Electrical Power and Energy Systems 178 (2026) 111961

Contents lists available at ScienceDirect

International Journal of Electrical Power & Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

AI systems in the energy sector: risk evaluation for business operations and regulatory oversight

Fabian Heymann [a,*], River Huang [b], Antoine Marot [c], Medha Subramanian [d], Russell McKenna [a,b]

[a] Chair of Energy Systems Analysis, Department of Mechanical and Process Engineering, ETH Zürich, Zürich, 8092, Switzerland
[b] Laboratory for Energy System Analysis, Centers for Nuclear Engineering and Sciences & Energy and Environment, PSI, Villigen, Switzerland
[c] Réseau de Transport d'Electricité (RTE), Paris, FR, France
[d] Electric Power Research Institute (EPRI), Dubin, IR, Ireland

ARTICLE INFO

Keywords:
Artificial intelligence
AI systems
Electricity transmission
Energy systems
MCDA
regulation

ABSTRACT

The growing use of Artificial Intelligence (AI) and rising awareness of system-specific benefits and risks triggered businesses, regulators and policy makers' interest in the evaluation and comparison of AI systems. In this work, we introduce an operationalizable AI systems' risk evaluation framework rooted in Multi-Criteria Decision Analysis (MCDA). Applying a PROMETHEE II-based approach coupled with the revised Simos method to six exemplary AI systems employed in electricity transmission, we show how AI systems can be ranked and compared from a utility standpoint and for regulatory oversight and enforcement. We demonstrate the robustness of retrieved rankings incorporating interval-based criterion-weight uncertainty via Monte Carlo simulations, resulting in a below 20% chance of seeing rank-position changes (by one rank) for two out of the six AI systems. A potential implementation of our framework is showcased using the logic of the EU AI Act.

1. Introduction

1.1. Digitalization, AI and new risks in the energy sector

Digitalization, the wide adoption of digital technology, business models and practices, is changing traditional planning and operation processes and outcomes in energy systems [1,2]. AI is a versatile digital technology which has received growing attention as it is seen as an enabler for improved productivity [3,4]. Hence, AI has become extensively utilized in all sectors of the economy [5,6]. As such, AI has also been identified as an essential tool for the energy sector [7,8], whose transition towards higher degrees of electrification and digitalization can greatly benefit from the capabilities of AI [2,9]. For example, estimates showed that AI could unlock 1.3 trillion US dollars (USD) by reducing the investment needs for power generation through flexibility [9], while AI-enabled control and balancing can reduce power system costs by 6–13% [9].

The merging of information and communication technology (including digital technologies) with energy and power systems is commonly referred to cyber-physical energy or power systems [10], where growing attention is also paid to new, emerging risks such as cyber security. While many works started to acknowledge emerging risks stemming from digitalizing energy and power systems, these typically include risks stemming from component failure, unforeseen change in demand or supply levels, strategic behavior of other market players (e.g. [11,12]), environmental risks [13] or malicious attacks [14]. Risks from the use of AI systems have so far not been addressed.

It has been stated that the use of AI systems could bring new risks to energy and power system planning and operation, for example with regard to model bias, lack of transparency or explainability of predictions (e.g., [15,16]). Risk has been defined by the National Institute of Standards and Technology (NIST) as a combined measure of how likely an event is to occur and how severe its consequences are [17]. As this paper investigates risk from an energy/ power system perspective, its focus lies on risks that could affect the system's functioning, in other words, systemic risk. In the specific context of AI systems, risks could arise for individuals, organizations, communities, society, or the environment [17,18].

Hence, it can be expected that AI could aggravate some of the current challenges in the energy sector. [19,20]. For example, the rising use of

* Corresponding author.
E-mail address: fheymann@ethz.ch (F. Heymann).

https://doi.org/10.1016/j.ijepes.2026.111961
Received 8 January 2026; Received in revised form 27 April 2026; Accepted 22 May 2026
Available online 2 June 2026

AI could require worldwide additional 85 TWh of electricity consumption by 2027 [21]. In addition, AI could be used to automate cyber-attacks [22] or to automate energy-related payments to consumers, with already reported large-scale failures [23].

To address the risks of AI use, a multitude of mitigation measures have been proposed. Among those are scientific approaches to improve the explainability of AI models [24], AI certification schemes [25] and risk assessment and management frameworks [17,26,27]. However, there is still a lack of understanding on how to compare and evaluate risks stemming from AI utilization. In other words, "How risky is a given AI system, also compared to another?". So far, none of the above-mentioned research addresses these questions comprehensively, i.e. building on documented AI risks frameworks.

Such a risk evaluation framework needs to be sufficiently abstract to be applicable across different AI applications. Likewise, it requires a level of detail that allows the consideration and comparison of the various aspects of AI systems such as the underlying data, employed algorithms, performed tasks or the environment in which the system is used.

1.2. Related work, research gaps and contributions

The work of [32] introduced a three-dimensional framework considering a detailed taxonomy of AI algorithms, criticality of tasks and model capabilities in terms of human senses. However, determining the correct model capabilities remains highly arbitrary, given the lack of an established and widely accepted mapping of AI algorithms against their intrinsic capabilities (and limitations). The OECD framework from 2022 proposes a very detailed approach for characterizing AI systems across all economic sectors or activities [33], detailing over 35 risk criteria within five dimensions related to AI system components (further explained in Section 2.3). However, many of these criteria are rather qualitative (e.g., user competency, impact on job quality, degree of employment) and their degree of relevance (high, medium, low) are to be determined by the user, which makes this framework highly subjective to the assessor. This in return limits the use of a cross-company, cross-regulatory authority or even a wider international comparison of AI systems.

While the analysis of risks [34] and their relation to uncertainty in energy and power system planning [35] is lengthily established and an integral part of system modeling (consider [36] for an overview of traditional risk assessment methods in energy), there has been very scarce research on the evaluation of AI-related risks in energy systems.

The study of [37] provides a first framework to determine both benefits and risk of AI systems *specifically* in energy systems, it lacks a more concise and detailed description of the AI model used in each system (e.g., specifying data type and algorithm).

In addition, several international, rather procedural standards and frameworks towards assessing risks and risks stemming from the use of AI systems (e.g. from ISO/IEC and NIST, [17,27,38]) already exist. However, these works address more the *what* (why risks shall be assessed, in what broader governance, in which categories) than the *how* (which methods, which data, system boundaries). None of the above-mentioned works addresses the risks AI systems introduce to the operation and planning of energy and power systems. While AI is increasingly used in the energy industry, businesses and policymakers currently lack the tools to assess and compare AI systems across different deployment contexts, with special emphasis on their potential risks. Hence, the identified research gaps are the following:

- A lack of a generic, scientifically sound risk evaluation methodology for AI systems in the energy sector.
- A missing investigation into how such methodology could be used in practice, e.g. within energy companies and ministries /regulatory authorities to assess and rank AI systems.

The methodologies' transferability and ease of implementation are important aspects here, as an unambiguous, easy to interpret and monitor methodology that is insensitive to manipulation is of interest to businesses and regulators (e.g. as stipulated in the RACER principle, mentioned in the EU Better Regulation Toolbox [39], p.346ff).

Similarly, as is argued in the opinion piece in [40], AI risk evaluations should align to the principle of proportionality, i.e. defining the minimum information needed for a valid evaluation of AI risks; develop methods to identify evaluations that are equally effective so less burdensome options can be used; and determine how to reduce resource demands or intrusiveness without compromising the informational value of the derived methodology, while accounting for the specific trade-offs of each evaluation.

Against this background and building on a first conceptualization published in previous research [41], we propose a transferable framework, which integrates an established risk taxonomy for AI systems with a widely used method from multi-criteria decision analysis (MCDA). Our methodology is then applied to six AI systems, both operational and futuristic, that handle various tasks in energy system planning and operation. The main contributions are:

1. Developing a flexible framework to assess and compare potential risks of a large number of AI systems in the energy sector.
2. Showcasing the applicability of our methodology and considering the selected AI systems and varying stakeholder preferences.
3. Discussing the practical application of our proposed methodology in regulatory oversights and business operations together with implementation challenges.

To our knowledge, we provide the first comprehensive framework to evaluate and compare risks that individual AI systems add to the planning and operation of energy systems.

The remainder of this manuscript is structured as follows: while Section 2 reviews AI risks and international regulatory frameworks, Section 3 outlines the proposed risk-evaluation methodology and introduces six AI systems compared in this study. Section 4 reports the results and sensitivity analyses, followed by Section 5 which discusses aspects of implementation, limitations and future extensions. Finally, Section 6 concludes with main findings, further research avenues and policy implications.

