fteval JOURNAL
for Research and
Technology Policy Evaluation

April 2025, Vol. 55, pp. e5, 1-31
DOI: 10.22163/fteval.2024.644

HOW DOES ARTIFICIAL INTELLIGENCE AFFECT THE EVALUATION SYSTEM?

IMPULSES TO SHAPE THE EVALUATION SYSTEM OF TOMORROW

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

ABSTRACT

This discussion paper presents initial reflections on potential changes brought about by artificial intelligence (AI) within an evaluation system. While previous debates on AI in this context have largely focused on data protection, ethics, and scientific or methodological reliability, this paper aims to introduce an additional, systemic perspective: it explores how the relationships between actors in an evaluation system may shift when especially generative AI enters the picture.

INTRODUCTION

In recent years, approaches based on Artificial Intelligence (AI) have rapidly expanded their capacity to produce human-like outputs and significantly increased public access to such technologies, most prominently through tools like OpenAI's ChatGPT (cf Dell'Acqua et al. 2023). As a result, strong narratives are currently emerging across various societal domains (from science to politics) shaping perceptions of AI. These narratives are difficult to ignore, regardless of the actual capabilities or applications of the technology.

Consequently, like many other systems, the field of evaluation is increasingly confronted with the need to develop constructive ways of engaging with this rapidly evolving technology. Generative AI is likely to have a profound impact on the scientific process—and, by extension, on many research-based data

collection methods used in evaluation (cf Van Noorden et al. 2023). This points to the general possibility of more substantial transformations in the years to come (Haupt et al. 2022; Chapinal-Heras and Díaz-Sánchez 2023; Stahl 2023; Konya and Nematzadeh 2024)

To date, public and academic debates on AI have focused largely on issues of data protection, ethics, and methodological reliability. With this discussion paper, we aim to broaden the discourse by introducing a systemic perspective. Specifically, we ask: How do the relationships between actors within an evaluation system change when the system is confronted with generative artificial intelligence[1]? (Hereafter, we will use "AI" as a shorthand.)

This question was the focus of a working group convened by the Austrian Platform for Research and Technology Policy Evaluation (fteval), which engaged with the topic—alongside its community between July 2023 and March 2024. Based on internal discussions and a targeted literature review, the present paper develops a conceptual model that can serve as a tool for reflection and dialogue among actors in the research, technology, and innovation (RTI) evaluation system.

The model is grounded in two core assumptions: (1) Generative AI, particularly in the form of large language models, is entering existing evaluation systems as a disruptive element; and (2) these tools have the potential to significantly reshape the relationships among actors within those systems. While our focus is primarily on the Austrian evaluation system, reflecting the practical experience of the working group, we believe the model may also be useful in other evaluation systems or sectors with comparable structures.

A CONCEPTUAL MODEL FOR REFLECTING ON AI IN EVALUATION SYSTEMS

To support a structured engagement with the potential impacts of AI on evaluation systems, a practice-oriented conceptual model was developed. It is designed to help actors and decision-makers reflect on changes in a step-by-step manner and to facilitate dialogue with other stakeholders. The model consists of three stages:

1 Generative (deep) artificial intelligence (AI) refers to algorithmic models trained to autonomously generate new data by approximating the probability distribution of a given dataset. In contrast to discriminative models, which are designed to predict class labels, generative models learn the underlying structure of the data and can use this to create entirely new instances, such as text, images, or audio (e.g., Ruthotto & Haber, 2021).

1. Users first define the attributes they assign to AI.
2. They then identify the relevant actors and the relationships between them.
3. Finally, they determine which dimensions of the evaluation system are particularly relevant in their specific context.

In this model, the characteristics ascribed to AI influence the relationships between actors, which in turn can affect various dimensions of the evaluation system. The introduction of AI may thus lead to both desirable and undesirable changes – initially at the level of relationships, and subsequently within the system itself and the processes it produces.

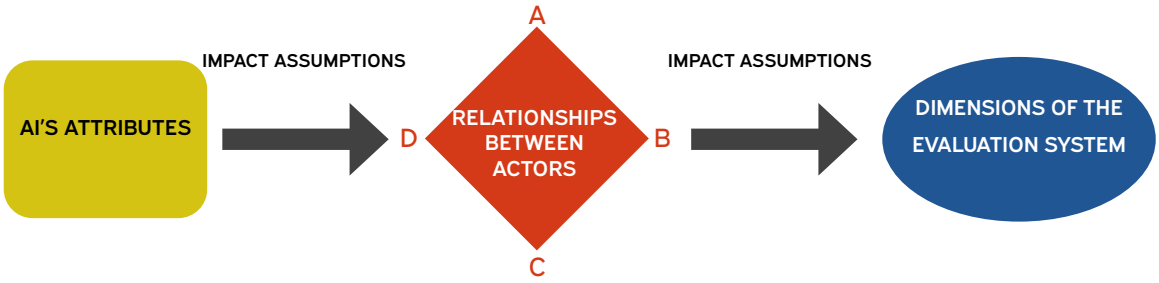

Figure 1: Schematic representation of how AI affects an evaluation system: AI's attributes shape actor relationships, which in turn influence key system dimensions.

In the following sections, we describe the key characteristics we have attributed to each component of the model and highlight central points of reflection that emerged in our discussions. Following the logic of the model, we proceed from left to right – beginning with the attributes of AI. A more detailed version of the model, including guiding questions, is provided in the appendix.

ATTRIBUTING CHARACTERISTICS TO AI

To make the discussion of AI's potential effects on evaluation systems more tangible, we developed a simplified, practice-oriented framing of AI's characteristics. For the purposes of our model, we focused on three core attributes that are intentionally abstract and general. This deliberate abstraction from specific applications – such as automated text generation – ensures that the model is not tied to the current state of technological performance. As such, it remains open to future developments and can be expanded to include more specific characteristics of AI where needed.

The selection of these characteristics is drawn on the work of Dell'Acqua et al. (2023). According to them, generative AI differs from classical machine learning in three important ways: (1) it enables surprising and unintended applications, (2) it has the potential to directly enhance individual performance, and (3) it exhibits a degree of opacity, as it can produce plausible but incorrect outputs. Together, these characteristics suggest that generative AI may reshape social and institutional relationships and significantly affect how work is organized. We now turn to a more detailed discussion of each of the three characteristics.

