DISCLAIMER

This is a translated draft of a working paper. The original German version was submitted as a contribution to the fteval Journal issue #55 and is currently under review. The translation was done with the support of Chat GPT 4.o and has not received a final editorial check.

If the article is accepted, it will be published in its final carefully edited English and German versions in the fteval repository here: https://repository.fteval.at/id/eprint/707/

Please cite this preliminary document as follows:

Thomas Palfinger, Felix Gaisbauer, Isabella Wagner and Susanne Beck (2024): How does artificial intelligence impact the evaluation system? Discussion points for designing tomorrow's evaluation system [Unpublished manuscript]. fteval Platform, Vienna.
If published, the final article will be made available here: https://repository.fteval.at/id/eprint/707/

Discussion Paper

HOW DOES ARTIFICIAL INTELLIGENCE IMPACT THE EVALUATION SYSTEM?

DISCUSSION POINTS FOR DESIGNING TOMORROW'S EVALUATION SYSTEM.

Thomas Palfinger, Open Innovation in Science Center, LBG
Felix Gaisbauer, DLR Project Management Agency
Isabella Wagner, fteval Platform
Susanne Beck, Warwick Business School

Contributions by:

Elisabeth Froschauer-Neuhauser, AQ Austria
Vitaliy Soloviy, AIT
Michael Strassnig, WWTF
And all the members of the AI Working Group

CONTENT

ABSTRACT

In this discussion paper, initial considerations are made on possible changes through artificial intelligence (AI) in an evaluation system using the example of research and innovation policy evaluation. Previous discourse on artificial intelligence (AI) has focussed primarily on data protection, ethics and scientific and methodological reliability. This discussion paper aims to introduce an additional, systemic perspective: It examines how the relationships between actors in an evaluation system change when this system is confronted with generative AI. A working group of the Austrian Platform for Research and Technology Policy Evaluation (fteval) addressed this question from July 2023 to March 2024. Based on internal discussions and accompanying literature research, a model was developed that serves as an instrument for reflecting on and jointly discussing the practical actions of the actors in the RTI evaluation system. The starting points are the assumptions that generative AI is entering the existing evaluation system as a disruptive element and that these tools can change the relationships between the actors. Although the focus is on the Austrian evaluation system, it is assumed that the model could also be useful in other evaluation systems or industries with similar structures.

INTRODUCTION

Approaches to artificial intelligence (AI) have recently expanded capabilities to produce human-like work and significantly increased public access to technology, such as OpenAI's ChatGPT (cf. Dell'Acqua et al. 2023). Consequently, strong narratives from various societal sectors (from science to politics) are emerging around this technology, which are hard to ignore regardless of the technology's actual capabilities and use.
As a result, the evaluation field, like many other systems, faces the need to find a constructive approach to the rapidly developing technology. The impact of generative artificial intelligence on the scientific process and thus on various research-based data collection methods of the evaluation process is potentially far-reaching (cf. Van Noorden et al. 2023), indicating significant upheavals in the coming years (Haupt et al. 2022; Chapinal-Heras and Díaz-Sánchez 2023; Stahl 2023; Konya and Nematzadeh 2024).
Discussions about AI have so far mostly focused on aspects of data protection, ethics, and scientific or methodological reliability in AI use. With this discussion paper, we hope to introduce an additional, systemic perspective into these discussions: How do the relationships between actors in an evaluation system change when this system is confronted with generative artificial intelligence (AI)? This question was posed by a working group of the Austrian Platform for Research and Technology Policy Evaluation (fteval) with and within their community from July 2023 to March 2024. Based on internal discussions and accompanying literature review, this discussion paper develops a model offered as an instrument for actors in the RTI evaluation system to reflect on their practice and discuss it collectively. The starting points are the assumptions that (1) generative AI enters the existing evaluation system as a disruptive element in the form of language models and (2) these AI tools can change the relationships between actors in this system. Although we will heavily refer to the Austrian evaluation system because the working group's experiential knowledge is greatest there, we assume that the model will also be useful in other evaluation systems or industries with similar structures.

THE EVALUATION SYSTEM AS A RELATIONSHIP SYSTEM: BASIC CONCEPTS AND ASSUMPTIONS

This discussion paper considers the impact of AI at the "relationship level" between different actors in the evaluation system. To reflect on the actors' framework of action regarding the system's entry of AI, we first describe the simplified status quo of the evaluation system before introducing AI as an extension and new technology with disruptive potential to this established relationship structure. In the following sections, we must define some basic concepts and make our assumptions transparent.

STATUS QUO: THE EVALUATION SYSTEM AND THE RELATIONSHIPS BETWEEN ITS ACTORS

Evaluation System: We understand the evaluation system as a practiced network of relationships with known actors, more or less codified standards, and the possibility of repeated interaction, which implies the possibility of relationship maintenance. The role distribution and the nature of the relationship between the actors are therefore considered very stable over long periods. We assume that the continuous interaction of selected actors significantly contributes to the emergence of an "evaluation culture" (in the sense of established interaction patterns) and thus to the evaluation system of an industry. Furthermore, for this discussion paper, we assume in a highly simplified manner that the evaluation system is closed in itself, deliberately excluding the significance of other actors than those considered here. Thus, it is also excluded how the evaluation system is exactly constituted or delineated from other societal systems. In this evaluation system, actors are (1) clients, (2) contractors, (3) evaluators, and (4) the evaluation object (cf. Figure 1).

Clients are, in this context, the program owners or responsible parties of an intervention – and therefore tend to be the clients of its evaluation. The client requires efficient and precise evaluation results to make informed decisions. There is a high need for clear communication channels and a clear interpretation of evidence. Besides data quality, trust in the evaluation practice is essential because the clients must rely not only on the data itself but also on meaningful contextualization in the policy field, good stakeholder management, and to some extent, the ability to anticipate a yet unmeasurable future and develop scenarios or recommendations.

