Die Präsentation vermittelte grundlegende Denkanstösse zur Weiterentwicklung der Evaluation und ihrer Rolle in Politik und Gesellschaft – dies jedoch nicht nur alternativ, sondern durchaus auch komplementär zu bewährten Praktiken herkömmlicher Evaluationen.

### A Fireside Chat with Kelly Beaver

*(Dienstag 3. Oktober 2023, Keynote 3)*

Die Keynote-Veranstaltung mit Kelly Beaver wurde in Form eines «Kaminfeuergesprächs» mit Fragen und Antworten durchgeführt. Als Vorsitzende von Ipsos UK und Irland hat Beaver ihre eigene Sicht auf verschiedene die Evaluationsgemeinschaft betreffende aktuelle Entwicklungen dargelegt. Ipsos ist ein weltweit tätiges führendes Marktforschungsunternehmen mit 18 000 Beschäftigten und Aktivitäten im Bereich Policy Evaluation. Ipsos war mit rund zehn Personen an der UKES-Konferenz vertreten.

Interessant waren insbesondere die folgenden Punkte zu Entwicklungen und Trends, welche die Evaluation betreffen:

- **Herausfordernde Kommunikation mit Politik:** In einem schwierigen, von Abstiegsängsten und verschiedenen Krisen (*declinism, polycrisis*) geprägten politischen Kontext möchten britische Kabinettmitglieder einerseits unbedingt wissen, welche Lösungen in der Realität funktionieren. Die Kommunikation mit der Politik bleibt für die Evaluation trotz einiger Fortschritte aber eine Herausforderung, weil Politiker/innen eher an konkreten Beispielen und Anekdoten als an detaillierten Evaluationsresultaten interessiert sind.
- **Weiterhin ungenügende Professionalisierung:** Eine Akkreditierung für Evaluierende und Evaluationsorganisationen wurde in der Vergangenheit bereits erfolglos angestrebt, wäre aber weiterhin wichtig. 70 % der heute tätigen Evaluierenden sind nach eigener Aussage «per Zufall» zur Evaluationsgemeinschaft gestossen.
- **Schwieriges Verhältnis zur Ökonomie:** Das Verhältnis zwischen der Wirtschaftswissenschaft und anderen für die Evaluationspraxis relevanten Sozialwissenschaften bleibt problematisch. Ökonom/innen haben einen anderen Umgang sowohl mit Personen (als Untersuchungsgegenstand) wie auch zu benachbarten Disziplinen. Beaver selbst ist Ökonomin, findet es aber schwierig, Ökonom/innen als Evaluierende zu rekrutieren. Diese haben meist weitere bzw. andere berufliche Perspektiven. Wichtig sind dabei auch Ökonom/innen, welche «nicht nur Ökonometrie betreiben».
- **Erwartete Umwälzungen durch Künstliche Intelligenz und «Big Data»:** Die Künstliche Intelligenz (KI) wird bereits in den nächsten fünf bis zehn Jahren Umwälzungen (*disruption*) für die Evaluationsbranche mit sich bringen. Der Umgang mit «Big Data» und KI dürfte Investitionen erfordern, welche eine Konsolidierung in der Evaluationsbranche (in Richtung grösserer Unternehmen) herbeiführen dürften. Mit Hilfe von KI wird es weniger Zeit brauchen, Ergebnisse zu erreichen. Unklar ist im Moment, wie KI auf ethische Art und Weise angewendet werden kann.
- **Weniger Erhebungen von Primärdaten:** Aufgrund der bessern Verfügbarkeit von Administrativdaten und Daten aus Sozialen Netzwerken dürfte die Erhebung von Primärdaten bei Evaluationen in Zukunft eine geringere Rolle spielen.

### What Makes a «Good Evaluation»?

*(Mittwoch 4. Oktober 2023, Session 2, Track Innovation)*

Während die Notwendigkeit «robuster» Evaluationen guter Qualität immer mehr Anerkennung findet, ist weniger klar, was eine «gute» Evaluation ausmacht. In einer Panel-Veranstaltung haben sich Vertreter/innen des öffentlichen Sektors (Evaluation Task Force, UK Cabinet Office / Treasury), des privaten Sektors (RSM Consulting) und der Wissenschaft (Dr. Ine Steenmans, University College London) miteinander und mit dem Publikum über die Eigenschaften «guter» Evaluationen ausgetauscht.