2. AI risks and current regulations

2.1. An overview over AI risks

AI systems can be evaluated from multiple perspectives, such as technological readiness [42] or economic potential [4]. At the same time, the rapid expansion of AI applications has intensified concerns about the risks associated with their development and use (e.g., [43]). An important step toward a systematic understanding of these risks was carried out by [44,45], who developed a comprehensive repository of documented AI risks. This living database so far analysed over 1000 reported risks across 43 different taxonomies. This is then filtered down into 2 main taxonomic branches: a causal taxonomy and a domain taxonomy. While the causal taxonomy is oriented towards the origin of risks (which entity, with which intent and timing), the so-called "domain taxonomy" sorts all reported risks into seven main overarching domains. These domains are particularly useful as they can serve as a lens through which different *types of risks* can be identified.

The domain taxonomy distinguishes seven broad categories of AI risk [45]**.** The first is **discrimination and toxicity**, which includes risks such as unfair treatment, misrepresentation, and exposure to harmful content. The second is **privacy and security**, covering harms arising from the collection, memorization, or disclosure of sensitive personal information. The third is **misinformation**, referring to the generation or dissemination of false or misleading information that may deceive users

or cause harm. The fourth is **malicious actors and misuse**, which concerns the intentional exploitation of AI for harmful purposes, such as manipulation or cyberattacks. The fifth is **human–computer interaction**, which captures risks related to reduced human autonomy or control in AI-supported decision-making and action. The sixth is **socioeconomic and environmental harms**, including inequality effects and the environmental burden associated with the energy use of AI systems and data centres. The seventh is **AI system safety**, failures, and limitations, which encompasses issues such as misalignment with human values, dangerous capabilities, unreliability, limited transparency, and related ethical concerns.

These domains have been, together with the previous work on AI system assessments (Section 1.1.), further developed into a set of evaluation criteria that will be described in more detail in Section 3.2.

2.2. Recent regulatory initiatives addressing AI risks

Globally, many regulatory initiatives have been appearing recently to address expected benefits and potential risks of AI systems [28]. While the interested reader can find a comprehensive review of international AI regulations in [29], we will briefly introduce the developments in the EU, China, India and US.

The European AI Act, implemented in summer 2024, is perceived as the first regional regulation of AI [46]. It is a horizontal regulation that spans all economic sectors with a risk-based framework that requires AI system developers and operators of deemed high-risk AI systems to comply with requirements, such as human oversight, extensive documentation and certification [46]. It also foresees the banning of systems with "unacceptable risk" such as systems that have face recognition and social scoring functionalities [46] and may extend to other jurisdictions beyond Europe where European-made or European-trained AI models are used [47]. Findings of [48] suggest the European AI Act is the most restrictive, compared to the AI regulations of the US, India and China.

In India, AI regulations are yet to be set up in full force [31]. However, consultations are on-going on a bill called the "Digital India" bill. From media coverage and interviews with the Indian Minister for Information Technology, it can be understood that the proposed regulation would span all economic sectors ("horizontal"), with the regulation primarily focused on user harm [49,50]. It is also expected that monitoring and enforcement will become obligatory, partially through new governmental bodies to be set up.

The Chinese approach towards regulating AI can be described as a rather hybrid approach, with non-binding (ethical) principles together with a few technologically oriented requirements [30]. The US AI regulation on the other hand can be described as a vertical approach and is decentralized for each economic sector [29]. In its current shape, it relies mostly on voluntary risk assessment, self-monitoring and existing soft law, such as the non-binding NIST AI risk management guidelines [29,51].

2.3. AI system components

To classify and evaluate the risk of AI systems, basic background information of the characteristics of each AI system is needed. The latest definition of the OECD defines an AI system as "*a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment*" [52]. In this work, we build on this definition and the earlier OECD framework for characterizing AI systems [33], which is currently the most comprehensive framework and taxonomy to evaluate and compare AI systems. It defines four major components of an AI system ([48], compare Fig. 1):

- **Data & input** refers to the data and/or expert input with which an AI model builds a representation of the environment. Relevant characteristics include the provenance of data and input, machine and/or human collection method, data structure and format, and data properties. Data & input characteristics can pertain to data used to train an AI system ("in the lab") and data used in production ("in the field").
- **AI Model** is the computational representation of all or some parts of the external environment of an AI system. Its key characteristics concern how the model is built (using expert knowledge, machine learning or both) and how the model is used (for what objectives and using what performance measures).
- **Task & output** describe what the system does and what it produces. This includes tasks such as personalization, recognition, forecasting or goal-driven optimization; its outputs; and the resulting action(s) that influence the overall context. Characteristics of this dimension include system task(s); action autonomy; systems that combine tasks and actions like autonomous vehicles; core application areas like computer vision; and evaluation methods.
- **Economic context:** captures the sectoral and organizational setting in which the AI system is employed. It usually pertains to applied AI systems rather than to a generic AI system and describes the type of organization and functional area for which an AI system is developed. It includes characteristics such as the sector of deployment (e.g. healthcare, finance, manufacturing), the business function and model; the critical (or non-critical) nature; its deployment, impact and scale, and its technological maturity.

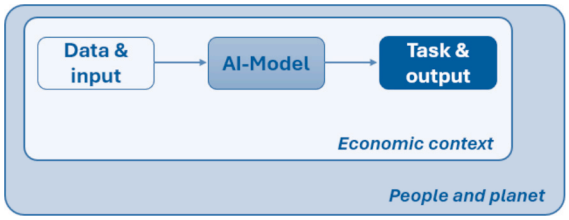

Fig. 1. OECD Framework to classify AI systems (based on [33]).

The AI system framework (particularly Data & input, AI Model and Task & Output) is further used to structure the set of derived risk criteria (Section 3.2). Economic context is defined by our domain area, which is the energy sector (critical infrastructure).

3. Risk evaluation of AI in energy systems

3.1. Risk evaluation framework

Building on emerging AI evaluation frameworks and AI regulations introduced above, we intend to provide a highly transferable and applicable framework to evaluate risks of AI systems in the energy sector and beyond. Our work is built on previous attempts to characterize AI systems presented in [33,37,53] and [32]. We synthesize and further develop the risk evaluation developed therein, to retrieve a specific, practical and generalizable framework that can be used to evaluate AI systems in practice, for example through a regulatory body or ministry. Our framework consists of three phases and can be transferred to any other sector where AI systems are used and need to be evaluated (Fig. 2).

The first phase "**Identification**" consists of identifying all AI systems subject to further analysis. This is not straightforward, as it requires a clear definition of what an AI system is, e.g. using an established taxonomy such as in [54], as well as mechanisms to retrieve information on new AI systems and those already operating in the respective sector. For the latter, a regulatory body mandated with sectoral oversight could foresee a scheme of self-declaration (with penalties) or mandatory registration/ notification, similarly to requirements for cyber security incidents. On the other hand, an energy company operating several AI systems may rely instead on internal documentation, governance and

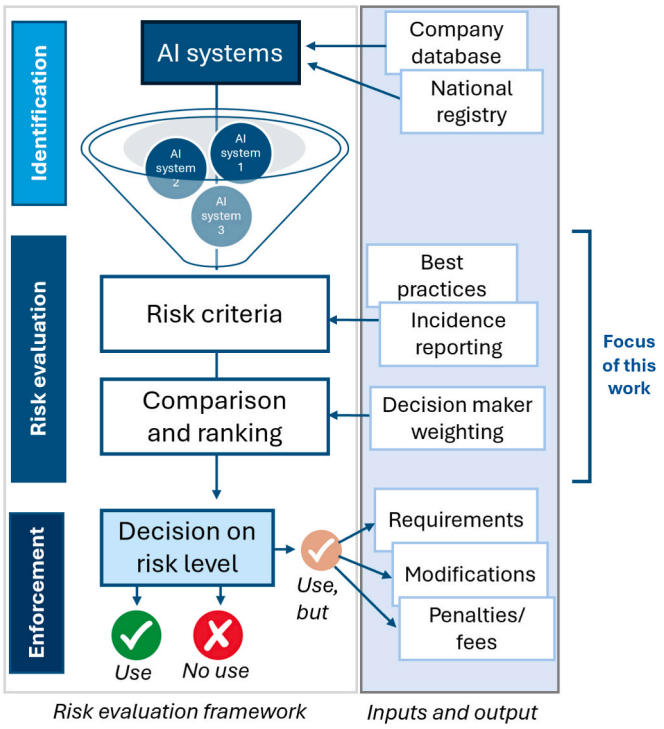

Fig. 2. AI risk evaluation framework.

registers that emerged from coordination or compliance requirements.