Contractors are understood here as the institutions that commission evaluations and are responsible for their execution. Contractors need resources and as comprehensive access to high-quality data as possible. Every evaluation activity aims to use the available personnel and budget resources economically or even increase them, for instance, by building knowledge and know-how or generating profits. Efficiency gains are therefore welcome, along with maintaining the quality of work, as the reputation regarding the expected trust in evidence production and the reliability of methods is crucial.

Evaluators: These are the actors who conduct the evaluation as individuals, generate, process, and analyze data, and provide recommendations regarding the evaluation questions. They are committed to both the client and the contractor as employers and may have expectations for their work methods and content, which implies aspects such as career ambitions, team dynamics, personal interests, and individual development. Due to their expert role and contextual knowledge, these actors often enjoy a high degree of autonomy and responsibility despite their dependence on the employer, and therefore, they can also be a starting point for developments.

Evaluatees/Owners of the Evaluation Object: Depending on the context, this can be the intervention or program itself, an institution or organizational unit, or the performance of one or more individual(s) being evaluated. Evaluatees must understand how AI is integrated into the evaluation process. Transparency and communication can be crucial in ensuring the trust of those affected by the evaluation.

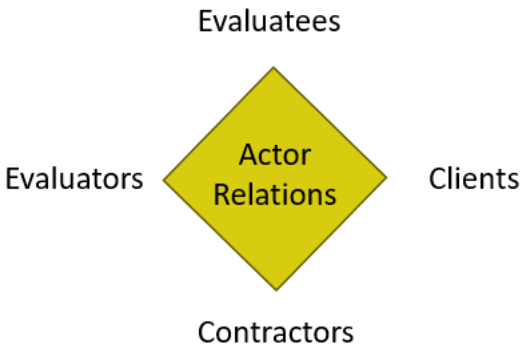

Figure 1: Sketch of the central actors and their relationships

Essentially, clients, contractors, evaluators, and the evaluatees themselves interact, with these groups being highly varied in their degree of formalization. For instance, evaluators are individuals with a much lower degree of organization than the institutions they work for. The evaluation object, in turn, can be an entire institution or a single program, usually understood as a collective effort by individuals within a framework provided by one or more institutions. The evaluation object thus lies somewhere between an individual and an institutional level, depending on what is being evaluated. This results in numerous interaction possibilities, making an analysis or conceptual presentation without simplifications impossible. Regarding the relationships of different actors, assumptions had to be made to reduce complexity without significantly minimizing the explanatory power of the statements.

As a simplifying assumption for the mutual interactions in this evaluation system, we assume that the connections between evaluators as individuals and the contracting institution play a subordinate role because they are mediated by a representative of the contracting institution. Similarly, there is no direct relationship between the contracting institution and the evaluatees because it is the evaluators themselves who interact here. These relationships are also sketched in Figure 1. The sum of the actors, their activities, and interactions

form the evaluation system. Based on the above description, we assume at least four dimensions of the evaluation system properties that affect all actors:

- Dimension 1: High demand for **trust** in processes that generate evidence. High expectations for trustworthy handling of data and evidence
- Dimension 2: **Reliability** of knowledge production and reproducibility
- Dimension 3: **Speed** and (cost) efficiency of evidence production for rapid decision-making
- Dimension 4: At the same time, hard-to-objectify **contextualization** and industry knowledge is needed, which has always been considered a "black box" in the evaluation system

We assume that applying AI impacts one or all actor groups and their relationships, raising the question of how AI affects the evaluation system. Therefore, in the next section, we address the fundamental properties we attribute to AI in this context.

EXPANSION: AI ENTERS THE EVALUATION SYSTEM AS AN ACTANT

For this article, we understand AI as an actant, loosely drawing on the ideas of Actor-Network Theory (Latour, 2005). In short, this theory suggests that non-human actors, as "actants," can also be active in an action relationship, even if they do not act themselves. Therefore, we attribute properties to AI – without anthropomorphizing it – that have the potential to influence relationships between actors. Thus, AI can change the network of relationships and action possibilities, which is why we will consider AI or AI-based tools as actants in the evaluation system.

We understand artificial intelligence (AI) as "the capability of a machine to imitate human abilities such as logical reasoning, learning, planning, and creativity."[1]. Technically, AI is based on various statistical methods. In "machine learning," patterns and regularities in large datasets are recognized, and each new task leads to further learning (e.g., online search engines refining their results). Large language models (LLMs) are models trained on large datasets that can independently generate new answers based on underlying statistical probabilities. "Generative AI" is more broadly defined and describes any AI capable of creating new texts, images, video, or audio clips. While all forms of AI are likely to impact existing systems in the coming years, the focus of this paper's considerations will be on practical generative AI, approximated in functionality to ChatGPT, LLama, Gemini, or Claude. Unlike, for example, image recognition AI, these can interact in real-time and assemble new information from these exchanges. For this reason, we attribute a particularly high potential to this AI type as actants capable of challenging or changing existing relationship systems, making them especially relevant for this article. Although the existing generative AIs serve as reference points, this article will not work through their specific capabilities but will attempt to conceptually capture them for the discourse.

Regarding the properties of (generative) AI, we must simplify significantly for clarity. Based on the work of Dell'Acqua et al. (2023), we attribute three central characteristics to generative AI that distinguish it from machine learning: Generative AI has (1) surprising unplanned application possibilities, (2) the ability to directly enhance employee performance, and (3) a relative opacity because it can generate incorrect but plausible results. Together, these three properties have the potential to influence social and institutional relationships and significantly impact their working methods. We will discuss these three properties in more detail below.