Vonseiten der Evaluation Task Force wurden fünf Punkte als entscheidend identifiziert:

- **Fragen:** Die richtigen Evaluationsfragen stellen
- **Einbettung:** Evaluation sollte bereits in der Konzeptionsphase der Politik präsent sein
- **Methoden:** Einsatz von verhältnismässigen und «robusten» Evaluationsmethoden
- **Mehrwert:** Von Beginn weg auf Mehrwert der Evaluation achten (insbesondere auf Beitrag zu «4 E»: Efficiency, Effectiveness, Economy, Equity)
- **Nutzung:** Resultate der Evaluation maximal nutzen

Die weiteren Diskussionsbeiträge gingen teilweise in eine ähnliche Richtung, fügten aber auch weitere Elemente «guter» Evaluationen hinzu. Dabei wurden insbesondere die folgenden Themen und Fragen angesprochen (grob eingeordnet nach der zeitlichen Abfolge von der Konzeption bis zur Kommunikation der Resultate von Evaluationen):

- **Form der Evaluationspraxis:** In den vergangenen Jahren hat die britische Regierung mit verschiedenen Formen experimentiert, wie Expertise und empirische Evidenz über herkömmliche Evaluationen hinaus genutzt werden können.
- **«Learning Plan»:** Die Evaluation sollte von einem Lernplan zum entsprechenden Evaluationsgegenstand geleitet werden.
- **Verhältnis zu Evaluierten:** Die Barriere zwischen Evaluierenden und Evaluierten sollte durchbrochen werden.
- **Evaluationsdesign:** Angestrebt werden sollte eine angemessene Kombination von (quasi-) experimentellen mit theoriebasierten Evaluationsmethoden.
- **Methoden:** Über die Standard-Methoden hinaus weitere mögliche Methoden berücksichtigen (dabei insbesondere «Magenta Book» des britischen Finanzministeriums beachten).
- **«Learning by doing» / «Best practice»:** Es ist wichtig, «Learning by doing» zu ermutigen, aber auch zu beachten, wie Methoden in anderen Evaluationen verwendet wurden. Es sollte ein Verzeichnis von Evaluationen mit verschiedener Methodik geben.
- **«Thinking like Sherlock»:** Immer offen sein für neue und originelle Ansätze, ein Evaluations-Problem zu lösen.
- **Datensammlung und Beurteilung:** Es sollte es einen reflektierten Umgang mit der Art der gesammelten Daten, mit der Kombination der Daten aus verschiedenen Methoden sowie mit den Beurteilungen aufgrund der gesammelten Daten geben.
- **Zugang zu Daten:** Evaluationen benötigen zum Teil auch ältere Daten, welche ein Jahrzehnt zurückreichen. Die Verfügbarkeit von Umwelt-Daten ist hier vorbildlich.
- **Visualisierung:** Gute Visualisierung der Daten ist wichtig, inklusive innovative Arten der Kommunikation von Fallstudien-Resultaten.
- **Kommunikation:** Wie kann die Evidenz am besten gegenüber den Entscheidungsträgern kommuniziert werden? Beispiel: Durch Einnahme der Perspektive der Entscheidungsträger und Verwendung einer angemessenen Sprache.

Die Panel-Diskussion gelangte erwartungsgemäss nicht zu einer abschliessenden Liste von Qualitätsmerkmalen einer «guten» Evaluation. Sie gab jedoch einen guten Eindruck dazu, wo die Teilnehmenden Herausforderungen sehen für eine gute Evaluationspraxis in Bezug auf die Methodik und weitere wichtige Aspekte von Evaluationen (wie deren Einbettung, Kommunikation und Nutzung).

*Last but not least* wurde in der Diskussion auch der Umgang mit «unrealistisch niedrigen Budgets» der (britischen) Regierung zur Finanzierung einzelner Evaluationen angesprochen. Vonseiten der Evaluation Task Force wurde in diesem Zusammenhang ein Gesamtbudget von 100 Mio. GBP für

200 Evaluationen in deren Aufsichtsbereich genannt, was immerhin verfügbaren Mitteln von durchschnittlich rund 600 000 CHF pro Evaluation entsprechen würde.

### Deepening Evaluative Practice through Doctoral Research

*(Mittwoch 4. Oktober 2023, Session 3, Track Teaching & Learning)*

Die Universität Bath bietet seit acht Jahren ein berufsbegleitendes Doktoratsprogramm in Policy Research and Practice an. Ein Drittel der bisher mehr als 80 Teilnehmenden dieses Programms haben bereits Berufserfahrungen im Bereich der Evaluation. Speziell an diesem von Professor James Copestake vorgestellten Programm ist, dass weder die betreuende Person (Doktormutter bzw. -vater) noch das Thema zu Beginn der Programmteilnahme bereits feststehen müssen.