The second phase ("**Risk evaluation**"), which we address in this work, focusses on evaluating the risks of AI systems within the sector / business activity (in our case: energy system) by using established methods, e.g. from MCDA. The set of risk criteria, through which individual AI systems are compared, can be established through expert judgement or previous, reported incidents/malfunctions of AI and should be regularly updated (for example, at the beginning of each regulatory period, like electricity network tariffs). Criteria can receive different weights, reflecting the legal focus and national policy priorities. While we opted in our methodology for a design with binary choices (Y/N), more complex qualitative metrics as suggested in [55] could be tested and used to extend our approach in future work. The second phase concludes with a ranking of all evaluated AI systems, based on a scoring system and considering adjusted weighting considering regulator and business operators stated preferences derived by the Simos method [56].

Building on these outcomes, the third phase focuses on "**Enforcement**". In other words, at this step, the regulator or a business organization that conducts monitoring and oversight of AI utilization decides which requirements an entity that uses an AI system would need to fulfil, based on the estimated risk level (outcome of phase 2).

Our contribution aims at providing a rigorous, scientifically backed solution that both business operators and regulators may use, given that the outcomes of our framework (and the MCDA) result in scores that could be used to design risk categories or cut-off values/thresholds above which respective regulatory requirements may become applicable. Towards the end of this study, we briefly outline a potential implementation procedure for regulators and businesses. However, the focus of this paper has been placed on the methodological framework to evaluate AI systems in the energy sector.

3.2. Risk criteria

For the energy sector, as part of the critical infrastructure, we deem several risk criteria relevant: How privacy sensitive is the input data used in the AI systems? Could there be bias in the input data or (racial/social) discrimination? Is the AI model itself robust and explainable? Or does it dynamically re-train, for example constantly adjusting neural network weights or hyperparameters (which could complicate documentation and the tracking model compliance)? How critical is the task performed? Can failure in accomplishing the task supported by the AI systems result in general human harm, system or asset failure or economic damage? In addition, what is the data collection latency; especially if data is collected close to real time and directly fed in a dynamic/online manner into the model, as this can influence vulnerability towards false data injections through cyberattacks [14]. In addition, AI model complexity is a relevant aspect as this will determine user-centered aspects such as explainability and trustworthiness[57,58].

Eventually, is the AI system itself integrated in an automated business process (which challenges human oversight) or is there a gap between the derived model outputs and further use in business functions. Following this reasoning, we eventually derive a set of risk criteria (Table 1) that build on the major AI risk domains extracted from the MIT risk repository [45] and group them into three AI system components "Data", "Model", "Task & output". To reduce the workload on AI system operator and regulator's side, while complying with the principle of proportionality on risk evaluation [40], the seven risk criteria are defined as binary (Yes – applicable, No – not applicable).

3.3. Comparing AI systems using PROMETHEE II

Given that the number of AI systems in a specific sector might scale with the number of companies (i.e., there could be hundreds of AI systems operated within a national energy system [51]), there is a strong need for a transferable method that can simultaneously evaluate and compare multiple AI systems.

MCDA has been widely used for weighing and comparing different problems in the energy sector or AI development (e.g., compare [35,59]). However, no work has so far evaluated the various risks of AI systems in the energy industry. In this work we use the applied set of risk criteria (Table 1) as input to the Preference ranking organization method for enrichment evaluation (PROMETHEE), which is one of the most widely applied outranking MCDA methods [60]. This method has been successfully applied across diverse fields such as supply chain management, logistics, engineering, manufacturing, and marketing, making it well-suited for various decision problems [61].

PROMETHEE operates by eliciting decision-makers' preferences through pairwise comparisons of the performance of alternatives on each criterion. In doing so, it not only aggregates individual criterion evaluations but also incorporates a partial-compensation property. This property is crucial in our context: it ensures that a poor performance in one risk criterion cannot be entirely offset by an excellent performance in another. Such an approach is particularly important with binary scores (like Yes/No). With binary evaluations, the granularity of the data is limited, and the partial-compensation mechanism prevents a single favorable score from completely masking a critical deficiency in another domain.

PROMETHEE evaluates the performance of alternatives through pairwise comparisons. Consider a scenario where m alternatives $A = \{$

Table 1
Risk criteria for evaluating AI systems.

AI model component	Risk criteria
Input data	Human sensed
	Real-time/ streamed
	Privacy sensitive
AI model	Black box
	Dynamic update
	Automation
Task & output	Potential for harm / damage beyond individual / local extent

$a_1, a_2, \cdots, a_m\}$ are assessed across n criteria $G = \{g_1, g_2, \cdots, g_n\}$. For each criterion, every pair of alternatives is compared to determine their relative performance. PROMETHEE offers a variety of preference functions tailored to different data types. In our analysis, since the evaluations are based on binary relations, we employ a straightforward preference function, commonly known as the Type I (usual) criterion preference function (REF):

$$P_j(a_i, a_{i'}) = \begin{cases} 1, if a_i > a_{i'} \\ 0, if a_i \leq a_{i'} \end{cases} \quad (1)$$

where a_i and $a_{i'}$ are the alternatives in the alternative set. The Type I preference function is appropriate for the binary scoring scheme adopted in this study. The risk criteria are negatively oriented: if a risk is present, the system is assigned "Yes" and scored 0; if no risk is present, it is assigned "No" and scored 1. With this coding, the Type I function yields full preference when an alternative scored "No" is compared with one scored "Yes," and indifference otherwise. The function thus represents the intended strict preference on each binary criterion without introducing arbitrary gradations.

This function compares the performance of alternatives on each criterion j. The overall degree of preference of alternative a_i over alternative $a_{i'}$ is expressed as:

$$\pi(a_i, a_{i'}) = \sum_{j=1}^{n} P_j(a_i, a_{i'}) \bullet w_j \quad (2)$$

where w_j represents the weight assigned to criterion j, reflecting its relative importance. This equation quantifies how much a_i is preferred over $a_{i'}$ across all criteria. Subsequently, two outranking flows are calculated to capture the overall preferences among all alternatives. These flows are defined as follows:

$$\begin{cases} \phi^+(a_i) = \frac{1}{n-1} \sum_{a_{i'} \in A} \pi(a_i, a_{i'}), \\ \phi^-(a_i) = \frac{1}{n-1} \sum_{a_{i'} \in A} \pi(a_{i'}, a_i), \end{cases} \quad (3)$$

where the positive outranking flow quantifies the degree to which an alternative a_i outperforms all other alternatives, while the negative outranking flow indicates the extent to which it is outperformed by the others. The PROMETHEE family comprises several variants for different decision-making contexts. In our study, we apply PROMETHEE II, which yields a net outranking flow to establish the final ranking of AI systems:

$$\phi(a_i) = \phi^+(a_i) - \phi^-(a_i) \quad (4)$$

This net flow, ranging from −1 to 1, serves as an indicator of preference: a higher net flow signifies a more desirable alternative. Specifically, if an alternative's net flow is greater than 0, it is, on average, preferred over others; conversely, a net flow below 0 indicates that the alternative is generally outranked by its competitors, rendering it less favorable in the overall ranking [52].

In the following, the application of the proposed framework is demonstrated using six exemplary AI systems from the power system domain.

3.4. Case studies: AI systems in electricity transmission

We exemplify the application of our AI evaluation framework considering six AI systems used in electricity system operation and planning (three of those systems have been introduced in [41]). The choice of all six AI systems was derived from the authors' views on promising applications of AI for energy systems (focus on electricity transmission systems). In addition, our reasoning for the selected group of AI systems was to cover both planning (AI systems 2, 4 and 5) and operational tasks (AI systems 1, 3, 6), long-to-short-term horizons and include both actual, documented, real-world applications (e.g. AI systems 1, 4 and 5) and futuristic applications (e.g. AI systems 3 or 6). The main AI risk characteristics of the six systems are shown in Table 2.