First, generative AI enables surprising applications for which it was not explicitly designed, and these uses may evolve quickly over time – either through improvements in model size and quality, or through shifts in user behavior. Although these systems are trained as general-purpose models, they often demonstrate specialized knowledge and capabilities during both training and everyday use (Singhal et al., 2022; Boiko et al., 2023). While there is ongoing debate at the technical level around the concept of "emergent capabilities" (Schaeffer et al., 2023), the effective capacities of AI systems are novel, unexpected, widely applicable, and increasing at a rapid pace. Recent studies show that AI can perform at a high level in professional domains ranging from medicine to law (Ali et al., 2023; Lee et al., 2023), thereby influencing those fields. Each new generation of AI models has shown substantial improvements, often giving rise to entirely new and unforeseen applications in real-world settings. As a result, both individuals and institutions may find these systems difficult to grasp or govern and hard to manage beyond setting general rules of engagement.

Second, AI can directly enable individuals to improve their domain specific performance without requiring technical expertise or major organizational support. Studies show measurable performance improvements when AI is used for writing tasks (Noy & Zhang, 2023), programming (Peng et al., 2023), and idea generation or creative work (Boussioux et al., 2024; Girotra et al., 2023). This suggests that generative AI could have a particular impact on professions dealing with complex tasks, offering individuals the ability to achieve noticeable performance gains independently of their employer or institutional context – gains they may choose to use for personal benefit or in service of the organization.

Third, generative AI is characterized by a relative opacity: it can produce outputs that are factually incorrect but appear plausible – commonly referred to as "hallucinations" or "fabrications." This opacity goes beyond technical or methodological issues. Even if the model's codebase is made transparent,

its outputs and interpretation would still not be fully explainable. As a result, AI may perform well in some contexts while failing unpredictably in others. This means that the effective and appropriate use of such systems cannot be fully defined in advance by their developers. Instead, it must be developed iteratively by users through trial and error. In practice, this often happens via peer exchange and the sharing of heuristics on community-based platforms such as, hackathons, Twitter threads, or YouTube channels. For organizations, this implies a fundamental uncertainty: they may never fully understand how an AI tool functions once it is integrated into their workflows and must learn to deal with this relative opacity in practice.

ANALYTICAL EXTENSION: POSITIONING AI AS AN ACTANT IN EVALUATION SYSTEMS

To analyze potential changes in evaluation systems, we understand generative artificial intelligence (AI) as an actant, drawing loosely on concepts from Actor-Network Theory (Latour, 2005). Central to this perspective is the idea that non-human entities, such as technologies, documents, or AI systems, can operate as effective elements within social networks. While they do not "act" in the traditional sense, they shape the actions of others through their characteristics, configurations, and contexts of use. Against this backdrop, we assign certain characteristics to AI that – without anthropomorphizing it – may influence the existing relationships between actors in evaluation systems. As an actant, AI can thus intervene in the system of relationships and affect the scope of action available to others.

From a technical standpoint, we define AI in the broadest sense as "the ability of a machine to imitate human capabilities such as reasoning, learning, planning, and creativity."[2] Many AI applications rely on statistical techniques, particularly machine learning. This discussion paper focuses specifically on generative AI, i.e. systems that can produce new texts, images, or other content – typically powered by large language models (LLMs), such as those used in ChatGPT, Gemini, Claude, or LLama.

Unlike other forms of AI (e.g. in image recognition or process optimization), generative models allow for direct and interactive use. They respond to input in real time, recombine information, and generate content that can be directly integrated into communication between actors. This makes generative AI

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

particularly relevant to our inquiry: as an interactive tool, it not only performs tasks but also has the capacity to shape communicative practices, interpretive patterns, and decision-making processes- thereby influencing relational dynamics in lasting ways.

The characteristics we assign to generative AI should therefore not be understood as a technical description of any specific tool, but rather as a conceptual generalization based on currently available systems. Our aim is to treat generative AI as a type whose role in evaluation systems can be meaningfully examined and discussed.

ASSUMED RELATIONSHIPS AND ACTORS

This discussion paper focuses on how AI affects the relational dynamics between different actors within the evaluation system. To reflect on how the emergence of generative AI might affect the evaluation system, we must first describe its existing structures and interactions. This description should be simplified enough to be workable yet sufficiently detailed to make potential changes imaginable.

At the core of the evaluation system considered here are four key groups: **clients**, **contractors**, **evaluators**, and **evaluatees/evaluation object**. These groups differ significantly in terms of formalization. Evaluators are usually individuals embedded in institutional contexts, but not fully representative of them. In contrast, clients and contractors are typically organizations that hold formal roles and mandates. The evaluation object itself can vary widely: from a single program to an entire institution. Often, it involves the collective performance of individuals operating within institutional structures.

This heterogeneity introduces a high degree of complexity into the relational structure of the evaluation system. To make this analyzable, a certain degree of conceptual simplification is necessary. Our model therefore relies on core assumptions that reduce the complexity of actor relationships without losing their essential meaning. Specifically, we assume the following simplified relationship pattern: In our simplified model, the evaluator has little direct contact with the client; communication typically runs through the contractor. Similarly, there is no direct relationship between the contractor and the evaluatees; rather, it is the evaluators themselves who interact with the evaluatees. This simplified network of relationships forms the basis of our model and is illustrated in Figure 2.

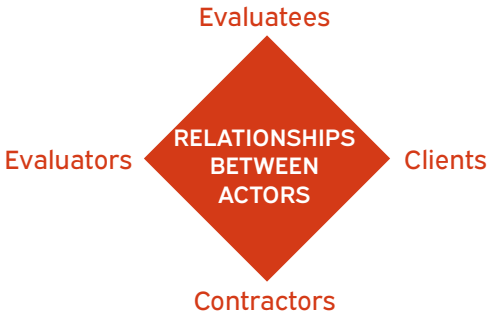

Figure 2: Schematic overview of the core actors and their relationships

With this structure in place, we can examine how generative AI enters the system as an actant and how it may reshape interactions and agency among the actors.

DESCRIPTION OF THE ACTORS IN THE ASSUMED EVALUATION SYSTEM

To analyze the potential effects of generative AI, it is necessary to identify the key actors in the assumed evaluation system. Our perspective is based on a simplified, but analytically useful, differentiation between four main actor groups, each with specific roles, perspectives, and scopes of action within the evaluation process:

Clients are the program owners or those responsible for an intervention – who are typically also the ones commissioning the evaluation. Clients need efficient and accurate evaluation results to make informed decisions. A clear need exists for reliable communication channels and an accessible interpretation of evidence. Beyond data quality, trust in the evaluation practice is crucial. Clients must be able to rely not only on the data itself but also on meaningful contextualization within the policy field, on thoughtful stakeholder engagement, and – at least to some extent – on the evaluator's ability to anticipate future developments and formulate scenarios or recommendations.