Die Studierenden werden ermutigt, bei der Entwicklung und Erstellung einer Forschungsarbeit auf ihre eigenen beruflichen Erfahrungen zurückzugreifen. Im Sinne der erforderlichen Interdisziplinarität vertiefen die Doktorandinnen und Doktoranden ihre Kenntnisse in verschiedenen relevanten Sozialwissenschaften.

Als wichtig wird betont, dass das Programm einerseits das Know-how zur Umsetzung von Evaluationen vermittelt, andererseits aber auch die kritische Reflexion von Erfolgen und Misserfolgen der Evaluationspraxis in verschiedenen institutionellen Kontexten fördert. Es gibt potentiell ein Spannungsverhältnis zwischen der technischen, der politischen sowie der ethischen Dimension von Evaluationen. Vor diesem Hintergrund ist gegenwärtig eine Suche nach einer «demokratischen Evaluationspraxis» im Gang.

## 3 Ausgewählte wiederkehrende Themen der Konferenz

*Im vorliegenden Abschnitt werden wiederkehrende Themen der verschiedenen besuchten Veranstaltungen identifiziert, welche auch im Kontext der Evaluationspraxis in der Schweiz von Interesse sein können. Dabei werden zunächst drei ausgewählte Themen detaillierter dargestellt, bevor zum Abschluss weitere Themen in Kurzform erläutert werden.*

### Reflexion zur Weiterentwicklung der Evaluationspraxis

Das Nachdenken und die Diskussion über die Weiterentwicklung der Disziplin der Politikevaluation nahm in verschiedenen Keynote-Präsentationen und in einzelnen weiteren Veranstaltungen einen wichtigen Platz ein. Dabei wird das bisher zumeist angestrebte Ideal der Politikevaluation als technokratisch-rationale Selbstoptimierung in einzelnen Politikfeldern durch verschiedene alternative oder komplementäre Sichtweisen der Rolle von Politikevaluation herausgefordert, deren gemeinsame Ambition eine ganzheitlichere Herangehensweise zu sein scheint.

Ein grosser Teil der im Rahmen der Konferenz präsentierten Evaluationen – wie auch die Anstrengungen von auftraggebenden und qualitätssichernden staatlichen Institutionen (so etwa der Evaluation Task Force der britischen Regierung) – zielen auf weitere Verbesserungen bei der Umsetzung herkömmlicher Evaluationsmethoden und -prozesse ab. Diese sind charakterisiert durch einen Fokus der Evaluation auf die in einzelnen Politikbereichen gesetzlich vorgegebenen Ziele, die Dominanz von vordefinierten Wirkungsmodellen sowie der «vier E» als Evaluationskriterien (*effectiveness, efficiency, economy, equity*), eine gewisse Hierarchie zwischen mehr bzw. weniger robusten Evaluationsmethoden und eine klare Aufteilung der Verantwortlichkeiten zwischen Auftraggebenden und Auftragnehmenden.

Diese Grundpfeiler bisheriger Evaluationspraxis werden von neueren Tendenzen infrage gestellt, die ihren Ursprung eher im akademischen Bereich sowie in den «Grassroots» einer gesellschaftlich und politisch idealistischeren Evaluationspraxis haben. Diese Tendenzen wurden in der Konferenz

in einer Präsentation von Emily Gates (Boston College) artikuliert und auf den Punkt gebracht.[4] Im Zusammenhang mit Veranstaltungen zur Evaluationsforschung und -lehre wurde generell für eine vermehrte Selbstreflexion zur Evaluationspraxis plädiert.[5]

In der Konferenz standen sich diese herkömmlichen sowie neuere Tendenzen der Politikevaluation immer wieder gegenüber, ohne dass sich daraus eine intensivere kontradiktorische Diskussion ergab. Stattdessen wurde verschiedentlich auf die mögliche Koexistenz und Komplementarität der verschiedenen Ansätze verwiesen.

Eine gewisse Einigkeit gab es bei der Tendenz, anstelle des enger verstandenen Begriffs der Evaluation vermehrt die als umfassender verstandenen Begriffe «evaluative practice» und «evaluative thinking» zu verwenden – und auch entsprechende Aufträge zu vergeben, welche diese breiter definierte Evaluationspraxis widerspiegeln.[6]

Als Hintergrund der Reflexion über die Weiterentwicklung der Evaluationspraxis wurden die verschiedenen politischen, gesellschaftlichen und ökologischen Herausforderungen der vergangenen Jahre genannt, welche Grossbritannien teilweise gleich, teilweise aber auch anders als andere Länder betrafen (*polycrisis* und *declinism* vor dem Hintergrund von Finanzkrise, Flüchtlingskrise, Brexit, Klimakrise, Covidkrise, Energiekrise, Ukrainekrieg und Inflation).