AI system 1: Forecasting electricity demand

The first case study represents the Demand forecasting tool (DFT) from ENTSO-E, the European Network of Transmission System Operators (TSO) for Electricity [62]. Electricity demand forecasting is increasingly challenging for TSOs as traditional load patterns evolve under the adoption of new appliances and assets (e.g., electric vehicles, heat pumps, demand response etc.) [63,64]. On the other hand, new data sources, decreasing granularity and enhanced modelling allow for the development of spatially and temporally more accurate load forecasts [65,66]. DFT is a load forecasting tool utilized throughout all European TSOs, encompassing a wide range of studies from extensive network expansion plans to resource adequacy studies in the shorter term (weeks to months in advance). With very little human input, the DFT leverages historical load time series, weather conditions, and future scenario data to generate load time series with hourly resolution [48]. The AI model used in DFT involves methods of machine learning, such as singular value decomposition and multilinear regression. Data sources are not streamed in real-time and do not possess any data privacy implications from an individual perspective. However, output might be used in an automated way in various short-term studies.

AI system 2: Interpretation tool for scarcity events

The second use case represents an AI system exploiting a rather interpretable algorithm. Building on previous works on rule-based energy system analysis [67,68], the study of [69] focused on a CN2 rule-based approach that effectively identifies scarcity events from a vast collection of electricity market simulations. Scarcity events are (expected) situations where supply cannot meet demand in a particular transmission system area, resulting in unserved energy. In this work, a rule-mining algorithm is utilized to analyze the outputs of ENTSO-E's Pan-European electricity market model, which encompasses 700 model scenarios, each spanning 8760 hourly time steps. Utilized data sets do not contain personal data, the model is static and requires rather high degrees of human input (labelling of scarcity events and interpretation of identified rule sets).

AI system 3: Dynamic security assessments (DSA)

The third AI system is used for DSA, which usually refers to the analysis performed to ensure a power system meets specified reliability and security criteria in both transient (ms timescales) and steady-state (minutes) time frames for a list of credible contingencies [70]. From the operational perspective, power system operating variables need to remain within operating limits, including static considerations (e.g., voltage and thermal limits), dynamics, and stability. The use of AI to perform such a task can usually be posed either as a classification problem (e.g., safe/unsafe) or a regression problem, in which case a continuous variable is estimated (e.g., the critical clearing time). Other implementations could use models from deep learning [70], thus relying on rather less interpretable algorithms. Given the scope of the system, close to real-time operation with streamed data even in automated context is likely for such systems.

AI system 4: Connecting renewable power plants

The fourth AI system represents an internal tool used to handle increasingly time-consuming requests for connecting new renewable power plants at the transmission system level [71]. The aim of such AI system is to accelerate TSO's response time, using a database of historical requests, studies and documentation on the transmission grid infrastructure for the area where the request stems from. In addition, the generative AI tool (built on a large language model) may also incorporate current information on the territory of interest, such as areas labelled with priority in municipal or regional development plans. Given its current shape, the tool may use human sensed, privacy sensitive data (e.g. on power plant projects and ownership, priority ratings) from previous requests, but is in general a low-risk, organizational tool for internal planning, without potential harm or safety critical aspects

Table 2
Evaluation of AI systems in TSO business (risk – Y; no risk – N).

Risk domain	Code	Polarity	Criteria	AI system 1	AI system 2	AI system 3	AI system 4	AI system 5	AI system 6
Data	g_1	Negative	Human sensed	N	N	N	Y	N	N
	g_2	Negative	Real-time/ streamed	N	N	Y	N	Y	Y
	g_3	Negative	Privacy sensitive	N	N	N	Y	Y	Y
Model	g_4	Negative	Black box	N	N	Y	Y	Y	Y
	g_5	Negative	Dynamic update	N	N	Y	N	Y	Y
	g_6	Negative	Automation	Y	N	Y	N	Y	Y
Task	g_7	Negative	Potential for harm / damage beyond individual / local extent	Y	Y	Y	N	N	Y

related to power system operation.

AI system 5: Real-time energy policy evaluation

The fifth AI system represents a real-world application run by the Swiss Federal Office of Energy. The AI system was developed to support Switzerland's 2022/23 energy-saving campaign and promote transparency in electricity consumption. The AI system consists of a model that predicts electricity savings using real-time streamed and analyzed smart meter consumption data from approximately 10,000 profiles across different consumer groups [72]. The data originates from various Swiss electricity providers and contains privacy sensitive electricity consumption profiles as well as aggregated sociodemographic data and weather data. While the system produces energy savings estimates in an automated manner without human oversight, no direct harm or damage would be expected. Detailed documentation on this system currently in operation can be found in [73].

AI system 6: Automating power system control

This rather futuristic AI system addresses the task of power grid topology reconfiguration, a typical, real-world task TSOs conduct on a daily basis [74]. The goal of the tasks is to optimize the power flows over several days at a 5-minute time-resolution, considering injections into the grid (e.g., through renewable electricity production), electric loads and the transmission grid topology. The authors [74] describe several AI systems (including the use of Deep Q-Networks) that could be deployed for this task. The described AI system uses streamed data of the system to optimize the flows in the transmission grid without human intervention. While it is yet a rather experimental AI system, grid operation through autonomous agents is often not seen as unlikely [75,76]. Although full autonomy might be achieved only on a longer timescale, AI assistants who guide grid operators (as described in 42]) might become operational much sooner. Such a system would be highly automated, dynamically updated and using also privacy-sensitive consumer data, whereas malfunctioning would have direct effects on the regional/national economy and societal wellbeing.

In the following, we will assess and compare the chosen six AI systems with our methodological risk evaluation framework, deliberately limiting our analysis to a focused, flexible evaluation methodology for AI risks rather than a wider, comprehensive cross-country, cross-sectoral or intra-sectoral comparison of AI applications along the energy value chain.

4. Results and sensitivities

A comparison of all AI systems with regard to the seven selected risk criteria is shown in Table 2, displaying already larger differences between the analyzed AI systems (e.g. compare AI system 1 that does not fulfill many risk criteria against AI system 6, which can be already understood as one of the "riskiest" applications). The final performances with equal weighted criteria (BAU case) as determined by the PROMETHEE II method (using the implementation in [77]) have been reported in Table 3. Each numeric value in Table 3 represents the pairwise preference score for a given system on a given criterion, ranging from −1 to +1. A score close to 1 means the system consistently outperforms its competitors on that criterion. A score close to –1 means it is consistently outranked on that dimension. Positive criterion scores imply that, on average, the alternative is preferred over the others; negative scores imply the reverse. The row "Final net flow" in Table 3 for each system is then obtained by aggregating its positive and negative flows using Eq. (2)–Eq. (4). It should be noted, because every criterion is scored on a binary (Yes/No) scale, the normalized values increase in discrete steps of 0.2, so the resulting net-flow scores appear uniformly spaced and visually similar.

Results show that AI system 2 poses the least risk to power system operation, followed by AI systems 4 and 1. These AI systems can be used offline, have no dynamic data and model update and are situated on the planning side. On the other hand, AI systems 3 and 5 are ranked riskier (i.e. show a lower performance in PROMETHEE II). This can be explained through the systems' use of real-time, streamed data and rather black-box models, where model outputs are used in an automated manner.

The riskiest AI system is the one where the whole control of a power system would be handed over, in a fully automated context (AI system 6). Such an AI system would pose an exceeding risk, compared to the other AI systems, originating from the little latencies between prediction, model outcomes and its direct use in highly critical tasks of real-time system operation. Hence, it is, as a very risky application, probably hard to deploy in the energy industry in the future. However, it may become more feasible in the form of an AI assistant, helping with the operator's workload and taking faster decisions (as described in [78]).

In our study, we initially applied equal weights across all evaluation criteria (base scenario), to maintain objectivity and avoid subjective bias. Because stakeholder priorities legitimately differ, we then exam-

Table 3
Output of the PROMETHEE II procedure.

Criteria	Code	AI system 1	AI system 2	AI system 3	AI system 4	AI system 5	AI system 6
Final net flow	–	0.2857	0.4571	−0.2286	0.1143	−0.2286	−0.4000
Human sensed	g_1	0.2000	0.2000	0.2000	−1.0000	0.2000	0.2000
Real-time/ streamed	g_2	0.6000	0.6000	−0.6000	0.6000	−0.6000	−0.6000
Privacy sensitive	g_3	0.6000	0.6000	0.6000	−0.6000	−0.6000	−0.6000
Black box	g_4	0.8000	0.8000	−0.4000	−0.4000	−0.4000	−0.4000
Dynamic update	g_5	0.6000	0.6000	−0.6000	0.6000	−0.6000	−0.6000
Automation	g_6	−0.4000	0.8000	−0.4000	0.8000	−0.4000	−0.4000
Potential for harm / damage beyond individual / local extent	g_7	−0.4000	−0.4000	−0.4000	0.8000	0.8000	−0.4000

ined whether alternative weight structures could change the preferred AI system. In particular, policy and economic perspectives often diverge. Accordingly, we elicited criterion weights from six experts, using the Revised Simos procedure (Table 4), and applied these weights, together with their uncertainty, through the PROMETHEE II analysis alongside the equal-weight baseline to stress-test the stability of the ranking. We adopted the Revised Simos procedure because it converts simple ordinal judgements into weights through transparent steps [56]. Experts perform a deck-of-cards exercise: they receive one card per criterion plus a set of blank cards and sort the criterion cards from least to most important, e.g., from left to right. Concretely, let the expert's ranking induce ordered sets $S_1 \prec \cdots \prec S_h \prec \cdots \prec S_H$. Criteria perceived as equally important are stacked in the same position, forming *ex aequo* sets. To express larger perceived gaps between two adjacent sets, the expert inserts $e_h \in \{1, 2, 3\cdots\}$ blank cards after set h. Finally, the expert specifies a global contrast ratio $z > 1$, interpreted as "how many times more important the most-important group is than the least-important group."