Contractors are the institutions responsible for managing and conducting evaluations. They depend on adequate resources and broad access to high-quality data. Their aim is to use staff and budgets efficiently—and ideally to expand them, for example by building expertise or generating revenue. Efficiency gains are therefore welcome, as long as the quality of work remains high. Reputation is key, since trust in the produced evidence and the reliability of methods are central expectations in this role.

Evaluators are the individuals who carry out the evaluation itself. They collect, prepare, and analyze data, and provide recommendations in response to evaluation questions. They are accountable both to the client and the contractor (typically their employer), while also holding their own expectations for how the work should be done. These may include career ambitions, team dynamics, thematic interests, or opportunities for personal and professional development. Given their expertise and contextual knowledge, evaluators often have considerable autonomy and responsibility despite their formal dependency on the employer, and can thus serve as important starting points for change.

Evaluatees or **owners of the evaluation object** may include, depending on the context, the intervention or program itself, a particular institution or organizational unit, or the performance of one or more individuals Evaluatees need a clear understanding of how the evaluation is conducted, along with trust in the underlying methods and their appropriate application. Transparency and communication are essential to ensuring the trust of those affected by evaluation.

Actor Group	Typical Role	Degree of Institutionalization	Typical Interests	Points of Contact with AI
Clients	Defining objectives, funding, and overseeing the evaluation process	High (organizations with clear mandates)	Relevant findings, strategic knowledge, legitimacy	Analytical support, decision-making templates based on AI outputs etc.
Contractors	Fulfilling formal contracts, managing resources, and interfacing with evaluators	Medium to high (organizations with evaluation expertise)	High-quality execution, project success, reputation	Integration of AI into workflows, coordination of AI usage etc.
Evaluators	Collecting data, analyzing and interpreting results, and reporting findings	Low to medium (mostly individuals or small teams)	Professional integrity, methodological quality, independence	Use of AI for text generation, data analysis, and synthesis etc.
Evaluatees	Providing data, interacting with evaluators, and being the subject of the evaluation	Highly variable (individuals, projects, or institutions)	Fair representation, influence on evaluation, transparency	Interacting with AI-supported data collection tools or assessments

Table 3: Overview of the actors in an evaluation system and their key characteristics

KEY DIMENSIONS WITHIN THE ASSUMED EVALUATION SYSTEM

We understand an evaluation system as a network of relationships characterized by recurring interactions between familiar actors, which has become stabilized over time. The established roles and patterns of interaction form the basis of an "evaluation culture" that shapes the evaluation system. For the purposes of this discussion paper, we adopt a simplified view of a self-contained system, considering only those actors explicitly included in our model. However, users of the model are free to adapt, expand, or reduce the set of actor groups as needed.

In any case, the evaluation system is defined by the totality of actors, their activities, and their interactions. In our model, we assume at least four characteristic dimensions of the evaluation system that affect all actors and shape their respective scopes of action:

- Dimension 1: A strong emphasis on **trust** in the processes that generate evidence, along with high expectations for responsible handling of data and results
- Dimension 2: **Reliability** of knowledge production and reproducibility of findings
- Dimension 3: **Speed** and (cost-)efficiency in the production of evidence to enable timely decision-making
- Dimension 4: The need for hard to objectify **context-specific** knowledge and sector expertise – elements that have always represented a kind of "black box" within evaluation systems

As outlined in the model, we assume that the use of AI affects one or more actor groups and their relationships, which in turn may influence one or more of these (or other) dimensions of the evaluation system.

AI AS A DISRUPTIVE ELEMENT IN EVALUATION SYSTEMS

Taken together, the AI characteristics defined in the model carry significant disruptive potential for established systems. Dell'Acqua et al. (2023) refer to this as a "jagged technological frontier", where the capabilities of AI extend unevenly across different domains of knowledge work. In some areas, AI may outperform human abilities, while in others it falls far short of expectations – often in unpredictable ways.

This asymmetrical performance profile of AI creates new scopes for action for individual users or organizations. Those who navigate this frontier skillfully may gain clear advantages – for example, by accelerating analysis or producing written content more efficiently. However, such advantages may come with systemic side effects: if they remain opaque or cannot be replicated by others, they may foster mistrust, uncertainty, and imbalances of power. Likewise, careless or inappropriate use of AI can undermine the quality of both relationships and evaluative judgments.

In this context, we understand AI, as previously described, as an actant that enters an existing relational system and reshapes it through its characteristics. The resulting effects are not the product of intentional behavior but rather emerge from how AI is used as a tool, how it is interpreted, and how it becomes embedded in social processes.

Evaluation systems are particularly sensitive in this regard. They rely heavily on established interaction patterns, trust, and reputation. At the same time, they face pressure to increase efficiency and embrace new technological possibilities. This tension between the need for stability and the pressure to innovate makes evaluation systems especially vulnerable to the disruptive effects of AI – which makes it even more important to examine these effects closely.

A MODEL FOR UNDERSTANDING HOW AI ENTERS THE EVALUATION SYSTEM

In this model, the three previously defined characteristics of AI, (1) unexpected application possibilities, (2) direct performance enhancement for staff, and (3) relative opacity, are positioned in relation to the four key dimensions of the evaluation system, (1) trust, (2) reliability, (3) speed, and (4) contextualization. Building on this framework, the previously described actors, along with their action and response logics in dealing with AI, can be integrated into an overall model of the evaluation system. This model is intended as a tool for reflection, enabling users to identify relevant questions, knowledge gaps, and areas for strategic development.

In our analysis, AI acts as an actant whose disruptive impact depends on the extent to which its characteristics are expressed. This degree is shaped not only by the AI's inherent properties and capabilities but also by how it is applied or permitted to be applied. As a result, the same AI tools may be introduced for different purposes and to varying extents. Therefore, the

disruptive potential of AI should not be viewed as uniform, but rather as scenario dependent. To reflect this variability, we define two ideal types of how AI may function as an actant in the system – one with limited, one with greater disruptive impact:

1. **"Simple AI application"**: AI tools are used to address straightforward tasks and to solve familiar problems more efficiently.
2. **"Complex AI application"**: AI models are used to address complex challenges or questions previously beyond scope, enabling entirely new approaches to problem-solving.