**Zukünftige Rolle von Künstlicher Intelligenz bei Evaluationen**

Die Künstliche Intelligenz war ein wichtiges und in den einzelnen Veranstaltungen der Konferenz oft wiederkehrendes Thema. Der Fokus dabei lag bei Anwendungen basierend auf «Large Language Models» (LLMs), welche in jüngster Zeit in der Öffentlichkeit grosse Aufmerksamkeit erhalten haben. Dazu gehören etwa «ChatGPT» und «Claude». LLMs dienen im Wesentlichen dazu, Texte maschinell zu «verstehen» und zu erstellen.

Thematisiert wurden primär die Anwendungsmöglichkeiten und die entsprechenden Chancen und Risiken im Zusammenhang mit Evaluationen, aber auch die möglichen Folgen für die Zukunft der Evaluationspraxis insgesamt.[7] Dabei wurden teilweise bereits Resultate aus experimentellen Test-Anwendungen dieser Technologie im Evaluationsbereich vorgestellt, nicht jedoch darüber hinaus gehende Evaluationsresultate im engeren Sinn.

Zusammengefasst werden die Chancen darin gesehen, dass mit Hilfe von LLM-Anwendungen Textdaten radikal schneller und günstiger als bisher ausgewertet werden können (z.B. durch Zusammenfassung, Klassifizierung und Übersetzung von Textpassagen). Dazu kommt die Unterstützung beim Verfassen von Texten. Nebst den auch bei anderen Anwendungsbereichen von IKT auftretenden Datenschutz-Problemen erscheint das (schwerwiegende) Hauptrisiko im Moment zu sein, dass bei LLM-Anwendungen die genaue Verarbeitung der Informationen und je nach Anwendungsart auch die verwendeten Quellen eine «Black Box» bleiben. Experimentelle Test-Anwendungen können hier punktuelle Schlaglichter auf einzelne Probleme werfen, ohne die «Black Box» jedoch wirklich transparent zu machen.[8] Die Qualitätskontrolle von KI-generierten Resultaten bleibt oft schwierig oder unmöglich, solange unklar ist, wie KI zu ihren Resultaten kommt.

Von KI-Anwendungen werden teilweise weitreichende Folgen bis hin zu grösseren Umbrüchen für die Evaluationsbranche erwartet. Dazu gehört auch eine mögliche Konsolidierung der

[4] Vgl. dazu die Detailangaben zur Keynote-Präsentation «Five Ways for Evaluation to Develop Value Amidst Complexity» von Emily Gates im Abschnitt 2.

[5] So zum Beispiel in den Präsentationen von James Copestake (Bath University) und Barbara Schmidt-Abbey (Open University) vom 4. Oktober 2023.

[6] Beispiel: Plädoyer zugunsten des Einbezugs verschiedener Formen von Fachexpertise durch James Collis (Evaluation Task Force) in der Veranstaltung «What makes a Good Evaluation?» vom 4. Oktober 2023.

[7] Vgl. dazu auch die Detailangaben zu den Veranstaltungen «Large Language Model Applications for Evaluation» sowie «A Fireside Chat with Kelly Beaver» im Abschnitt 2. Eine weitere Veranstaltung, die ganz dieser Thematik gewidmet war, fand am 3. Oktober 2023 unter dem Titel «Encounters with Artificial Intelligences» statt.

[8] Beispiel: LLMs können zu unterschiedlichen (probabilistischen) Antworten auf mathematische Probleme kommen, wo es tatsächlich nur eine (deterministische) Lösung gibt.

Evaluationsbranche in Richtung grösserer und/oder innovativer Auftragnehmer mit dem Willen und der Fähigkeit, die neuen Technologien bei Evaluationen zu nutzen.

**Anwendung theoriebasierter Evaluationsansätze**

Das Konzept der theoriebasierten Evaluation (*theory-based evaluation*) und dessen Anwendung war eines der wiederkehrenden Themen der Konferenz. Trotz seiner häufigen Verwendung scheint der Begriff von der Evaluationsgemeinschaft nicht immer einheitlich verstanden und verwendet zu werden. Im bereits weiter oben thematisierten «Magenta Book» der britischen Regierung spielt die theoriebasierte Evaluation eine wichtige Rolle und wird wie folgt definiert (vgl. HM Treasury 2020a, S. 36):[9]

> Theory-based impact evaluations draw conclusions about an intervention's impact through rigorous testing of whether the causal chains thought to bring about change are supported by sufficiently strong evidence and that alternative explanations can be ruled out. Theory-based evaluation is explicitly concerned with both the extent of the change and why change occurs; it tries to get inside the black-box of what happens between inputs and outcomes, and how that is affected by wider contexts.