Generative AI has, first, surprising application possibilities for which it was not specifically developed, and these can rapidly evolve over time. This happens both when the model's size and quality improve and when user behavior changes. Although trained as general models, generative AIs exhibit specialized knowledge and skills as part of their training process and during normal use (Singhal et al., 2022; Boiko et al., 2023). While there is still considerable debate on the concept of emergent abilities at the technological level (Schaeffer et al., 2023),

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

the effective capabilities of AIs are novel and unexpected, widely applicable, and significantly increase over short periods. Current work shows that AI can be used at a high level in professional contexts from medicine to law (Ali et al., 2023; Lee et al., 2023), thereby also influencing these contexts. So far, each generation of AI models has shown significant improvements in their capabilities, suggesting that with each iteration, new surprising and unplanned application possibilities could emerge from practical use. This situation also means that generative AIs and their design capabilities can be difficult for individuals and institutions to grasp and manage beyond basic usage principles.

Because generative AI is fundamentally capable of solving domain-specific problems, it secondly has the ability to directly enhance the performance of individuals using these systems without requiring extensive organizational or technological investments or technical expertise on the part of the users. Studies suggest direct performance enhancements through AI use, particularly in writing tasks (Noy and Zhang, 2023) and programming (Peng et al., 2023), as well as in idea generation and creative work (Boussioux et al., 2023; Girotra et al., 2023). Therefore, AI is likely to impact professions dealing with complex tasks, providing individuals with the opportunity to achieve performance improvements quickly and independently of their employer, which they can use for themselves or the company.

As the third relevant characteristic, generative AI is characterized by relative opacity, which means that AI models can produce incorrect but plausible results (hallucinations or fabrications) and thus make factually false claims. This relative opacity extends beyond a technical or methodological understanding because even revealing the source code cannot fully explain the derivation of the results. Therefore, AI can be effective in some tasks while failing unpredictably in others. This means that the possibilities for sensibly using these AI systems cannot be comprehensively presented by their developers from the outset but must be worked out through continuous experimentation and error-making by users, for example, by sharing experiences and heuristics across various online forums (such as user groups, hackathons, Twitter feeds, and YouTube channels). For organizations, this characteristic of AI means they can never fully know what kind of tool they are introducing into their work processes and must learn to deal with this relative opacity.

AI AS A DISRUPTION FOR THE EVALUATION SYSTEM

These three characteristics of generative AI combined mean that the pros and cons of AI for employees and organizations could be challenging to grasp. Dell'Acqua et al. (2023) show that generative AI can cover an uneven spectrum of knowledge work capabilities, creating a "jagged technological frontier": While AI capabilities increasingly overlap with, extend, or even surpass human capabilities in some areas, they can remain entirely useless or counterproductive in others. The challenge is to determine which is the case against a specific system or problem background. However, since AI capabilities are rapidly evolving and often poorly understood, even experts find it challenging to determine exactly where this "jagged frontier" currently lies.

Skilled navigation at this frontier can yield significant productivity advantages for users or institutions working with AI. This can significantly impact relationships within a system because individuals and institutions can gain advantages over the pre-AI state that may be opaque and hard to replicate for other actors in the system. Additionally, mistrust can arise in a system if AI performance decreases when applied outside this frontier. Therefore, it can be concluded that both successful and unsuccessful navigation of the "jagged frontier" can have a disruptive impact on a system, especially if the use of the technology is desired systemically and thus not inherently stigmatized.

Given this background, the evaluation system is a particular case because it is likely to struggle with disruptive changes due to its structure (e.g., often long-term negotiated relationships between actors) and orientation (e.g., focus on trust and reliability) as a reputation-based system. Simultaneously, it is always under pressure to justify resource use more efficiently (e.g., speed), which in the medium term speaks for the imminent and widespread use of AI. Against this backdrop, the previously identified characteristics of generative AI are

particularly relevant for the disruptive potential of generative AI and the relationships within the evaluation system.

MODEL FOR UNDERSTANDING THE ENTRY OF AI INTO THE EVALUATION SYSTEM

The three previously derived characteristics of AI - (1) surprising capabilities, (2) direct performance increase, (3) relative opacity - are contrasted with four dimensions we have defined for the evaluation system - trust, reliability, speed, and contextualization. In this consideration, the actors and their action and reaction logics in dealing with AI can be added to form an overall model of the evaluation system. The model is intended to serve users as a reflection tool to identify knowledge gaps and design possibilities through relevant questions.

We assume that AI's disruption potential for the evaluation system as an actant varies depending on the expression of AI's characteristics. What this expression looks like depends partly on the properties and capabilities of the AI and partly on how it is used or allowed to be used. Therefore, it is conceivable that the introduction of the same AI tools will occur for different purposes and to varying extents, meaning AI's disruption potential should be thought of in scenarios rather than uniformly. To account for this, we assume two idealized positions of AI as an actant in a system – a weaker and a stronger one:

- "Simple application of AI": AI tools are used to solve simple tasks and make questions that have already been addressed more efficiently solvable.
- "Complex application of AI": AI models are used to solve complex tasks and questions that were previously unaddressable, enabling entirely new ways of problem-solving.

The terms "simple" and "complex" are not used here as properties of the applied technology itself. The same technology can be used in both applications, but the handling requirements differ: While in "simple" applications, routine tasks can be outsourced to AI for efficiency gains, and the reliability of the results should be somewhat predictable for experienced evaluators, the assumption is that for questions previously unaddressable without AI, the possibilities for reliability checks are lower.

PRESENTATION AND KEY QUESTIONS OF THE MODEL

To simplify, we assume in our model a one-sided impact of AI on the evaluation system and the relationships within it. In this model, AI is a new actant entering an existing system, potentially triggering change or disruption processes for existing relationships. We focus on the impact of AI on relationships between actors and the resulting influences on aspects of the evaluation system.