The terms "simple" and "complex" do not refer to the abilities of the technology itself. In both scenarios, the same technology may be used. The difference lies in the demands placed on its use. In simple applications, AI can support routine tasks aimed at improving efficiency, and experienced evaluators can reasonably assess the reliability of the outcomes. In contrast, complex applications deal with questions that could previously only be addressed with significant effort – or not at all. In such cases, the ability to verify reliability may be limited.

AI'S EFFECTS ON ACTORS IN THE EVALUATION SYSTEM AND THEIR MODES OF REACTION

Because the actors within the evaluation system are interrelated, they may be affected by AI even if they do not use it themselves or are not allowed to. This creates a situation in which all actors, regardless of their position on AI, are required to engage with the topic and develop informed perspectives or strategies. A constructive process requires shared understanding of which changes and adaptations are necessary, desirable, or undesirable.

To support this process, the next section presents illustrative assumptions – based on the conceptual model and previously defined AI characteristics – about how different actor groups might be affected. The focus here is on identifying potential response strategies for each group of actors. No normative judgement is made about whether a more "defensive" or more "proactive" approach is preferable.

Each subsection begins by stating the assumptions made. Specifically, which dimension(s) of the evaluation system, which characteristic of AI, and which relationship between actors were considered particularly relevant and

therefore informed the analysis. The result is a set of tables illustrating possible positive and negative impacts (for both simple and complex AI applications), along with strategic options for each actor group. These are not definitive conclusions, but suggestions intended to spark discussion. Users may arrive at different conclusions based on different assumptions – the role of the model is to make these assumptions explicit and structured, which is essential for a solution-oriented debate.

CLIENTS

For policy makers or intervention owners, the challenge lies in the fact that they have little direct control over which actors use AI, where it is applied, and how it is used. At the same time, they are particularly dependent on maintaining high levels of trust in the evaluation system, as evaluation results serve as the basis for decision-making (e.g. whether to continue or discontinue a program). This trust is especially important in the relationship between clients and evaluatees, as these two groups often work together over extended periods of time. In this context, the inappropriate use of AI poses a significant risk to that relationship. Clients may therefore feel compelled to introduce rules or controls to mitigate this risk. However, such measures may lead to increased costs, limit the developmental potential of new technologies, or trigger negative reactions from other actors in the system – such as evaluators.

Effects of AI's entry into the evaluation system on the actor group "Clients"			
		POSITIVE	**NEGATIVE**
Disruptive potential	**„simple" application of AI**	The **direct performance enhancement** of individuals – particularly evaluators – can lead to faster delivery of evaluation results. In addition, larger volumes of data can be processed more efficiently, which may ultimately result in cost savings across the evaluation system.	Even in simple applications, AI's **relative opacity** can lead to a loss of trust. As data volumes grow, it becomes increasingly difficult for clients and evaluatees to understand how results are generated – even when AI is used in straightforward ways. This may prompt clients to intervene in evaluation practices by introducing formal rules or controls. However, such interventions could have unintended consequences for the long-term adoption of AI in the evaluation system.
	„complex" application of AI	AI makes it possible to better understand and assess aspects that were previously difficult to grasp. This allows programs to be examined from entirely new, but relevant perspectives, potentially increasing their overall impact. To make use of this potential, clients must either develop a solid understanding of how AI can be applied, including seeking out **unexpected applications** – or rely on contractors to identify and share such insights with them.	The combination of **unexpected application possibilities** and **relative opacity** may compromise the integrity of AI-generated outputs, making them unsuitable as a basis for decision-making by clients. To use these results safely and effectively, clients would need to build substantial internal capacity for interpretation and contextualization.

Strategic approaches of the actor group "Clients" in responding to AI	
OFFENSIVE	**DEFENSIVE**
Clients actively create incentives to encourage internal engagement with AI and to build in-house capacities for its application. They also establish forums for dialogue with evaluatees to exchange perspectives on the opportunities and limitations of AI, and to identify particularly relevant use cases for specific programs. In doing so, clients take on a shaping role—for example, by defining success indicators and goals for the use of AI in the evaluation system. These are then monitored over time, in collaboration with other actors in the field, to detect emerging challenges at an early stage. Clients advocate for the use of AI and provide transparent options for action in cases where its application does not produce the desired outcomes.	Clients take a passive role in the introduction of AI into the evaluation system, leaving it to other actor groups to bring the technology into active use – if trust in the system is not undermined as a result. To ensure this, an ongoing exchange with other actors is necessary, particularly to maintain an understanding of the current state of trust within the evaluation system. While clients do not adopt a proactive, shaping role, they do provide regular training for their staff on known applications of AI.

CONTRACTORS

Contractors occupy a unique position: they are in competition with other contractors, must meet the requirements set by clients, and may also enter a new dependency dynamic with their employees in the context of AI. Two system dimensions are particularly relevant for their role: **speed** (efficient evidence production) and **contextualization** (access to expertise and domain knowledge). In many cases, employees can use AI models more quickly than their organizations, as these tools are freely available and can be used independently of the employer. This dynamic may increase the contractors' dependence on highly skilled staff. At the same time, contractors can build structural capacities for AI use – for example, by training their own models or preparing datasets tailored to specific evaluation contexts. Such tasks require resource investments that individuals alone cannot manage. In the competitive landscape among contractors, the ability to effectively integrate AI into workflows and to leverage resulting efficiency gains is likely to be essential. Access to high-quality data and the ability to interpret AI-generated results will be key to success in the evaluation market and may foster competition rather than cooperation at this level.