Die in dieser Definition erwähnte «hinreichend starke Evidenz» dürfte bei einzelnen Evaluationen gerade auch dank experimenteller oder zumindest quasi-experimenteller Methoden zustande kommen. An verschiedenen anderen Stellen im «Magenta Book» wird die theoriebasierte Evaluation jedoch explizit als eine Alternative zu experimentellen und quasi-experimentellen Evaluationsmethoden dargestellt. Punktuell wird immerhin auch von einer sinnvollen Kombination dieser Methoden gesprochen (vgl. HM Treasury 2020a, S. 36). Auch weitere in entsprechenden Publikationen zu findende Definitionen von theoriebasierter Evaluation beinhalten eine Abgrenzung zwischen theoriebasierter Evaluation und (quasi-)experimentellen Methoden.[10]

Im «Magenta Book» werden als einzelne Methoden theoriebasierter Evaluationen die folgenden Ansätze thematisiert (vgl. HM Treasury 2020a, S. 45): *Realist evaluation, contribution analysis, process tracing, Bayesian updating, contribution tracing, qualitative comparative analysis, outcome harvesting, most-significant change.*

Im Hinblick auf die Anwendung theoriebasierter Methoden und deren Kombination mit quasi-experimentellen Methoden war an der Konferenz die weiter oben bereits thematisierte Präsentation «Theory Based Evaluation Meets Econometrics» besonders interessant. Dabei wurde unter anderem auch die Anwendung der (quasi-experimentellen) Methodik der *difference-in-differences* (DiD) in die theoriebasierte Evaluationsmethodik integriert. Darüber hinaus blieb jedoch in verschiedenen Veranstaltungen der Konferenz die konkrete Vorgehensweise bei der Anwendung der Methodik der theoriebasierten Evaluation oft etwas vage.[11] In einer Veranstaltung zu theoriebasierter Evaluation wurde kritisiert, dass diese teilweise praktiziert werde ohne ein gemeinsames Verständnis dazu, was darunter genau zu verstehen sei.[12]

In mehreren der besuchten Veranstaltungen wurde die im «Magenta Book» als spezifische Methode theoriebasierter Evaluation aufgeführte «Qualitative Comparative Analysis» (QCA) thematisiert. In drei Präsentationen wurde QCA als angewendete Hauptmethodik vorgestellt.[13] Nachdem Charles Ragin bereits 1987 den Grundstein für QCA gelegt hatte, scheint sich diese

[9] Das «Magenta Book» widmet der theoriebasierten Evaluation einen Abschnitt im Leitfaden (vgl. HM Treasury 2020a, Abschnitt 3.4) sowie den ersten von fünf Anhängen (vgl. HM Treasury 2020b, Annex A.1).

[10] Vgl. Abstract von Devaux-Spatarakis (2023): «Theory-based evaluation was developed in response to the limitations of experimental and quasi-experimental approaches, which do not capture the mechanisms by which an intervention produces its impacts. This approach consists of opening the "black box" of public policy by breaking down the different stages of the causal chain linking the intervention to its final impacts. The hypotheses thus formulated on the mechanisms at play can then be tested empirically.»

[11] Ein Beispiel einer expliziten und gut verständlichen Beschreibung zu einer konkreten Anwendung des theoriebasierten Evaluationsansatzes findet sich in Umweltbundesamt (2020), Abschnitt 1.3, S. 28–31.

[12] Panel «Delivering Theory-Based Impact Evaluation» (4. Oktober 2023, Session 4).

[13] Präsentationen von Jonathan Cook und Ben Baruch am 2. Oktober 2023 (Session 5) sowie Präsentation von Saad Mufti am 3. Oktober 2023 (Session 5).

Methodik langsam und mit einiger Verzögerung im Bereich der Politikevaluation zu verbreiten. Auf kritische Aspekte der Methodik wurde in diesen Veranstaltungen kaum eingegangen. Die vom Autor dieses Konferenzberichts gestellte Frage nach der statistischen Signifikanz bzw. der Zufälligkeit von QCR-Resultaten konnte dabei nicht geklärt werden.

**Weitere wiederkehrende Themen der Konferenz**

Nachfolgend werden einzelne weitere Themen in Kurzform aufgeführt, welche jeweils in mehreren Veranstaltungen oder informell am Rande der Konferenz zur Sprache kamen.