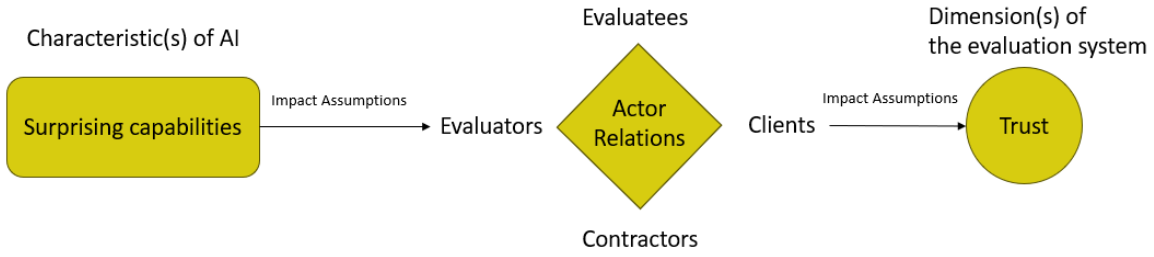

Figure 2 Interaction of AI, Relationships, and System

Figure 2 graphically depicts the basic mechanisms of the model. Starting from a defined characteristic of AI (in the example shown: unexpected application possibilities), we make assumptions about this characteristic for selected actors and consider how these assumptions might affect their relationships. From there, we can then work out the impact on one or more of the four dimensions of the evaluation system described by us (in the example: trust). Thus, starting from AI characteristics, we can derive possible impacts of AI's system entry

through intermediate steps. In applying this model, we repeatedly asked the following questions that guided us through the entire process of dealing with AI:

1. Do we assume that a specific AI characteristic influences the existing relationship practices of relevant actors in a system?
2. What assumptions can be made about the impact of AI, and which of these appear particularly relevant?
3. How do the impact assumptions influence the existing relationship framework between actors?
4. What are the consequences of changes in the relationship framework on one or more dimensions of the evaluation system?
5. Are these consequences desirable from our perspective? How would they need to be designed to be desirable?

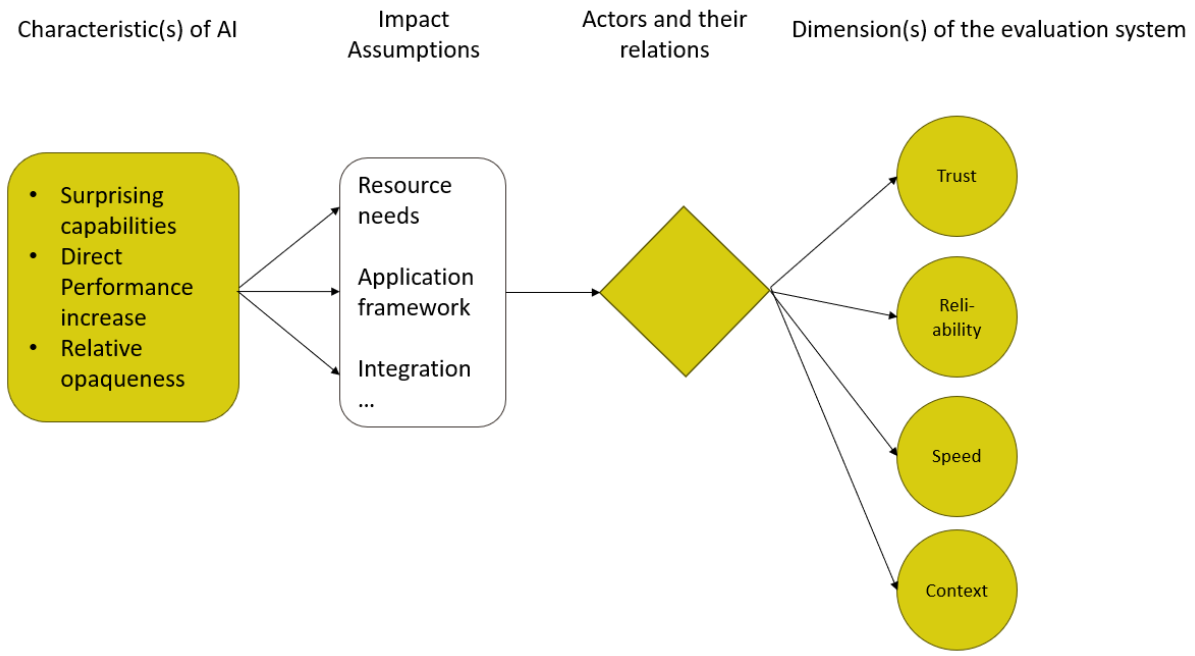

Figure 3: Schematic Overview of the Model

The complexity of the model can be varied or expanded as needed. As in Figure 3, one can extract a single AI characteristic and make impact assumptions applied to individual actors and a selected dimension of the evaluation system. Alternatively, as in Figure Y, multiple AI characteristics, actors, and dimensions can be played out in relation to each other, or one can start with simpler considerations and gradually make them more complex.

By filling in our model step by step, we aim to enable a structured discussion in the context of limiting parameters to engage with the many possibilities despite the complexity. The model is suitable for two tasks: it can help raise relevant questions or assist in formulating strategic considerations (both for individual actors and systems of multiple actors). Both can increase the actors' ability to act in the evaluation system concerning new developments in AI and help make coordinated decisions for the system.

IMPACT OF AI ON THE ACTORS AND POSSIBLE REACTIONS

Because the actors in the evaluation system are interconnected, they will be affected by AI even if they do not use it themselves (or are not allowed to). This creates a situation where all actors, regardless of their position on AI, must engage with AI and develop positions or strategies. To support this process, we present generalized assumptions about AI's impact on individual actors in the next section based on the model described above.

The focus was on formulating possible strategies for each actor group. At the beginning of each section, the assumptions made are disclosed, and which dimensions of the evaluation system, AI characteristics, and actor relationships are considered particularly relevant and therefore particularly included in the analysis. The result is completed tables with possible positive and negative impacts (for both simple and complex applications of AI) and resulting strategic options for each actor group.