Effects of AI's entry into the evaluation system on the actor group "**Contractors**"			
		POSITIVE	**NEGATIVE**
Disruptive potential	**„simple" application of AI**	The **direct performance gains** enabled by AI – for example, in report writing -allow for more efficient use of available resources. These resources can either be redirected to other parts of the evaluation or invested in building additional capacities and skills, thereby contributing to a lasting improvement in evaluation quality.	When **performance gains** from AI are not shared with the organization but used by individual employees for their own advantage, contractors may begin to distrust their staff. This can lead to the introduction of new monitoring measures. Under such conditions, ensuring quality becomes increasingly difficult – even though quality assurance is especially important due to AI's **relative opacity**. In some cases, the organization may also be forced to mediate between conflicting interests among employees, which can tie up additional resources.
	„complex" application of AI	The complex application of AI opens opportunities to work with high-quality data for sophisticated analyses. As domain expertise is increasingly supported – or even partially replaced – by AI, new professional roles may emerge that are characterized by cross-cutting skill sets. This facilitates collaboration across different areas within and beyond an organization and enables new approaches and forms of evaluation.	Adapting to the use of complex AI models may require significant restructuring within an organization. New professional roles may entail changes to employment contracts or require new training programs. In addition, organizations must invest in technical infrastructure and expertise – often without certainty about the actual long-term benefits. It also remains unclear whether employees will use **unexpected application possibilities** – like direct performance gains – for their own advantage. This creates a heightened risk for organizations that their investment may not pay off.

Strategic approaches of the actor group "**Contractors**" in responding to AI	
OFFENSIVE	**DEFENSIVE**
AI is seen as an opportunity to actively experiment with new methods and approaches, and employees are provided with an environment that supports such exploration. The focus of these efforts is on developing better solutions for existing problems and identifying newly accessible solutions for issues that were previously out of reach. Contractors rely on strong relationships with experts and therefore implement trust-building measures to reduce the stress that these developments may cause among staff – and to prevent resistance. Resources are invested to build internal expertise, and the trust-building efforts are designed to actively encourage employees to share new ways to apply AI within the organization.	Contractors largely leave the application and handling of AI to their employees, providing only minimal guidance or restrictions. Active investments in this area are kept limited, with a continued emphasis on "traditional" forms of evaluation. This approach is also positioned as a distinguishing feature in the competitive landscape – intended to build trust among clients and evaluatees in the methods used, and among employees in the organization itself. As a result, some contracts may even include explicit prohibitions on the use of AI by staff.

EVALUATORS

Evaluators can use AI to independently streamline their daily work and delegate tasks to these systems. The better they are at navigating AI, the greater the potential benefits, making it likely that they will take a strong interest in identifying **unexpected application** cases. However, this also places a significant amount of responsibility on them. If no institutional structures are in place, evaluators must themselves deal with the **relative opacity** of AI and decide whether AI-generated outputs meet the standards of the evaluation system – even in situations where such standards and practices may not yet exist. An unstructured adoption of AI led by evaluators may result in them becoming key decision-makers regarding the **reliability** and **trustworthiness** of the system, which could lead to them being overwhelmed by this responsibility. A strong relationship between contractors and evaluators, and the development of appropriate support structures could help mitigate this shift in responsibility. At the same time, the personal interests of evaluators must be considered. For instance, they may choose not to share efficiency gains with their employers to maintain relative advantages within their team or to advance their own career paths. Conversely, some evaluators may perceive the use of AI tools as a threat to their job security, potentially leading to tensions with employers who are enthusiastic about the technology.

Effects of AI's entry into the evaluation system on the actor group "**Evaluators**"			
		POSITIVE	**NEGATIVE**
Disruptive potential	**„simple" application of AI**	The **direct performance gains** enabled by AI can relieve evaluators of routine tasks through automation. The nature of this technology also enables them to implement and tailor it to their specific needs. This creates the opportunity to focus more on engaging aspects of their work, such as conducting complex analyses and generating deeper interpretations.	**Direct performance gains** are only accessible to individuals who actively engage with the technology, which may lead to "generational conflicts" within teams or organizations. Individuals seeking to use AI for more efficient workflows may encounter resistance from colleagues who reject the technology, slowing progress for the former and creating a sense of threat for the latter.
	„complex" application of AI	The complex application of AI can enable evaluators to engage in tasks that go well beyond their core expertise, allowing them to continuously expand their sector-specific knowledge. Ongoing exchanges with colleagues can help mitigate challenges such as AI's **relative opacity** and promote the broader sharing of **unexpected application** cases. This may foster a new understanding of how work is organized – one that offers greater autonomy and new possibilities.	The complex application of AI can lead to overload, making it more likely that AI's **relative opacity** results in hard-to-detect errors. If evaluators are unable to exchange experiences (e.g. due to organizational AI bans) or unwilling to do so (e.g. because of personal advantages), this corrective mechanism is lost. The lack of clarity over who holds responsibility for introducing and using AI-supported analyses results in either no use at all or covert application.

Strategic approaches of the actor group "**Evaluators**" in responding to AI	
OFFENSIVE	**DEFENSIVE**
Evaluators become drivers and shapers of AI application. They actively contribute to the development of both codified guidelines and informal norms based on real-world institutional and system-level practices. As a result, routine tasks can be reliably delegated to AI, freeing up time for more demanding work. Their personal interest and capacity for shaping AI use foster a culture in which regular training and updates on AI developments are well received, and peer exchange is actively maintained.	Evaluators come under pressure in a competition for efficiency and feel compelled to use AI as a shortcut in ad-hoc and unreflective ways to cope with these challenges. AI is applied only in areas where increasing pressure leaves no alternative. Exchange between individuals is hindered not only by time constraints but also by the absence of official AI deployment: those who benefit from using AI may avoid sharing their practices for fear of violating internal rules or losing access. Under these conditions, there is little willingness to participate in training opportunities, further compromising the quality of evaluation.

EVALUATEES

For evaluatees, both **expected and unexpected** applications of AI may give rise to new forms of evaluation that allow for a more nuanced and comprehensive assessment of their activities. This, in turn, can increase **trust** in the evaluation system by presenting program achievements in a more differentiated way. However, evaluatees must be convinced of the **reliability** of these new forms of evaluation. At the same time, AI can enable the internal development of robust, novel monitoring systems that support self-reflection and provide a solid foundation for future evaluations. Such systems could also be used by evaluatees to create counter-narratives to external evaluations, potentially complicating the relationship between evaluatees and evaluators. This may undermine **trust** in the evaluation system, especially if clients are uncertain about which sources of evidence to rely on. To avoid deepening such divides, a careful balance must be struck between the interests of evaluatees and clients – one that supports constructive collaboration with contractors and evaluators.