Assume H is the number of ranks in which criteria are ordered, let e'_h be the number of blank cards between the ranks h and $h+1$. The following set of constraints is defined:

$$\begin{cases} e_h = e'_h + 1, \forall h = 1, \cdots, H-1 \\ e = \sum_{h=1}^{H-1} e_h \\ u = \dfrac{z-1}{e} \end{cases} . \tag{5}$$

Then, non-normalized weight $k(h)$ of each criterion in rank h is then calculated as:

$$k(h) = 1 + u(e_0 + \ldots + e_{H-1}) \text{with} e_0 = 0 \tag{6}$$

The normalized weight of criterion g_i can be derived as:

$$w_i = \frac{k(h_i)}{\sum_{i=1}^{n} k(h_i)} \tag{7}$$

In Equation 8, h_i denotes the rank of criterion g_i. The criteria weights were solicited during interviews with six domain experts (energy industry, energy policy): Expert 1 is the research director of a large international corporation specializing in energy management and automation solutions, whereas Expert 2 has over five years of experience in public administration leading several initiatives as a national domain expert on the use of AI and future regulation in the energy sector. Experts 3 and 4 are early career professionals, each with roughly 2 year appointments in energy industry, whereas Expert 5 and 6 work on energy policy and strategy related questions, with 3 and 1 years of experience respectively. The interviews took place between November 2025 and March 2026, and the following orders of importance have been obtained according to the DM, also considering the positioning of blank cards between the criteria, and the z value:

Expert 1 : $L_1 = \{g_5\},\ e_1 = 0,\ L_2 = \{g_6\},\ e_2 = 0,\ L_3 = \{g_2, g_4\},\ e_3 = 1,$
$L_4 = \{g_1, g_3\},\ e_4 = 0,\ L_5 = \{g_7\},\ z = 3.$

Expert 2 : $L_1 = \{g_4, g_5\}, e_1 = 1, L_2 = \{g_1, g_3\}, e_2 = 1,$
$L_3 = \{g_2, g_6\}, e_3 = 1, L_4 = \{g_7\}, z = 10,$

Expert 3 : $L_1 = \{g_1\}, e_1 = 0, L_2 = \{g_4, g_7\}, e_2 = 0, L_3 = \{g_5\}, e_3 = 0,$
$L_4 = \{g_2\}, e_4 = 1, L_5 = \{g_6\}, e_5 = 0, L_6 = \{g_7\}, z = 7,$

Expert 4 : $L_1 = \{g_1\}, e_1 = 0, L_2 = \{g_2, g_4\}, e_2 = 1, L_3 = \{g_3, g_7\}, e_3 = 0,$
$L_4 = \{g_6, g_7\}, z = 10,$

Expert 5 : $L_1 = \{g_2\}, e_1 = 0, L_2 = \{g_5\}, e_2 = 0, L_3 = \{g_4, g_6\}, e_3 = 1,$
$L_4 = \{g_1\}, e_4 = 0, L_5 = \{g_3\}, e_5 = 0, L_6 = \{g_7\}, z = 5,$

Expert 6 : $L_1 = \{g_2, g_5\}, e_1 = 1, L_2 = \{g_4\}, e_2 = 1, L_3 = \{g_6\}, e_3 = 2,$
$L_4 = \{g_1, g_3\}, e_3 = 0, L_5 = \{g_7\}, z = 10,$

where L_h presents the set of criteria for the h rank. Based on the elicited inputs, Table 4 presents the criterion weights derived from the revised Simos method. Compared with the equal-weight baseline of 0.14 for each criterion, the expert-specific profiles show a clear common tendency together with substantial variation. In particular, g_7("Potential for harm/damage beyond individual/local extent") is assigned the highest weight by five of the six experts and remains highly weighted by the sixth, indicating broad agreement on its importance.

Beyond this common pattern, the profiles differ markedly. Experts 1, 5, and 6 emphasize g_1 and g_3; Expert 2 gives strong importance to g_2, g_6, and especially g_7; Expert 3 prioritizes g_6 and g_7; and Expert 4 places the greatest emphasis on g_5 and g_6. In contrast, g_4 and g_5 often receive relatively low weights, although not consistently across all experts. These results highlight the diversity of stakeholder perspectives and provide the basis for the subsequent robustness analysis of the PROMETHEE II ranking across alternative weighting schemes. For implementation, though, it remains an interesting question how stakeholders will become selected and interviewed; in other words, which organization is eventually responsible for selecting (energy sector) stakeholders and eliciting their preferences on AI risks.

Based on the six expert-derived weight sets, we recalculated the PROMETHEE II net flows for all AI systems. Fig. 3 shows the resulting distribution of net flow scores across these weighting scenarios (6 experts + BAU). The variation in criterion weights reflects differences in stakeholder perspectives, such as those of business operators and policymakers, and may also capture differences in regulatory priorities across AI governance frameworks [29,48].

The results indicate that the overall pattern is robust at both extremes, while some middle positions remain sensitive to the weighting perspective (compare Fig. 3). To further assess the robustness of the ranking, we defined an interval for each criterion weight using the minimum and maximum values observed across all experts and BAU scenarios. We then performed a Monte Carlo simulation by repeatedly sampling weight sets within these intervals and recalculating the PROMETHEE II ranking. Fig. 4 reports the percentage of simulation runs in which each AI system attains each rank position. The results confirm that the ranking is highly robust at the extremes. AI system 2 is ranked first in almost all simulations and only rarely appears in second place, while AI system 6 is ranked last in all simulation runs. At the lower end of the ranking, AI system 5 is most frequently placed fourth, with a smaller proportion of fifth-place outcomes, whereas AI system 3 shows the opposite pattern and is predominantly ranked fifth, with occasional fourth-place positions. The main uncertainty concerns the middle of the ranking. AI system 4 is most often ranked second, but it also frequently appears in third place and only marginally reaches first place. AI system 1 alternates between second and third place, with third place occurring more often. Overall, the Monte Carlo analysis indicates that the best and worst alternatives are highly stable, while rank uncertainty is concentrated mainly between the second and third positions and, to a lesser extent, between the fourth and fifth positions.

Our proposed evaluation framework is straightforward, robust and

Table 4
The base case criteria weighs and weights elicited by six experts.

Criterion	Weight						
	Base scenario	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
g_1	0.1429	0.1831	0.1176	0.0400	0.0233	0.1803	0.2302
g_2	0.1429	0.1268	0.2059	0.1600	0.0756	0.0492	0.0264
g_3	0.1429	0.1831	0.1176	0.0800	0.1802	0.2131	0.2302
g_4	0.1429	0.1268	0.0294	0.0800	0.0756	0.1148	0.0943
g_5	0.1429	0.0704	0.0294	0.1200	0.2326	0.0820	0.0264
g_6	0.1429	0.0986	0.2059	0.2400	0.2326	0.1148	0.1283
g_7	0.1429	0.2113	0.2941	0.2800	0.1802	0.2459	0.2642

Monte Carlo rank-position percentages for the six AI systems under interval-based criterion-weight uncertainty.