- **Professionalisierung / Zertifizierung als weiterhin offenes Desideratum:** Die Problematik der (mangelnden) Professionalisierung und Zertifizierung der Evaluations-Disziplin bzw. einzelner Evaluierenden beschäftigt die Evaluationsgemeinschaft auch in Grossbritannien. Über grössere Fortschritte in diese Richtung wurde dabei nicht berichtet.
- **Evaluations-Guidelines der Regierung als methodischer Fokuspunkt:** Die Diskussionen in einzelnen Veranstaltungen legen den Schluss nahe, dass das britische Finanzministerium mit seiner erst drei Jahre alten «Central Government guidance on evaluation» (auch als «Magenta Book» bezeichnet) ein Referenzwerk geschaffen hat, an dem sich zumindest ein Teil der Evaluationsgemeinschaft in Grossbritannien orientiert.[14]
- **«Lernagenden» für Regierungsbehörden:** In den Vereinigten Staaten scheinen «learning agendas» und «learning plans» für einzelne Bundesbehörden an Bedeutung zu gewinnen. Auf deren Grundlage sollen gezielt Wissenslücken in Bezug auf Kernziele der einzelnen Behörden identifiziert und reduziert werden.
- **Suche nach «guten» Evaluationen:** Die Suche nach «guten» Evaluationen beschäftigt die Evaluationsgemeinschaft und deutet auf eine gewisse (Selbst-)Unsicherheit in Bezug auf die richtigen Qualitätskriterien und deren Umsetzung in der Praxis hin.
- **«Mind the gap» (I) – Anwendung fortgeschrittener Evaluationsmethoden:** Insbesondere in einzelnen informellen Gesprächen am Rand der Veranstaltungen kam zum Ausdruck, dass es (auch) in Grossbritannien weiterhin einen substantiellen Graben gibt zwischen den verfügbaren fortgeschrittenen Evaluationsmethoden und deren Anwendung, welche lückenhaft bleibt.
- **«Mind the gap» (II) – Verlässlichkeit der Resultate von Evaluationen:** Punktuell wurde scharfe Kritik geäussert zur Diskrepanz zwischen positiven Evaluationsresultaten (insbesondere bei ex-ante und begleitenden Evaluationen) und eher ernüchternden Erkenntnissen einige Jahre später. Ein Teilnehmender forderte, die Voraussagen und Resultate von Evaluationen sollten nach einigen Jahre systematisch überprüft werden. Der Intervenierende «lacht sich kaputt», wenn er einige Voraussagen und Beurteilungen aus Folgenabschätzungen und Evaluationen vergangener Jahre liest.
- **Verbesserung der Kommunikation gegenüber der Politik:** In verschiedenen Veranstaltungen wurde die Verbesserung der Kommunikation von Evaluationsresultaten gegenüber der Politik thematisiert. Die Diskussion ging in die Richtung, dass sich die Sprache der Evaluierenden noch weiter an die Bedürfnisse der Empfänger anpassen muss, ohne aber die Qualitätsansprüche der Evaluation zu sehr zu vernachlässigen. Es wurde festgestellt, dass Minister und generell Politiker eher an für sie nützlichen illustrativen Beispielen als an systematischer und umfassender Evaluationen interessiert sind.

[14] Vgl. Veranstaltung «Improving Evaluation in Government» (Abschnitt 2).

# 4 Fazit – Das Ganze ist mehr als die Summe der Einzelteile

Der **Mehrwert der Teilnahme** an der UKES-Konferenz aus Sicht eines Nicht-Mitglieds aus einem anderen Land bestand vor allem im Gesamteindruck der wiederkehrenden sowie einzelner neuer oder weniger bekannten Themen, welche die Evaluationsgemeinschaft in Grossbritannien aktuell beschäftigen. Aufgrund der schnellen Abfolge relativ kurzer Präsentationen (oft nur 25 Minuten inkl. Diskussion) mit teilweise längeren Folienpräsentationen in kleiner Schriftgrösse war es manchmal schwierig, die präsentierten Themen hinreichend zu erfassen. Ein individueller Zugang zu den einzelnen Präsentationen in elektronischer oder Papier-Form war nicht vorgesehen. Aufgrund verhältnismässig kurzer Kaffee- und Lunchpausen war die Zeit für informelle Gespräche am Rande der Konferenz kürzer als bei vergleichbaren anderen Veranstaltungen.

Die beiden Themenbereiche **transdisziplinäre bzw. innovative Evaluationsansätze** bildeten den Schwerpunkt der mitverfolgten Veranstaltungen. Sie repräsentierten zwei der vier Themenbereiche, nach denen die Organisatoren die Konferenz strukturierten. Auch aufgrund der begrenzten Möglichkeiten, sich während der Konferenz in einzelne vorgestellte Evaluationen zu vertiefen, blieb es danach allerdings schwierig, besonders gelungene Beispiele transdisziplinärer Zusammenarbeit und innovativer Evaluationsmethoden zu erkennen und zu würdigen.