Effects of AI's entry into the evaluation system on the actor group "**Evaluatees**"			
		POSITIVE	**NEGATIVE**
Disruptive potential	**„simple" application of AI**	In addition to faster availability of results, AI-driven **performance gains** can help evaluatees to translate evaluation findings into more actionable formats, thereby facilitating their quicker integration into new practices.	The **relative opacity** of AI means that evaluatees may even in cases of simple AI applications no longer be able to understand how evaluation results were produced. This effect can be amplified by the pseudo-objective language often used by AI systems. To navigate this situation, evaluatees are forced to build their own internal capacities, diverting resources away from their core activities without necessarily generating additional value.
	„complex" application of AI	The interplay between internal monitoring and external evaluation – mediated by AI – opens up entirely new forms of collaboration between evaluatees and evaluators. Both sides can benefit from each other's as-yet-undiscovered or **unexpected applications**, while jointly working to reduce the risks associated with AI's **relative opacity**.	In the case of complex AI applications, the effect described above may intensify. If evaluatees find themselves at the "mercy" of evaluations – due to a lack of oversight or safeguards from the clients – trust in the evaluation system may rapidly decline. As an immediate response, evaluatees might withhold or stop generating data that would be of particular interest for AI-based methods.

Strategic approaches of the actor group "**Evaluatees**" in responding to AI	
OFFENSIVE	**DEFENSIVE**
Evaluatees use AI both to enhance their own capacity for reflection and learning, thereby building an understanding of the technology's potential applications, and to engage constructively with external evaluators who integrate AI into the evaluation. Thereby, they actively help shape the conditions for its use. A participatory and transparent process has been developed to incorporate AI into the evaluation process. There is a culture of open communication about how AI is used and how its outputs are interpreted and applied.	Evaluatees use AI partly for internal monitoring but do not share these results with external evaluators. This reduces their demand for external evaluation as a source of learning but also risks institutional blind spots or increased dependence on AI-savvy staff. In cases of external AI use, the processes remain vague and poorly defined. As a result, complex AI-based evaluations are less likely to occur externally, and no visible, constructive approach to the technology is developed. This passive stance may lead to misunderstandings about AI use and can decrease overall openness within the organization. If clients were to insist on an AI-based evaluation, the organization might be forced to quickly establish processes that could be prone to error.

The considerations presented here are neither exhaustive nor final. However, they illustrate the wide range of potential reflections across different actor groups and highlight the value of breaking them down into simpler components. Our model provides a framework for doing so in a structured manner. Building on the strategies outlined here, a possible next step would be to explore where strategic overlaps exist between actors and where their interests regarding the use of AI may diverge significantly. This creates a solid foundation for initiating a discourse on shared solutions.

KEY QUESTIONS

As previously outlined, our model can be used to identify system-relevant questions about the use of AI in evaluation contexts. Throughout our exploration, one perspective emerged as particularly central: viewing the evaluation system as a social structure shaped by stable relationships. The entry of AI as an *actant* changes this structure, often subtly and not immediately visible to all actors.

One characteristic of generative AI proved especially relevant: its tendency to enable *unexpected applications*. Unlike the often-discussed opacity or performance gains, this potential has received relatively little attention – despite having significant implications at the individual level. Creative or unintended uses may shift agency from the institutional to the individual level, creating new opportunities but also risks for trust, transparency, and

fairness. Evaluators, for example, might conceal their discoveries of *unexpected applications* of AI to gain efficiency advantages, especially in environments where AI is viewed with skepticism. At the same time, they would greatly benefit from peer exchange, which requires organizational support and clear frameworks from contractors. Principles for AI use should therefore account for the importance of experimentation while also ensuring transparency within the social system – beyond purely technical perspectives. This leads to a double challenge for contractors: they are expected to harness the potential of AI, yet are reliant on informal use by their staff, while bearing institutional responsibility. A lack of shared norms and principles around AI could, in the long term, destabilize the entire evaluation system.

This example illustrates how technological innovation can (unintentionally or not) reshape social relationships. Based on this, the fteval community should consider the following key questions:

1. What might a desirable AI-infused evaluation system look like?
2. At what level is competition over unexpected applications (or AI in general) desirable and innovation-enhancing – and at what level is it harmful?
3. Who should ultimately benefit from AI applications in the evaluation system? Evaluators? Contractors? Clients? Evaluatees? How can this be made explicit and fair?
4. What is required to foster trust-based relationships between evaluators and contractors – or across actor groups more generally?
5. How can exchange on emerging AI applications be organized across the evaluation system?
6. What conditions must contractors establish to encourage evaluators to openly share the unexpected AI use cases they discover?
7. How can institutional actors benefit from the experimental spirit of their staff without compromising trust in the results?
8. How should responsibility be structured so that risks to the system or individual actors are minimized – without undermining the willingness to experiment?
9. How can covert exploitation of new applications be reduced while still rewarding individuals and institutions for their engagement?

10. What should knowledge sharing about new AI practices in evaluation look like—and how can that goal be achieved?

11. How does AI change the way individuals perceive feedback they seemingly receive from colleagues or supervisors?

All these questions have ethical and legal implications, but extend far beyond them. If actors in the evaluation system fail to develop a trust-based approach to working with this new technology, negative externalities for the system are likely to increase. As a result, confidence in evaluation results may decline, and clients may be less inclined to use such evaluations as a basis for decision-making. Addressing these questions successfully will be a key challenge for the years ahead and one that has yet to be fully explored.

LOOKING AHEAD: WHAT SHOULD BE TAKEN INTO ACCOUNT NOW?

This discussion paper aims to encourage a structured reflection on how AI may affect relationships within the evaluation system. All of us will be affected by developments in this field. At the same time, generative AI inherently turns us into active agents of this development. This dual role holds both great potential and considerable risk for the evaluation system. From our perspective, the overarching question in the coming years will be: Which relationships between actors should remain as they are, and which ones should evolve in the interest of a better evaluation system?

AI should not be seen merely as a technological development, but also as an expression of a broader societal will for transformation. While it may still be possible to resist the adoption of AI as a technology for some time – perhaps for years – AI as a social force will increasingly find its way into diverse systems, often with considerable momentum and regardless of the consequences. This combination of rapid technological change and social momentum makes it necessary to formulate a future vision for evaluation – independent of AI's current technical capabilities and regardless of one's personal stance toward the technology. Otherwise, even fundamentally "positive" features of AI may unintentionally disrupt the relationships within a system and trigger undesired forms of disruption.

This work, in the form of a schematic model accompanied by guiding reflection questions, provides a foundation for a structured discussion about the

integration of AI into the field of evaluation. For every actor in the system, it is crucial to engage early with potential changes and to actively help shape a future-proof evaluation practice. There is no central authority that oversees the evaluation system as a whole. Instead, it emerges from the sum of interdependent actors who influence one another – knowingly or unknowingly. This makes conscious engagement with these dynamics even more essential. We therefore invite all interested and affected parties to join the conversation and to collaboratively set the course for an effective and ethical evaluation practice in the age of AI.