CLIENTS

For policy or intervention owners, the challenge is particularly significant as they have no direct control over which actors use AI at which point and how. However, they are particularly dependent on the trust in the evaluation system being high, as the evaluation results form the basis for decisions (e.g., for or against continuing a program). This trust is especially important for the clients towards the evaluatees, as these two groups usually work together over long periods, making the misuse of AI a significant risk for this relationship. Clients may therefore feel compelled to introduce rules or controls, which, however, incur costs or stifle the development potential of the new technology and may be negatively received by other actors like evaluators in the system.

Impacts for the "**Clients**" Actor Group upon AI's Entry into the Evaluation System			
		Positive	**Negative**
Disruption Potential	**"Simple" AI application**	The direct enhancement of individuals – especially evaluators – allows evaluation results to be provided faster. Larger datasets can also be processed more efficiently, potentially leading to cost savings in the evaluation system.	Even with a simple AI application, the identified relative opacity in the results can lead to trust losses. Especially with increasing data volumes, it becomes harder for clients and evaluatees to understand how the results were achieved. Clients may therefore intervene in the evaluation practice and impose rules, with unknown consequences for the long-term introduction of AI in the evaluation system.
	"Complex" AI application	AI allows previously hard-to-grasp aspects to be better understood and evaluated. This can illuminate programs from entirely different but relevant perspectives, enhancing their impact. For this, the client must either develop a good understanding of the application possibilities or trust that contractors share this information with them.	The high significance of unexpected application possibilities, combined with relative opacity, can lead to the statements of these models lacking integrity, making them unsuitable as a decision basis for clients. Clients need to build substantial capacities for interpreting and assessing the results to use them safely.
Strategies for the "**Clients**" Actor Group in Dealing with AI			
		Offensive	**Defensive**
		Clients actively create incentives to encourage engagement with AI within their organization and build capacities for AI application. They also establish forums with evaluatees to exchange on AI's application possibilities and limitations, identifying particularly interesting application areas for their programs. Clients take a proactive role and define success indicators and goals for AI use in the evaluation system, monitored over the coming years with other actors to identify issues early. Clients advocate for AI use and establish clear escalation paths for	Clients take a passive role in AI introduction in the evaluation system, leaving it to other actor groups to bring this technology into active use as long as trust in the evaluation system is maintained. To monitor trust levels, clients need to exchange with other actor groups. Clients do not take an active role but regularly train their employees on AI's known application possibilities.

	other actors if AI is not ideally used in specific cases.	

CONTRACTORS

Contractors are in a unique position as they face competition with other contractors, must meet clients' requirements, and could enter a new dependency relationship with their employees regarding AI. Speed (efficient evidence production) and the ability to contextualize (availability of expertise/industry knowledge) are two relevant dimensions to remain competitive in the evaluation system. However, they may adopt AI models more slowly than their employees, as these are freely and easily accessible and can be applied without employer involvement, potentially increasing dependency on these specialists. On the other hand, contractors can build structural capacities for AI, enabling them to solve specific contextualized tasks better (either by training their own models or preparing data for AI), requiring resource investment that individuals cannot bear. In competition with other contractors, effectively integrating AI into workflows and leveraging efficiency gains is essential. Access to high-quality data and the ability to interpret AI results can be critical to their success in the evaluation market, leading to competition rather than cooperation at this level. There may also be concerns about new market entrants unfamiliar with systemically ingrained practices (such as fteval membership).

Impacts for the "**Contractors**" Actor Group upon AI's Entry into the Evaluation System			
		Positive	**Negative**
Disruption Potential	**"Simple" AI application**	AI application Direct performance enhancement through AI for employees – such as in report writing – leads to more efficient resource use. These resources can be redirected to other parts of the evaluation or into building additional capacities/skills, increasing evaluation quality long-term.	Direct performance enhancement through AI is not shared with the organization but used for individual gain, increasing mistrust of contractors towards their employees and potentially leading to new monitoring measures. Quality assurance becomes more difficult for contractors under such conditions, despite the relative opacity making it particularly necessary. The organization may also need to mediate between employees' differing interests, tying up additional resources.
	"Complex" AI application	Complex AI application offers the possibility of utilizing high-quality data for demanding analyses. Since expertise is increasingly covered by AI, new job profiles with cross-sectional competencies may emerge, facilitating collaboration across different areas within and outside an organization and enabling new approaches and evaluation forms.	Adapting to complex AI models may require substantial restructuring within an organization (new job profiles needing new contracts and/or training). Additionally, technical infrastructure and expertise must be developed without certainty about the ultimate benefit. It remains unclear if employees use unexpected application possibilities (similar to direct performance enhancement) for their advantage, increasing the risk for organizations that their investment does not pay off.
Strategies for the "**Contractors**" Actor Group in Dealing with AI			
		Offensive	**Defensive**
		AI is seen as an opportunity to experiment with new methods and approaches, providing employees with an appropriate environment. The focus is on developing better solutions for existing problems and	Contractors leave AI application and management largely to their employees, conveying minimal principles for its application or prohibitions. Active investments in this area are kept limited,

	identifying newly achievable solutions for previously unsolvable issues. Contractors build good relationships with experts, taking trust-building measures to reduce stress among employees due to these developments (thus avoiding resistance). Resources are allocated to build corresponding expertise within the organization. Trust-building measures are designed to ensure that new application possibilities can be actively shared within the organization.	relying on "traditional" evaluation methods. This approach is seen as a distinguishing feature in the field and aims to give clients and evaluatees the necessary confidence in the applied methods and employees' trust in the organization. Consequently, AI use may be prohibited for employees on specific assignments.