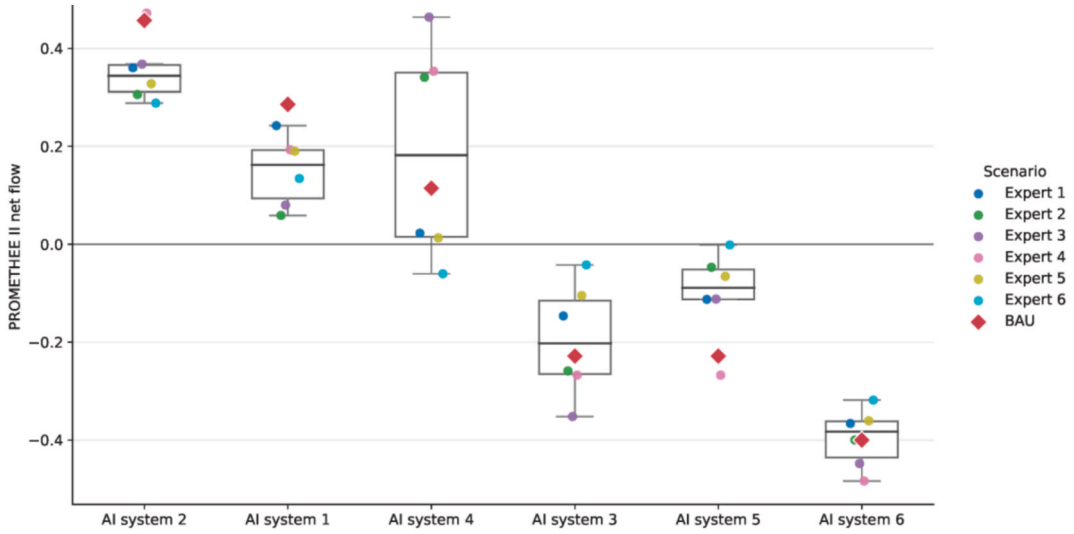

Fig. 3. PROMOTHEE II net flows spread under alternative weight scenarios.

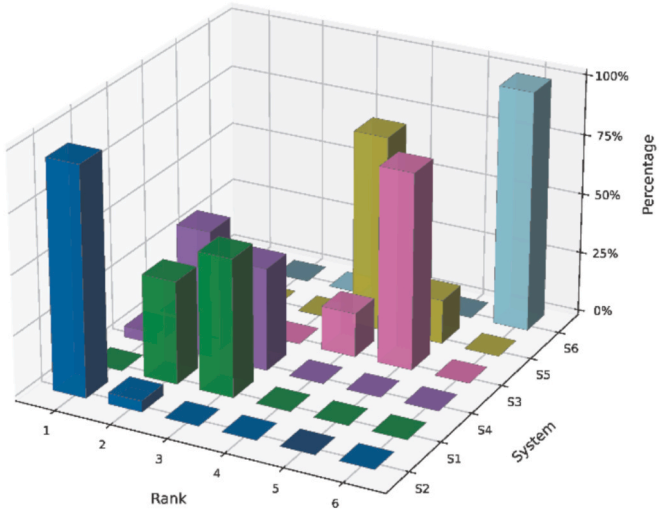

Fig. 4. Monte Carlo rank-position percentages (vertical axis) for the six AI systems under interval-based criterion-weight uncertainty.

with a high degree of transferability, so that it can be applied in practice by any public administration or AI system operator/user with limited background in MCDA or a shallow technical understanding of AI systems themselves. This "ease-of-use" is in our view fundamental, as the evaluation of hundreds to thousands of AI systems per sector and country[1] may be required in practice, through organizations or market participants with limited technical capacity, constrained human resources or lack of experience.

For example, users of AI systems within a company or regulatory body may have limited information about the functioning (e.g., algorithms) of an AI system, as these could have been developed by a third-party provider or another business division. In addition, many upcoming regulations may require documentation on how AI systems are used. Thus, one requirement is also to establish a flexible framework that allows the documentation of the main characteristics of AI systems and analysis of the applications' compliance.

5. Implementation and future development

5.1. Implementation of the proposed risk evaluation

In the following, we briefly discuss how our proposed methodology could be used in practice, using the European AI Act as an example. As such, we will outline a potential implementation of the developed AI risk evaluation methodology both for regulatory oversight and business operations (Fig. 5), followed by a few observed challenges and envisioned future developments.

[1] For example, if each Swiss electricity market participant (>600) would use 5–10 AI systems, up to 3,000–6,000 AI systems would need to be compared on national level).

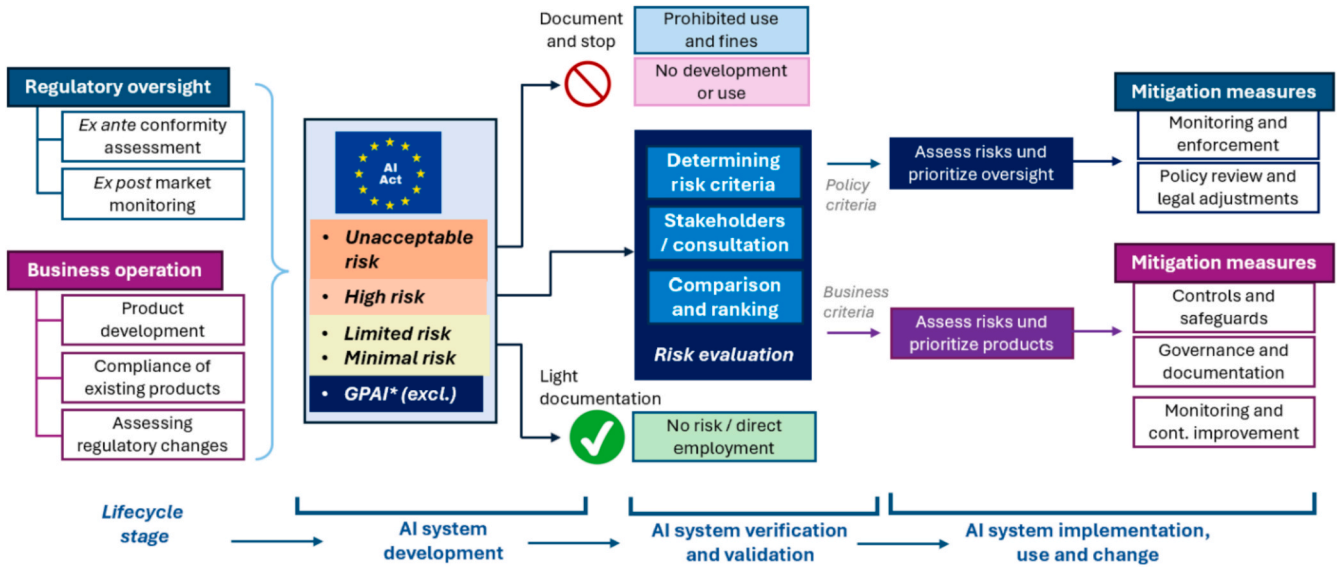

Fig. 5. Potential use of methodology for regulatory oversight and business operations, for example under EU AI Act.

Regulatory authorities. As foreseen for example under the EU AI Act [51], regulators might begin by conducting an *ex-ante* conformity assessment, ensuring that AI systems meet legal and ethical requirements before they are deployed. Once the AI system is on the market, regulators perform *ex-post* market monitoring of its use, reporting potential and risks from real-world performance (within the energy sector). If, after an initial screening, an AI system is found to be prohibited, regulators document the findings, and the system is not certified for use. If it is otherwise found that the AI system poses no risk, the system may also be documented without further regulatory action.

During the risk evaluation phase, regulators assess the level of risk associated with each AI system and prioritize oversight within the total pool of AI systems accordingly. Focus is placed on higher-risk systems. For systems that proceed into active use, regulators continue with monitoring and enforcement to ensure ongoing compliance. Insights from these activities, particularly on new or emerging risks and even unknown benefits, feed into policy reviews and legal adjustments, informing future regulatory revisions.

Most likely, regulatory authorities will rely on self-declarations of energy companies that deploy AI systems (similar to cyber security measures [20,79]). To mitigate the risk of this information asymmetry and potential under-reporting, some mitigative measures such as independent oversight mechanisms (third-party conformity assessments, standardized documentation requirements, and post-market monitoring obligations) could be pursued. For example, the EU AI Act foresees external conformity assessments prior deployment and risk management and incident reporting ([26]), while frameworks like the NIST AI Risk Management Framework encourage users of AI system to adhere to standardized transparency and documentation practices [17]. Together, these measures aim to reduce the degree of information asymmetry by enabling external verification and continuous regulatory oversight beyond developer self-reporting.

Business operators. On the side of an energy company that intends to use a self-built or procured AI system, the process starts with the definition of product requirements from the business side, where companies design and build AI systems while incorporating compliance and ethical standards. They also test/simulate the compliance of existing products under possible, regulatory changes that may affect current or future AI systems operated in the company.