ACKNOWLEDGEMENTS

This work is based on conceptual reflections by Thomas Palfinger, Open Innovation in Science Center at the Ludwig Boltzmann Society (LBG) and Susanne Beck, formerly of the LBG, now affiliated with Warwick Business School, University of Warwick. These ideas were further developed in the context of evaluation within the working group on Artificial Intelligence in Evaluation, hosted by the Austrian Platform for Research and Technology Policy Evaluation (fteval). A sub-group focusing on changing actor relationships, active from July 2023 to May 2024, included: 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 Center), Vitaliy Soloviy (AIT), Michael Strassnig (WWTF), and Isabella Wagner (fteval). The broader working group additionally included representatives from AIT, aws, FFG, FWF, KMU Forschung Austria, ÖAWI, Technopolis Group | Austria, WIFO, and ZSI. Sincere thanks to all for their ideas, contributions, and valuable feedback!

REFERENCES

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

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

APPENDIX – THINKING IN PROGRESS

DETAILED PRESENTATION AND GUIDING QUESTIONS OF THE MODEL

To simplify matters, our model assumes a one-sided influence of AI on the evaluation system and the relationships within it. In this model, AI is conceptualized as a new actant entering an existing system – potentially triggering changes or disruptions in established relationships. We focus specifically on how AI affects the relationships between actors and based on these, how it influences certain aspects of the evaluation system.

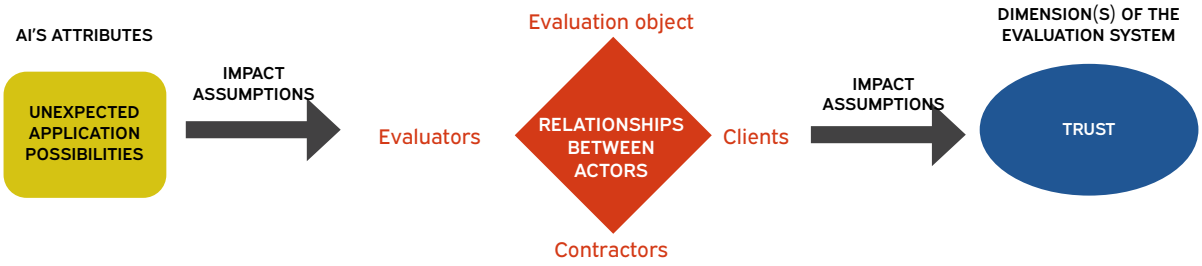

Figure 3: The Interplay between AI, Relationships, and the Evaluation System

Figure 3 illustrates the basic mechanisms of the model. Starting from a defined characteristic of AI (in the illustrated example: unexpected application possibilities), we formulate assumptions about how this characteristic might affect selected actors. We then examine how these assumptions could influence their mutual relationship. Based on this, we can determine potential effects on one or more of the four dimensions of the evaluation system described in the respective chapter (in this case: trust). This enables a step-by-step understanding of how specific AI characteristics might influence the system upon entering it. While applying the model, we continuously asked ourselves the following questions, which have guided our entire engagement with AI:

1. Do we assume that a certain characteristic of AI influences the existing relational practices between relevant actors in a system?
2. What assumptions can be made about the impact of AI, and which of these seem particularly relevant?
3. How do these assumptions affect the existing relational structures between actors?

4. What are the consequences of these changes in relationships for one or more dimensions of the evaluation system?

5. Are these consequences desirable from our perspective? What would need to be in place to make them desirable?

These questions guide users step by step from left to right through the model, helping to focus on the key aspect at each stage. The fifth question already leads beyond the model itself and into broader discourse – encouraging reflection and, as we did in the discussion paper ourselves, the formulation of strategic positions and visions for a preferred future.

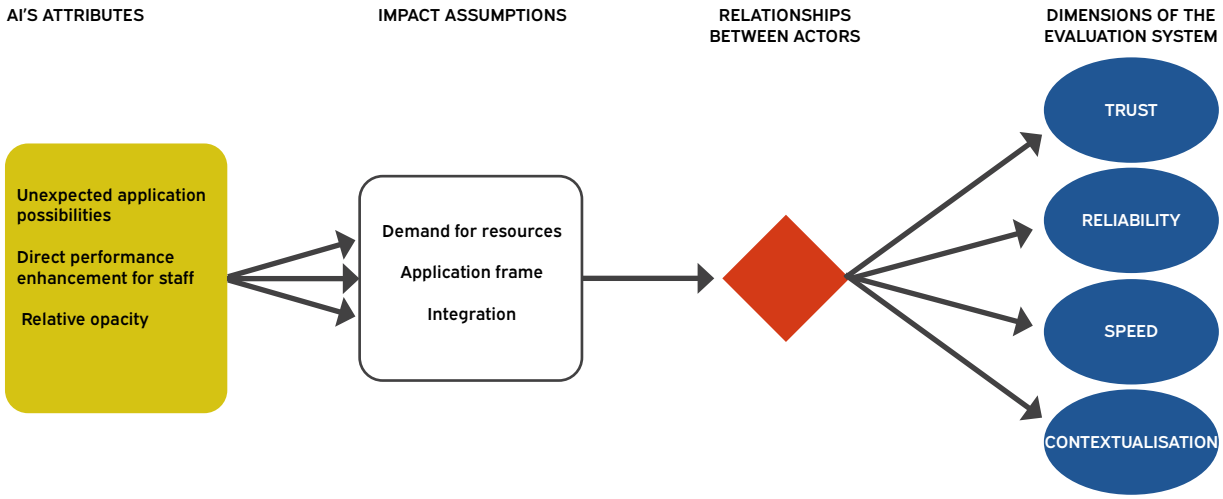

Figure 4: Schematic Overview of the Model

The model's complexity can be scaled or expanded as needed. As shown in Figure 3, one may choose to focus on a single AI characteristic, develop related assumptions, and explore their impact on selected actors and a single dimension of the evaluation system. Alternatively, as shown in Figure 4, one might examine multiple AI characteristics, actors, and dimensions in combination, or begin with simpler considerations and gradually increase complexity.