EVALUATORS

For evaluators, AI offers the possibility to independently design their workday more efficiently and outsource tasks to AI. The better evaluators navigate AI, the greater the potential benefit, so they may have a strong interest in identifying unexpected application possibilities. However, they may also bear considerable responsibility if there are no institutional solutions, as they must manage AI's relative opacity and decide whether AI-generated results meet the evaluation system's standards (sometimes in cases where there are no exact standards and practices yet). An unstructured introduction of AI by the evaluators themselves can lead to them largely determining the reliability and trust in the evaluation system, potentially overburdening them. A good relationship between contractors and evaluators and the establishment of appropriate structures can mitigate this redistribution of responsibility. However, the personal interests of evaluators must also be considered, such as not wanting to share efficiency gains with employers for relative advantages within the team or career paths. Conversely, employees may see their jobs threatened by AI tools, leading to conflicts with employers keen on AI.

Impacts for the "**Evaluators**" Actor Group upon AI's Entry into the Evaluation System			
		Positive	**Negative**
Disruption Potential	**"Simple" AI application**	Direct performance enhancement can relieve evaluators from routine tasks through automation, which they can implement largely independently and tailored to their needs. This would allow a focus on more interesting tasks such as complex analyses and interpretations.	Direct performance enhancements are only accessible to individuals engaging with the technology, potentially leading to "generation conflicts" within a team or organization. Those wanting to use AI for more efficient work may encounter resistance from those refusing, causing delays for the former and feelings of threat for the latter.
	"Complex" AI application	Complex AI application can enable evaluators to conduct far-reaching projects beyond their core expertise, continually acquiring new industry knowledge. This ongoing exchange with colleagues can mitigate challenges such as AI's relative opacity and share unexpected application possibilities more broadly. A new understanding of work design with new freedoms and possibilities emerges.	Complex AI application leads to overwhelm, with AI's relative opacity increasingly resulting in distorted but hard-to-identify errors. If evaluators cannot exchange (due to AI prohibition by the organization) or do not want to (personal advantage), this corrective falls away. Uncertainty about who bears responsibility for AI-supported analyses leads to no or only covert use.
Strategies for the "**Evaluators**" Actor Group in Dealing with AI			
		Offensive	**Defensive**

	Evaluators become drivers and designers of AI application. They advance the development of codified guidelines and implicit behavioral rules based on real-world practice within institutions and the system. This allows safe outsourcing of routine tasks to AI, freeing up time for demanding tasks. Personal interest and design capability lead to regular training and updates on AI developments being welcomed, and exchange formats with colleagues are maintained.	Evaluators face efficiency competition pressure and feel compelled to use AI as an "shortcut" ad hoc and unreflectively to cope with this challenge. AI application only occurs where efficiency pressure leaves no other option. Exchange between individuals is hampered by time pressure and the lack of official technology use, with those benefiting from AI not wanting to lose advantages due to bans. Reluctance to participate in training under these conditions further affects evaluation quality.

EVALUATEES

For evaluatees, new forms of evaluation could emerge through expected and unexpected AI application possibilities, allowing for a multifaceted and broader evaluation of their activities. This can increase trust in the evaluation system as the program's successes can be presented more differentiatedly. Evaluatees must be convinced of the reliability of these new evaluation forms. At the same time, AI potentially enables the internal introduction of their robust, novel monitoring systems, allowing evaluatees to reflect on their activities and providing a good basis for evaluation. However, these internal monitoring systems can also be used by evaluatees to create "counter-reports" to external evaluations, complicating the relationship between evaluatees and evaluators. Additionally, trust in the evaluation system can be diminished if clients cannot assess which evidence to trust. Therefore, a good balance must be struck between the interests of evaluatees and clients, facilitating good collaboration with contractors and evaluators.

Impacts for the "**Evaluatees**" Actor Group upon AI's Entry into the Evaluation System			
		Positive	**Negative**
Disruption Potential	**"Simple" AI application**	Besides faster availability of results, AI through direct performance enhancement can help evaluatees translate evaluation results into a more operational form, quickly finding their way into new practices.	Relative opacity of AI leads to evaluatees not understanding how the results were achieved, even with simple AI applications. This effect can be amplified by AI's pseudo-objective language. To cope, evaluatees must build their capacities in this area, diverting resources from core activities without generating greater added value.
	"Complex" AI application	The interplay between internal monitoring and external evaluation mediated by AI opens entirely new forms of collaboration between evaluatees and evaluators. Both sides can benefit from each other's unexpected application possibilities and collectively reduce the risk of relative opacity.	The above-described effect can intensify with complex AI application. If evaluatees are "at the mercy" of evaluations because there are no control measures from clients, trust in the evaluation system can quickly erode. Evaluatees may respond by not creating or making data available for AI methods.
Strategies for the "**Evaluatees**" Actor Group in Dealing with AI			
		Offensive	**Defensive**
		Evaluatees use AI to enhance their reflection and learning ability, building an understanding of the technology's application possibilities. They also embrace external efforts to use AI in evaluation and actively shape the framework conditions. A	Evaluatees use AI partly for internal monitoring but do not share these results with external evaluators, reducing their demand/need for learning from external evaluations – risking institutional blindness or higher dependency on AI-savvy employees. External processes remain unclear and

	participatory and transparent process has been developed to integrate AI into the evaluation process. There is a culture of clear communication about AI use and how the results are used.Einsatz von KI und wie die Ergebnisse verwendet werden.	undefined. Consequently, fewer external evaluations with complex AI use occur, but no positive approach to the technology is developed externally. This passive stance can lead to misunderstandings about AI use, reducing the organization's overall openness to the technology. If clients insist on AI-based evaluation, the organization may need to quickly develop processes that may be error-prone.

The considerations presented here are by no means final or complete. However, they show the diverse considerations of different actors and highlight the benefit of breaking them down into simple aspects. Our model offers a way to do this in a structured manner. Based on the strategies outlined here, the next step could be to identify strategic overlaps between actors and where their interests differ significantly regarding AI use. This creates a good foundation for a discourse on common solutions.

