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.

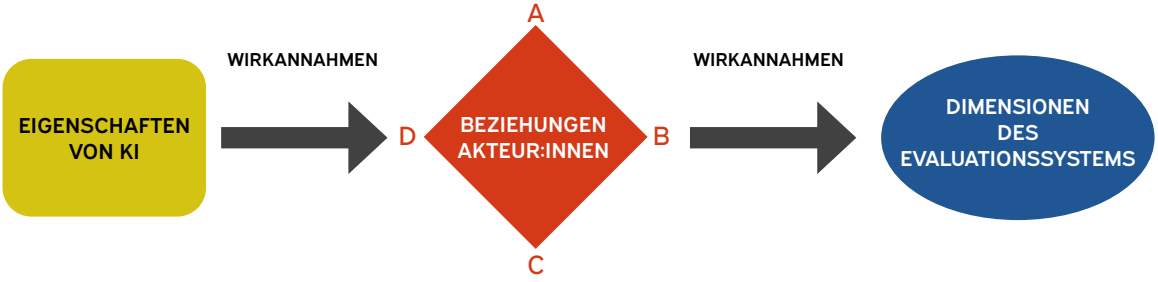

Abbildung 1: Schematische Darstellung des Wirkmodells von KI auf ein Evaluationssystem: KI hat Eigenschaften, die die Akteur:innenbeziehungen verändern, was wiederum Auswirkungen auf die unterschiedlichen Dimensionen des Evaluationssystems hat.

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.	Bei der komplexen Anwendung von KI kann sich der oben beschriebene Effekt weiter verstärken. Sind die Evaluierten den Evaluationen dann „ausgeliefert“, weil es beispielsweise keine Kontrollmaßnahmen von Seiten der Auftraggebenden gibt, kann dies schnell zu einem starken Sinken des Vertrauens in das Evaluationssystem führen. Unmittelbar könnten Evaluierte darauf reagieren, indem sie für KI-Methoden interessante Daten nicht erstellen, oder nicht verfügbar machen.

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 3 Zusammenwirken von KI, Beziehungen und System

In Abbildung 3 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.

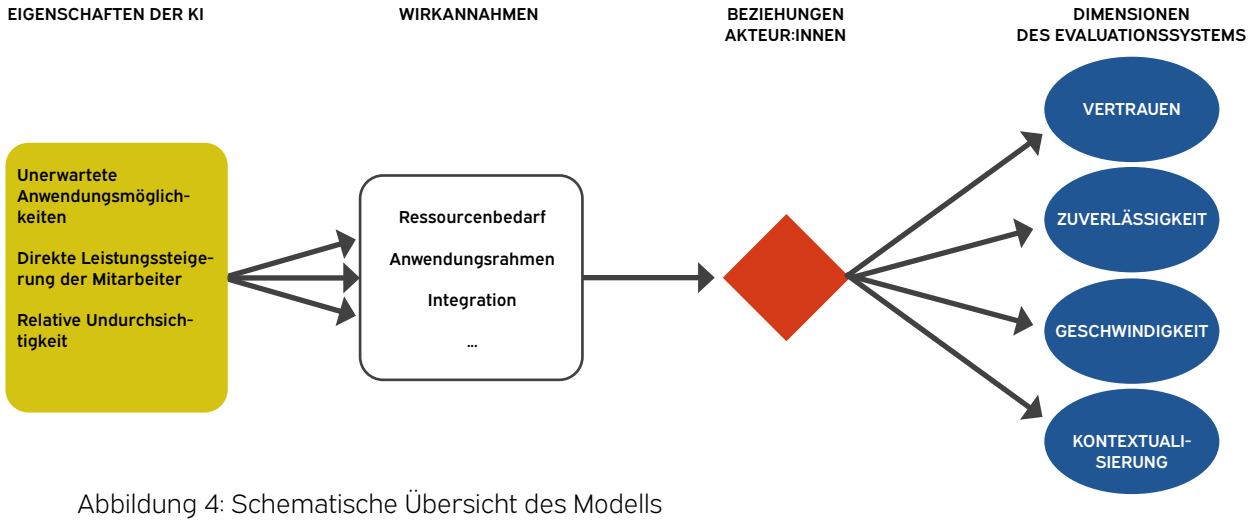

Abbildung 4: Schematische Übersicht des Modells

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 4 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.