Thus, the AI system development/procurement will be accompanied with internal risk assessments and compliance checks; if a system is eventually deemed too risky or prohibited, the company may stop the development or use of the said AI system and document the decision. In the risk evaluation phase, while using our proposed methodology, businesses would assess and prioritize their AI systems, identifying the ones which would require enhanced governance, mitigative measures or additional safeguards.

Once AI systems are implemented, companies then apply controls and mitigation measures to manage identified risks (internal controlling). The overall aim in this phase of AI system implementation is to maintain robust governance and documentation to prove accountability if needed and facilitate internal and external audits. Finally, through ongoing monitoring and continuous improvement, businesses may refine and adapt their operated AI systems, ensuring they remain safe, compliant, and effective throughout the overall lifecycle.

5.2. *Limitations and future work*

While the presented methodology is certainly a first step to systematically assess and compare AI-related risks in energy systems, it could benefit from further extensions that address the identified limitations.

The presented evaluation framework and its potential implementation in regulatory agencies and businesses can be further developed and refined through the addition of more techno-economic detail in terms of risks and potential economic or human harm. Currently, the number of criteria is limited (7) and criteria itself are only qualitative, where all risks being equally weighted (base scenario) with weighting variants from a couple of domain experts which aim at reflecting energy policy and energy business considerations (and its variance across different organizations). While the adverse effects on energy systems, and eventually, society might strongly vary across the set of criteria and across geographies (such as the Value of Loss of Load [80] of energy from blackouts), future work could further dive into system-specific weighting of AI risks Violating data privacy or the failure of an AI system with streamed data in the human resources department will pose less serious threats to society than a black-box model in power system operation with a high vulnerability towards cyber-attacks. While the first system could cause a violation of data privacy or employee discrimination, the latter may lead to an hour-long blackout. Some of these aspects are addressed through the emerging concept of AI integrity, whose further development might also bring more holistic mitigation measures towards intertwined AI risks [81].

Furthermore, future extensions should include more graded evaluations (e.g. grades 1–5) for a subset or all criteria, for example similar the

approach presented in [33] For example criteria such as "Automation" could differentiate different autonomy levels that are already used to describe different human–machine configurations in power systems [75] and policymaking [82]. On the other hand, data-related criteria could be mapped to different data products used in the energy system or emerging ontologies [83,84]. When experts are uncertain about assigning precise scores, a fuzzy extension of PROMETHEE can be applied. For instance, instead of selecting a single value for a risk level, they may express their assessment using a triangular fuzzy number [85]. Uncertainty can also be considered in the elicited criterion weights, since the information provided by experts is inherently subjective and may itself be imprecise. In this context, the revised Simos card procedure can be incorporated into an optimization model that explicitly accounts for uncertainty in the inputs provided by experts, such as the number of blank cards and the z-value [86]. However, all these extensions beyond binary criteria will require more experiences and observations with realized risks through AI systems (on the methodology side) and trained AI risk assessment personel with clear documentation manuals with little space for misinterpretations or subjective judgment (on the implementation side).

Policy and decision makers may further differentiate the requirements that AI systems in different risk categories must comply with. Linked to this analysis is the question of the costs of regulation, contrasting to expected benefits using AI systems. It has been stated that the European AI Act will directly affect the use of AI in the electricity industry, as well as companies developing such AI systems [48,87]. A recent study [51] covering the electricity systems of Austria, Greece and Switzerland found that regulatory requirements may constrain the use of so-called «high-risk» AI systems, resulting in non-neglectable administrative burdens (e.g., documentation, human oversight, certification). The study estimated aggregated compliance costs of up to 200 million Euros per country, depending on AI utilization and different risk scenarios. A comprehensive study over a full jurisdiction (or country) could shed more light on the cost-benefit analysis of AI across various applications and subsectors.

In addition, future research may investigate further details on the full implementation and enforcement phase of the proposed methodology. Relevant open questions are, for example, under which paradigms and within what type of stakeholder processes will regulatory agencies determine their preferences towards risks. This includes a transparent procedure to identify high-risk AI systems and separate those from less-risky AI systems. In the European AI Act, it is stated that a high-risk AI system is one that is part of a "safety critical component", which would roughly translate to a subset of our AI risk criteria (g_2, g_6, $g_7 = Y$). In our current framework, we limit the criteria to binary (Y/N) states because as we defend that its wide and transferable use requires clarity and the smallest possible room for subjectivism, as self-declarations on the user/developer level and bottom-up evaluations might precede any wider analysis.

Given the potentially large number of AI systems in a particular sector, will regulators need to use percentiles to address the highest ranked AI systems through regulatory measures such as documentation requirements and human oversight? Or will *ex ante* determined, fixed cut-off values be used to draw the line between AI systems subject to measures and those deemed less critical, with less or no additional requirements to fill (if compared to AI systems in non-critical sectors)? Modelling oversight thresholds is an important next step, which could be illustrated through a worked example applying various thresholding methods (e.g., quantiles, absolute cut-offs, or clustering). Such thresholds could be used to categorize systems across weight sets derived from pairwise comparison and link these to regulatory actions, which would necessarily require a larger, well documented sample of AI systems operating in the energy industry, which is currently lacking. In addition, in this work we intentionally avoided fixed universal cut-offs, as these should be defined case by case depending on regulatory and business monitoring capacities. At this stage, the framework is designed to identify the most critical systems rather than to definitively classify them.

Ultimately, it remains fundamental to assess how risk mitigation measures such as risk management and human oversight over critical AI systems could be effectively monitored, given that policy makers will most likely rely on information from self-declaration of AI system operators. The latter may favor not declaring all risk-carrying applications, given the regulatory burden and costs these add to business operations. Besides, utilities may simply lack full oversight over the decentralized use (and development) of AI systems within a larger organization. Thus, AI regulations might require monitoring and testing of AI systems, a research stream has recently received more attention (e.g. [88]).

6. Conclusions and outlook

AI is transforming the planning and operation of energy and power systems. Harnessing its benefits is however not risk-free. While the energy industry is often described as risk-focussed, there currently exist no methods to assess risks stemming from the wide use of AI in energy systems. In this context, the present study provides a practicable, MCDA-based frameworks to compare and rank AI systems with regard to the risk they add to energy systems. Exemplifying the use of this framework on six realistic examples of AI systems applied to operational and planning tasks in electricity transmission, we found that all analyzed AI systems introduce new risks to the energy sector, but to different extents. Using a sensitivity analysis that resembles different, realistic regulatory /policy and business priorities, we further showed how risks are accentuated for specific systems, across all stated preferences. While the analyzed AI systems tried to cover different tasks and a variety of time horizons, it should be remembered that our assessment relies on external documentation and assumptions on their usage. This is not a limitation as under both implementation schemes, the evaluation of AI systems will rely on bottom-up data, from AI system developers and users. In that sense, the proposed assessment was not intended to indicate in which tasks AI systems are less risky to use today, but rather showcases how heterogeneous AI applications can be evaluated consistently.

Future work should aim towards improved quantification of risks and potential damage through AI systems. In addition, more research will be needed to shed some light on the current and future use of AI in the energy industry beyond isolated case studies and on effective regulatory benefit exploitation and risk mitigation strategies. Furthermore, we currently assume that all AI systems are known. However, new developments (and the "retirement/discontinuation" of older AI systems will require adjusting the methodological framework towards a dynamically changing pool of AI systems under scrutiny. This should be reflected extending our work with approaches that allow for benchmarking all analyzed AI systems against an optimal solution ("zero risk") or replacing the MCDA-core with some distance measure (from an ideal solution). Eventually, the implementation of risk-focussed AI regulations or business operations will require an agreement on transparent and undisputable rules that can be used to locate AI systems either within or beyond the space subject to more extensive regulatory or mitigative measures.

CRediT authorship contribution statement

Fabian Heymann: Writing – review & editing, Writing – original draft, Project administration, Investigation, Formal analysis, Conceptualization. **River Huang:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Antoine Marot:** Validation, Formal analysis, Data curation, Conceptualization. **Medha Subramanian:** Writing – original draft. **Russell McKenna:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the co-financing provided by the Swiss National Science Foundation (SNSF) under the project Locational Optimization of Data Centers in Multi-Energy Systems (LOCO-DC), Grant No. 10.006.283. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions, as well as their colleagues Vivian Chen, Franziska Schmidt, Wendy Reyes, and Tania Lopez Garcia for their support during the weight elicitation process using the revised SIMOS method. Finally, the first author would like to thank his beloved Cedric, Olivia and Elvis, for reminding him day-to-day of the real-world relevance of risk evaluation, regulation and oversight.

Annex 1 – AI risk criteria and decision rules.