Relevant Questions

As mentioned, the model can also be used to identify relevant questions addressing the challenge. In developing and discussing the model within our working group, we repeatedly encountered such questions. The central perspective was to understand the evaluation system as a social system with relationship networks now being influenced by a new actant. Understanding and considering this will be crucial for a smooth transition to an evaluation system where AI plays an essential but not always visible role for all actors. The unexpected application possibilities of AI seem to play a central role, unlike relative opacity and the ability to directly enhance individuals' performance, which have received relatively little attention in the discourse on AI introduction. Particularly in the implementation and relationship between actors, this AI aspect is an exciting factor because it is not entirely clear what is being implemented and how it will interact with individual and institutional actors.

The characteristic of unexpected application possibilities of generative AI comes into play primarily during ongoing interaction between humans and AI, especially at an individual level. This characteristic holds the potential to shift action power from the institutional to the individual level. When introducing or allowing generative AI, institutions cannot a priori know the full application field of the respective AI, as it can only be grasped through interaction with AI as navigation of the "jagged technological frontier." Evaluators might hide their discoveries of AI's unexpected application possibilities to use resulting efficiency gains for themselves, especially in an environment skeptical of AI. Nonetheless, they could significantly benefit from exchanging with their "peers," requiring higher organizational structure and corresponding framework conditions, for example, from contractors. It would help if principles for using AI took this into account, allowing for experimentation while ensuring transparency in the social system (from a non-technical perspective). Contractors face the challenge of wanting to leverage this application's potential but are heavily reliant on their employees' individual handling of AI while bearing institutional responsibility for negative consequences. If contractors and individual evaluators fail to develop common norms, it can have consequences for the entire evaluation system.

This example quickly shows how AI can influence social interactions and established relationships far beyond its actual technical application, without necessarily being the target of the respective application. Actors can react very differently to AI use. Based on the scenario described above, the following questions could be derived, which we as the fteval community should consider:

- How do we envision a positive AI-permeated evaluation system?
- At what level is competition for unexpected application possibilities or AI, in general, desirable and innovation-promoting in the evaluation system? At what level is it harmful?

- Who should ultimately benefit from AI's application possibilities in the evaluation system? Evaluators? Contractors? Clients? Evaluatees? How can this be established?
- What is needed for a trustworthy relationship between evaluators and contractors or among actor groups in general?
- How can exchange processes about new AI application possibilities be organized in the evaluation system?
- What framework conditions must contractors create to ensure that evaluators openly share their findings of AI's unexpected application possibilities?
- How can institutional actors best benefit from their individual employees' experimental use of AI without taking excessive risks affecting trust in the results?
- How should responsibility be organized to minimize risks for the evaluation system or specific actors without destroying the desire to experiment?
- How can the secret exploitation of new application possibilities be reduced while rewarding individuals or institutions for their efforts?
- How should new practical knowledge of AI in the evaluation system be shared? How can the intended goal be achieved?

All these questions have ethical and legal implications but go far beyond them. If actors in the evaluation system fail to build a trusting working relationship with the new technology, negative externalities for the entire system will increase. This could lead to a decline in trust in the results, making it less attractive for clients to rely on this type of decision basis. Successfully dealing with these issues seems a highly relevant and not yet fully comprehended topic for the coming years.

CONCLUSION: WHAT NEEDS TO BE CONSIDERED NOW?

This discussion paper aims to encourage structured engagement with AI's impact on relationships in the evaluation system. We will all be affected by developments in this field. At the same time, it seems an inherent part of generative AI that we are all drivers of this development. This holds much potential but also significant risk for the evaluation system. From our perspective, the overarching question in the coming years will be: Which relationships between actors do we want to maintain as they are, and which should change for an improved evaluation system?

AI should not only be understood as a technology but also as a societal will for change. While one can probably avoid AI as a technology for years – if not longer – AI as a societal will shall forcefully (and regardless of the consequences) gain access to many different systems. The rapid technological change combined with the societal will for application necessitates creating a future vision, independent of AI's current capabilities and detached from one's stance on the technology. Otherwise, even fundamentally "positive" aspects of AI can disrupt actors' relationships in a system and lead to unintended disruptions.

This work, in the form of a schematic model with accompanying reflection questions, provides a foundation for a structured discussion about integrating AI into the evaluation landscape. As an actor, it is crucial to engage early with the potential changes and participate in shaping a future-proof evaluation process. We invite all affected and interested parties to actively participate in this discussion and jointly set the course for effective and ethical evaluation practice using AI.

ACKNOWLEDGEMENTS

This work is based on conceptual considerations by colleagues Thomas Palfinger and Susanne Beck from the Open Innovation in Science Centre of the Ludwig Boltzmann Gesellschaft, and the latter from Warwick Business School, University of Warwick. These were further developed for the evaluation context within the working

group on artificial intelligence in the evaluation of the Austrian Platform for Research and Technology Policy Evaluation (fteval). In the subgroup, which dealt with the changed stakeholder relationships from July 2023 to May 2024, Alexander internal relationships included Alexander Daminger (WIFO), Charlotte D'Elloy (Technopolis Group | Austria), Elisabeth Froschauer-Neuhauser (AQ Austria), Felix Gaisbauer (DLR Project Management Agency), Tina Olteanu (FWF), Thomas Palfinger (LBG-OIS), Vitaliy Soloviy (AIT), Michael Strassnig (WWTF) and Isabella Wagner (fteval). The entire working group also consisted of (further) representatives from AIT, aws, FFG, FWF, KMU Forschung Austria, ÖAWI, Technopolis Group | Austria, WIFO and ZSI.
Many thanks for all the ideas, contributions and 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.