Die beiden verbleibenden Themenbereiche im Zentrum der Konferenz betrafen **ethische Fragen und Anliegen sowie Lehre und Unterricht** zur Praxis der Evaluation. Vor allem im Zusammenhang mit ethischen Fragen und der diesbezüglich erhobenen Forderung nach «equality, diversity and inclusion» (EDI) findet offenbar aktuell eine Suche nach einem neuen Paradigma für die Evaluationspraxis statt. Die immer rascher aufeinander folgenden Krisen der vergangenen zehn bis fünfzehn Jahre (Finanzkrise, Flüchtlingskrise, Brexit, Klimakrise, Covid, Ukrainekrieg, Energiekrise, Inflation) beeinflussen auch die Diskussion über das Selbstverständnis der Evaluierenden. Die «Polycrisis» scheint in Grossbritannien dazu beizutragen, dass in der Evaluationsgemeinschaft nach Möglichkeiten gesucht wird, die Evaluation stärker auf ethische Kriterien sowie auf breitere und übergeordnete politische Ziele auszurichten. Dabei wird auch ein gewisses Spannungsfeld zwischen der Nähe zu den Entscheidungsträgern (im Zusammenhang mit Auftragsvergabe und Kommunikation der Resultate) und der Nähe zu den Betroffenen politischer Entscheidungen (inklusive benachteiligte Gruppen) bei der Evaluationstätigkeit sichtbar.

Vor diesem Hintergrund lässt sich der Eindruck gewinnen, dass in der britischen Evaluations-gemeinschaft und darüber hinaus gerade eine **vertiefte Auseinandersetzung um die Kultur der Evaluationspraxis** stattfindet. Auf der einen Seite steht dabei das über die vergangen Jahrzehnte verfolgte Ideal einer Politikevaluation, welche in der Tradition der Aufklärung für transparente Grundlagen für einzelne Entscheidungen in einem als grundsätzlich rational verstandenen politischen Prozess sorgen will. Auf der anderen Seite steht eine im kulturgeschichtlichen Sinn «romantische» Sichtweise der Rolle der Evaluation, welche in einem als komplex, widersprüchlich und nur bedingt rational empfundenen politischen Kontext zur Verwirklichung übergeordneter gesellschaftlicher Ideale beitragen soll. Hinter der ersten Sichtweise steht ein eher reduktion-istisches Paradigma, bei dem Evaluierende als neutrale Beobachtende mittels Wirkungsmodellen den politischen Prozess in seine Einzelteile zerlegen, bevor sie diese wieder zusammenfügen und dazu ihre Schlussfolgerungen ziehen (Analyse und Synthese). Die zweite Sichtweise steht für ein eher holistisches Paradigma, bei dem Evaluierende als Akteure mit eigenem ethischen Kompass mit allen verfügbaren Mitteln der Evaluation die Gesamtheit des evaluierten Bereichs sowie den weiteren gesellschaftlichen Kontext zu verstehen und zu verbessern versuchen.

Das Stattfinden und die Teilnahme an einer solch umfassenden und grundlegenden Selbstreflexion innerhalb einer der grossen Evaluationsgemeinschaften in Europa machte den «unique selling point» dieser Konferenz aus – mehr als die Präsentation einzelner Evaluationen und ihrer jeweiligen Methodik. Zumindest in diesem Sinn konnte somit der holistische Ansatz gegenüber dem reduktionistischen Ansatz bei der UKES-Konferenz in London ein «one–nil» erzielen.

## Quellenhinweise

### Publikationen


Devaux-Spatarakis, Agathe (2023), Theory-based Evaluation, in Policy Evaluation: Methods and Approaches, Éditions science et bien commun (Link)

Giel, Susanne (2013), Theoriebasierte Evaluation: Konzepte und methodische Umsetzungen, Münster: Waxmann, 2013

HM Treasury (2020a), Magenta Book, Central Government guidance on evaluation, March 2020 (Link)

HM Treasury (2020b), Magenta Book, Annex A, Analytical methods for use within an evaluation, March 2020 (Link)

National Audit Office (2021), Evaluating government spending, 2 December 2021 (Link)

National Audit Office (2022), Evaluating government spending: an audit framework, 12 April 2022 (Link)

Ragin, Charles (1987), The comparative method: moving beyond qualitative and quantitative strategies, Berkeley: University of California Press (Link)

Umweltbundesamt (2020), Evaluation des Nationalen Programms für Nachhaltigen Konsum, Texte 210/2020, November 2000 (Link)

Wolfram, Stephen (2023), «What Is ChatGPT Doing ... and Why Does It Work?», Stephen Wolfram Writings, 14 February 2023 (Link)

Weiss, Carol (1997), «Theory-based Evaluation: Past, Present, and Future», New directions for evaluation, 76: 41–55