This step-by-step model construction is designed to enable structured engagement within a framework of constraining parameters, helping to address complexity in a manageable way. The model serves two main purposes: it can help identify relevant questions, and it can support the development of strategic considerations – either for individual actors or for constellations involving multiple actors. Both uses can strengthen actors' capacity to act in response to emerging developments in the field of AI and foster more coordinated decision-making in support of the evaluation system.

EXPLORING THE IMPACT OF AI ON THE EVALUATION SYSTEM

Dimension 1: A strong emphasis on trust in the processes that generate evidence, along with high expectations for responsible handling of data and results.

1. **Unexpected applications:** This characteristic of AI challenges the need for trust, as the concept of the „jagged frontier" implies a constant process of exploration regarding which tasks can and should be meaningfully performed by AI, and whether—and where exactly—a functional division of labor between humans and AI is actually implemented in the evaluation process.

2. **Direct performance enhancement of staff**: This characteristic of AI shifts agency—and potentially responsibility—from the institutional level to individual employees. As a result, an organization's reputation may increasingly depend on the perceived trustworthiness of individual staff members rather than the institution as a whole.

3. **Relative opacity**: This characteristic also challenges trust. The „black box" nature of AI constantly carries the risk that plausible but fabricated or incorrect claims go undetected.

Dimension 2: Reliability of knowledge production and reproducibility of findings

1. **Unexpected applications**: This characteristic could increase the reliability of knowledge production, as more tools for collecting and analyzing data—especially large volumes of qualitative data—are available and relatively easy to apply.

2. **Direct performance enhancement of staff**: This characteristic could reduce the transparency of knowledge production, as individual staff members may become less reliant on methodological exchange with colleagues. However, this has also been the case in the past when employees possessed expertise and skills that were not redundantly available within an institution. At the same time, it is conceivable that using AI methods for repetitive data processing could free up time – enabling more in-depth methodological and substantive exchange.

3. **Relative opacity**: This characteristic fundamentally contradicts the notion of reliability. For instance, GPTs are designed never to generate identical responses, making reproducibility in the traditional sense unattainable.

Moreover, the tools themselves—and thus the results they can produce—are in constant flux. Freely accessible GPT interfaces are operated by private providers, who may restrict or withdraw access at any time. As a result, individual users may also face challenges in ensuring reproducibility.

Dimension 3: Speed and (cost-)efficiency in the production of evidence to enable timely decision-making.

1. **Unexpected applications:** A highly positive characteristic with the potential to rapidly increase speed and efficiency.

2. **Direct performance enhancement of staff:** The key to realizing efficiency gains lies in the ability of staff to recognize and implement new application possibilities – ideally faster than their competitors within the system.

3. **Relative opacity:** The actual efficiency gains may be difficult to capture, as they could be offset by high costs for verifying data reliability or for technical infrastructure and support.

Dimension 4: The need for hard to objectify context-specific knowledge and sector expertise – elements that have always represented a kind of "black box" within evaluation systems.

1. **Unexpected applications:** Both clients and contractors can use GPTs to reflect on industry-specific questions, such as for brainstorming purposes. A wide range of monitoring options and, to a certain extent, even automated analyses could emerge through low-threshold, accessible AI methods. A more speculative outlook is that internal monitoring itself may evolve and be partially replaced by local AI systems providing analyses that, while not perfect, are acceptable.

Direct performance enhancement of staff: Individuals now theoretically could access written, explicit domain knowledge, potentially acquiring or simulating parts of industry-specific expertise in ways that previously required years of professional experience. At the same time, neither AI nor evaluators or policymakers with limited industry backgrounds are capable of understanding and interpreting implicit knowledge—such as social norms, group dynamics, or historical trajectories. These forms of contextualized knowledge will continue to require the contribution of experienced actors. The challenge for institutions lies in understanding when such knowledge is essential and thereby implementing checkpoints within new workflows. There is also evidence suggesting that the relevance of domain expertise is diminishing, as has been studied in the context of online search engines (cf Fisher et al. 2022).

Relative opacity: Even if it is possible to integrate checkpoints into the evaluation process, it will remain very difficult – despite proper documentation – to discern which aspects of a given work originate from an AI and which from a human, and to determine which of these constitute legitimate contributions and which do not.

EXPLORING THE IMPACT OF AI ON RELATIONSHIPS AND PROCESSES

The integration of AI-based tools into the evaluation landscape of research, technology, and innovation (RTI) policy in Austria will not only affect existing processes but also create new dynamics among actors and patterns of interaction. Based on our initial assumptions, we attempt to anticipate the fundamental ways in which the entry of AI could impact relationships between actors. These reflections are intended to stimulate further discussion around the technology and are by no means conclusive.

Clients – Contractors:
Traditionally, clients have commissioned evaluations and received reports from the contractors. As a result, contractors are accountable for the methods and technologies used in data processing. However, with the advent of new technologies, clients could become more directly involved in the evaluation process by shaping certain aspects of the AI models themselves.

Closer collaboration between clients and contractors during the design and development of AI models could lead to more tailored monitoring and evaluation systems that are better aligned with the clients' needs.

Contractors – Evaluators:
Contractors are responsible for carrying out evaluations, while evaluators as individuals traditionally enjoy considerable autonomy in how they work. Due to AI's potential for unexpected applications and its direct use by employees, employers may be unaware of whether and how evaluators are using AI. If the use of AI remains opaque, internal competition among staff could intensify, as performance pressure increases. Colleagues who fail to adopt AI in time or sufficiently may fall behind in the professional peer dynamic.

Evaluators will need to expand their skillsets to include not only traditional evaluation competencies, but also the ability to work with AI systems. Contractors, in turn, must provide appropriate frameworks and support measures to enable this development.

Evaluators – Evaluatees:
Evaluatees may refuse to provide data if they have doubts about how transparently it will be processed. This makes it essential for evaluators to communicate clearly and openly about how AI is integrated into the evaluation process. Doing so may require a stronger involvement of evaluatees in the assessment itself.

Participatory approaches that grant evaluatees insight into the functioning of the AI tools being used – and offer them the opportunity to give feedback – could lead to more robust and widely accepted evaluation outcomes.

Evaluatees – Clients:
As automation enhances monitoring and impact assessment capabilities, the expectations of clients may rise accordingly. This could increase the pressure on agencies and programmes to deliver evidence and demonstrate accountability. In response, evaluatees may resist the expanded use of AI-driven monitoring tools, particularly if they perceive them as intrusive or overly demanding.