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

ANNEX - WORKING GROUP'S TOPIC ARCHIVE

CONSIDERATIONS ON THE IMPACT OF AI ON THE EVALUATION SYSTEM

CHARACTERISTIC 1: HIGH DEMAND FOR TRUST IN PROCESSES THAT GENERATE EVIDENCE. HIGH EXPECTATIONS FOR TRUSTWORTHY HANDLING OF DATA AND EVIDENCE

1) Unexpected application possibilities: This characteristic of AI challenges trust because a constant exploration process is underway in the sense of the "jagged frontier" to determine which tasks can be sensibly taken over by AI and whether and where exactly a functional division of labor between humans and AI can actually be implemented in an evaluation.

2) Direct performance enhancement of employees: This characteristic of AI potentially shifts agency and thus responsibility from an institution to the employee level. The reputation of a company could increasingly be replaced by the presumed trustworthiness of individual employees.

3) Relative opacity: This characteristic also challenges trust. The "black box" AI always carries the risk that plausible but fabricated or incorrect claims are not recognized as such.

CHARACTERISTIC 2: RELIABILITY OF KNOWLEDGE PRODUCTION AND REPRODUCIBILITY

1) Unexpected application possibilities: This characteristic could increase the reliability of knowledge production by providing more tools for data collection and analysis, especially large volumes of qualitative data, which can be relatively easily applied.

2) Direct performance enhancement of employees: This characteristic could reduce the transparency of knowledge production because individual employees may be less dependent on methodological exchange with colleagues. However, this has also been the case so far if there are employees with expertise and skills that are not redundantly available at an institution. It is also conceivable that using AI methods for repetitive data revisions could save time for methodological and content exchange.

3) Relative opacity: This characteristic fundamentally contradicts reliability. For example, GPTs are programmed never to deliver the same answers, making reproducibility in the traditional sense impossible. Moreover, the offers and thus the generable results are constantly evolving, and freely accessible GPTs can be withdrawn or throttled by private operators at any time, potentially causing reproducibility issues for individual users.

CHARACTERISTIC 3: SPEED AND (COST) EFFICIENCY OF EVIDENCE PRODUCTION FOR RAPID DECISION-MAKING

1) Unexpected application possibilities: A highly positive characteristic with the potential to quickly increase speed and efficiency.

2) Direct performance enhancement of employees: Key to realizing efficiency gains. Employees must recognize and implement new application possibilities quickly(er) than the competition in the system.
The actual efficiency gains may be difficult to grasp, as high costs for data reliability checks or technical infrastructure or support may arise.

CHARACTERISTIC 4: AT THE SAME TIME, HARD-TO-OBJECTIFY CONTEXTUALIZATION AND INDUSTRY KNOWLEDGE IS NEEDED, WHICH HAS ALWAYS BEEN CONSIDERED A "BLACK BOX" IN THE EVALUATION SYSTEM

1) Unexpected application possibilities: Both clients and contractors can use GPTs to reflect on industry-specific questions, such as brainstorming. Various monitoring possibilities and semi-automated analyses with low-threshold available AI methods may also develop. A speculative outlook could be that internal monitoring itself will change and be replaced by imperfect but acceptable analyses from a local AI.

2) Direct performance enhancement of employees: Individuals now theoretically can access written and explicit industry knowledge and possibly acquire or simulate parts of industry knowledge, which previously required years of work experience. At the same time, it is impossible for both AI and industry-remote evaluators or policymakers to understand and interpret implicit industry knowledge, practices, group dynamics, or historical courses. These types of contextualized knowledge will continue to be provided only by experienced actors. The challenge for institutions is to understand when what is needed and possibly to introduce control points in new workflows. There are also indications that the importance of industry knowledge is decreasing, as researched using online search engines (cf. Fisher et al. 2022).

3) Relative opacity: Even if it is possible to effectively incorporate such control points into the evaluation process, it will still be challenging to recognize which aspects of a work are generated by AI and which by humans and to judge what constitutes legitimate statements and what does not.

CONSIDERATIONS ON THE IMPACT OF AI ON RELATIONSHIPS AND PROCESSES

Integrating AI-based tools into Austria's RTI policy evaluation landscape will not only influence existing processes but also create new actor dynamics and interaction patterns. Based on our assumptions, we try to anticipate the fundamental impact of AI's entry on the relationships between actors. These considerations aim to stimulate further discourse on the technology and are not exhaustive.

Clients-Contractors:

Traditionally, clients have commissioned evaluators and received reports from them. Contractors are therefore also accountable for the methods and technologies used for data processing. With new technologies, clients could directly intervene in the evaluation process by co-designing certain aspects of AI models. Closer collaboration between clients and contractors during the conception and development of AI models could lead to customized monitoring and evaluation systems better tailored to clients' needs.

Contractors-Evaluators:

Contractors are responsible for conducting the evaluation while evaluators have considerable freedom. Due to AI's characteristic of unexpected application possibilities and its direct use by employees, employers may be unclear if and how evaluators use AI. Intransparent handling could increase competition among employees due to heightened performance pressure. Those who fail to use AI timely or adequately could fall behind in the collegial competition. Evaluators need to expand their skills to develop not only traditional evaluation competencies but also knowledge in dealing with AI systems. Contractors must create meaningful framework conditions and support measures for this.

Evaluator-Evaluation Object:

Evaluatees might refuse to provide data if they have doubts about transparent processing. Evaluators must communicate transparently and openly how AI is integrated into the evaluation process. This could require greater involvement of evaluatees in the evaluation process. Participatory approaches, where evaluatees gain insight into AI's functioning and provide feedback, could lead to more informed and accepted evaluation results.

Evaluation Object-Clients:

Clients' demands for monitoring and impact assessment could increase with better automation possibilities. Agencies and programs could face higher pressure regarding accountability, causing evaluatees to resist new monitoring possibilities.