### Internetseiten


BAKOM – Erste Evaluation der Leitlinien zur künstlichen Intelligenz
https://www.bakom.admin.ch/bakom/de/home/digital-und-internet/strategie-digitale-schweiz/datenpolitik/kileitlinien.html

MERL Tech – Natural Language Processing Community of Practice (NLP-CoP)
https://merltech.org/nlp-cop/

SBFI – Leitlinien Künstliche Intelligenz für den Bund
https://www.sbfi.admin.ch/sbfi/de/home/bfi-politik/bfi-2021-2024/transversale-themen/digitalisierung-bfi/kuenstliche-intelligenz.html

UK Cabinet Office / HM Treasury – Evaluation Task Force
https://www.gov.uk/government/organisations/evaluation-task-force

UK Evaluation Society (UKES)
https://www.evaluation.org.uk/

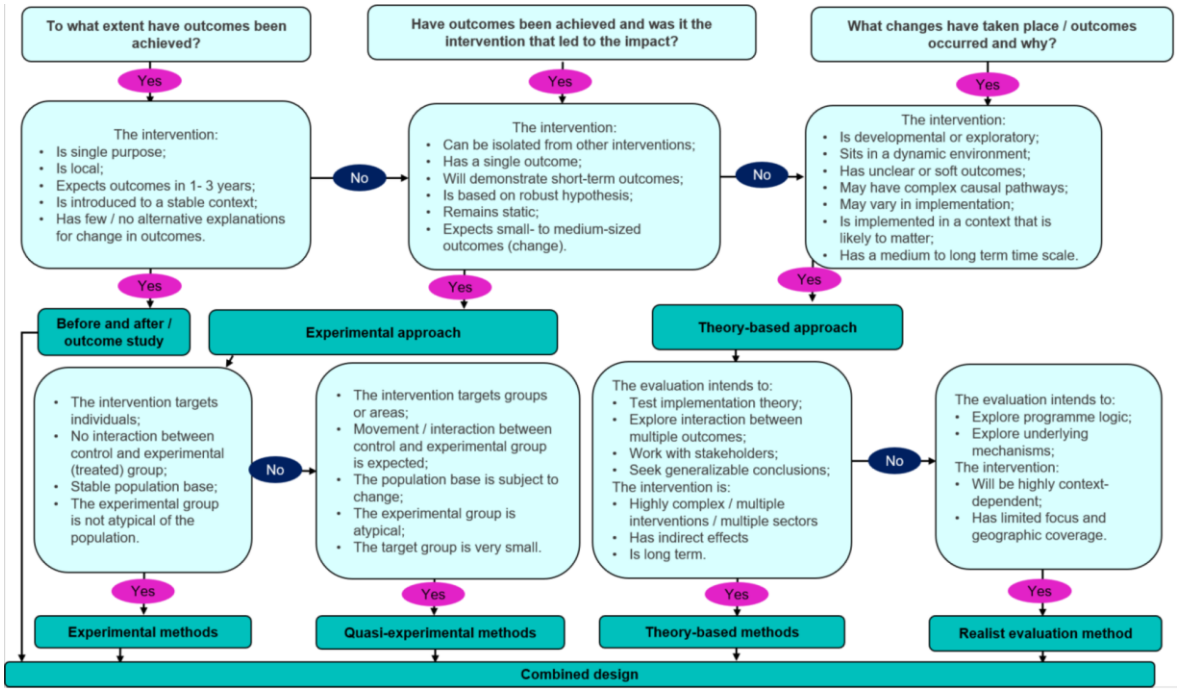
UK Evaluation Society (UKES) – Annual Conference 2023
https://www.evaluation.org.uk/conference-2023/

UK Evaluation Society (UKES) – Annual Conference 2023 – Programme
https://www.evaluation.org.uk/app/uploads/2023/09/RisingtoChallenges-2023-Agenda-Second-Release.pdf

## Anhang – Theoriebasierte und (quasi-)experimentelle Evaluationsansätze

*Auszüge aus «Magenta Book» zur Auswahl von Evaluationsansätzen sowie zum Prozess einer theoriebasierten Evaluation (HM Treasury 2020a, S. 33, S. 37)*

Figure 2.4. Selecting the approach for impact evaluation, based on the evaluation questions to be answered [1]




[1] Reproduced from: Hills, D. and Junge, K. (2010). *Guidance for transport impact evaluations: Choosing an evaluation approach to achieve better attribution.* [pdf]. The Tavistock Institute in consultation with AECOM. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/525806/transport-impact-evaluations.pdf [Accessed 5th November 2019]

Figure 2.5: Process followed using a theory-based approach