AI risk criterion	Definition	Decision rule (Y/N)	Examples (including boundary cases)
Human-sensed data	AI system processes inputs originating from human observations, reports, or behavior (rather than purely physical sensors or system KPI).	Y if any operational model input originates from humans (operators, customers, field staff, crowdsourced reports). N if inputs are exclusively machine-generated telemetry or simulations.	Y: AI system with model for outage prediction using customer calls that can linekd to entries in own database. Boundary Y: AI system for situational awareness using social media reports. N: Fault detection using only phasor measurement units; load forecasting based exclusively on metering and weather data.Boundary N: AI system trained on manually annotated data by system operator.
Real-time / streamed data	The AI system processes instantaneously arriving data and produces operational decision.	Y if inference occurs on live or near-real-time and outputs are used for operational actions. N if the model runs offline or in delayed execution mode for planning / analysis.	Y: Real-time stability monitoring AI system using phasor measurement units. Boundary Y: Grid state estimation updating every few seconds. N: AI system used for annual grid expansion planning or long-term asset maintenance prediction. Boundary N: Hourly performed load forecast for analyzed system.
Privacy-sensitive data	The AI system processes data that could identify individuals or reveal personal behavior patterns.	Y if inputs contain personally identifiable information or fine-grained consumption patterns traceable to households or individuals. N if data are aggregated, anonymized, or infrastructure-only.	Y: AI system is used to analyze household smart-meter data to infer occupancy patterns. Boundary Y: AI system using 15-minute household smart meter data. N: AI system used for transmission asset monitoring. Boundary N: AI system processing feeder-level aggregated demand across hundreds of customers.
Black-box model	AI system's model cannot be readily interpreted or audited by domain experts based on results without further analysis or computations.	Y if the AI system's model relies on non-interpretable algorithms (e.g., deep neural networks) and explanations do not reveal decision logic. N if the AI system's model structure and reasoning are transparent or interpretable.	Y: AI system deploying deep neural networks to predict cascading failure risk.Boundary Y: AI system runs with deep neural networks that allow for post-hoc explainability only. N: AI system using linear regression to perform load forecasting or decision-tree-based fault classification. Boundary N: AI system using gradient boosting with clear feature importance.
Dynamic model update	The system changes parameters or decision logic after deployment without a human validation.	Y if the AI system's model learns online, retrains automatically, or adapts parameters autonomously after deployment. N if updates occur only through manual retraining and controlled redeployment.	Y: Adaptive load forecasting model is updating continuously from new data streams.Boundary Y: AI system incorporates continuous online learning from operational events, e.g. through streamed data. N: Asset failure prediction model retrained annually by engineers.Boundary N: AI system is monthly retrained with human approval.
Automation	Extent to which AI outputs directly trigger operational actions without human intervention.	Y if AI system initiates or executes operational actions automatically (closed-loop control). N if the system provides decision support and requires human approval before action.	Y: AI system autonomously adjusts inverter reactive power setpoints in real time.Boundary Y: AI performs autonomously system operations, such as system balancing. N: Operator decision-support dashboard recommending switching actions.Boundary N: AI system infers recommendations for system operation, but requiring operator confirmation
Potential harm beyond local extent	Malfunctioning of the AI system propagates beyond a local perimeter and produces system failure or causes other larger, regional or societal impacts.	Y if AI system's erroneous outputs could affect large energy/power system areas, many customers, or affect other critical infrastructure. N if AI system- caused impacts remain localized and quickly reversible.	Y: AI system is used to coordinate transmission switching which affects regional grid stability.Boundary Y: AI system's failure affects whole transmission-level dispatch. N: AI system is used to optimize building-level heating, cooling and air ventilation.Boundary N: AI system optimizes local microgrid whose failure would affect only one facility.

Data availability

Data will be made available on request.

References

[1] IEA, Digitalization & Energy, Paris, France, 2017. doi: 10.1787/9789264286276-en.

[2] Heymann F, Milojevic T, Covatariu A, Verma P. Digitalization in decarbonizing electricity systems – Phenomna, regional aspects, stakeholders, use cases, challenges and policy options. Energy Jan. 2023;262. https://doi.org/10.1016/j.energy.2022.125521.

[3] H. Agarwall, C. P. Das, and R. K. Swain, Does Artificial Intelligence Influence the Operational Performance of Companies? A Study, Proceedings of the 2nd International Conference on Sustainability and Equity (ICSE-2021), vol. 2, pp. 1–11, 2022.

[4] D. Czarnitzki, G. P. Fernández, C. Rammer, and K. U. Leuven, Artificial Intelligence and Firm-Level Productivity Artificial Intelligence and Firm-level Productivity, ZEW Discussion paper, pp. 1–38, 2022.

[5] McKinsey Analytics, The State of AI in 2021, 2021. [Online]. Available: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021.

[6] A. Misra, J. Wang, S. Mccullers, K. White, and J. L. Ferres, Measuring AI Diffusion: A Population-Normalized Metric for Tracking Global AI Usage, Nov. 2025.

[7] T. Ahmad et al., Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities, Mar. 20, 2021, Elsevier Ltd. doi: 10.1016/j.jclepro.2021.125834.

[8] F. Heymann, H. Quest, T. Lopez Garcia, C. Ballif, and M. Galus, Reviewing 40 years of artificial intelligence applied to power systems – A taxonomic perspective, Energy and AI, vol. 15, p. 100322, Jan. 2024, doi: 10.1016/j.egyai.2023.100322.

[9] WEF, Harnessing Artificial Intelligence to Accelerate the Energy Transition, 2021.

[10] R. V. Yohanandhan et al., A specialized review on outlook of future Cyber-Physical Power System (CPPS) testbeds for securing electric power grid, Mar. 01, 2022, Elsevier Ltd. doi: 10.1016/j.ijepes.2021.107720.

[11] Isfakul Anam M, Nguyen TT, Vu T. Risk-based preventive energy management for resilient microgrids. International Journal of Electrical Power and Energy Systems 2023;154:Dec. https://doi.org/10.1016/j.ijepes.2023.109391.

[12] Liu M, Wu FF. Risk management in a competitive electricity market. International Journal of Electrical Power and Energy Systems Nov. 2007;29(9):690–7. https://doi.org/10.1016/j.ijepes.2007.05.003.

[13] Mokarram M, Shafie-khah M, Aghaei J. Risk-based multi-criteria decision analysis of gas power plants placement in semi-arid regions. Energy Rep 2021;7:3362–72. https://doi.org/10.1016/j.egyr.2021.05.071.

[14] Hug G, Giampapa JA. Vulnerability assessment of AC state estimation with respect to false data injection cyber-attacks. IEEE Trans Smart Grid 2012;3(3):1362–70. https://doi.org/10.1109/TSG.2012.2195338.

[15] B. Dattner, T. Chamorro-Premuzic, R. Buchband, and L. Schettler, Hiring And Recruitment The Legal and Ethical Implications of Using AI in Hiring, Harv. Bus. Rev., vol. April, pp. 1–10, 2019, [Online]. Available: https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-...

[16] Nelson GS. Bias in Artificial Intelligence. N C Med J Jul. 2019;80(4):220–2. https://doi.org/10.18043/ncm.80.4.220.

[17] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), Jan. 2023. doi: 10.6028/NIST.AI.100-1.

[18] Körner MF, Sedlmeir J, Weibelzahl M, Fridgen G, Heine M, Neumann C. Systemic risks in electricity systems: a perspective on the potential of digital technologies. Energy Policy May 2022;164. https://doi.org/10.1016/j.enpol.2022.112901.

[19] Belkhir L, Elmeligi A. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. J Clean Prod 2018;177:448–63. https://doi.org/10.1016/j.jclepro.2017.12.239.

[20] Heymann F, Henry S, Galus M. Cybersecurity and resilience in the Swiss electricity sector: status and policy options. Util Policy Dec. 2022;79:101432. https://doi.org/10.1016/j.jup.2022.101432.

[21] de Vries A. The growing energy footprint of artificial intelligence. Joule Oct. 2023;7(10):2191–4. https://doi.org/10.1016/j.joule.2023.09.004.

[22] IEA, Enhancing Cyber Resilience in Electricity Systems, Paris, France, 2021.

[23] Digital Future Society, Algorithms in the public sector: four case studies of ADMS in Spain, 2023.

[24] R. Machlev et al., Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities, Aug. 01, 2022, Elsevier B.V. doi: 10.1016/j.egyai.2022.100169.

[25] LNE, Certification of processes for AI, https://www.lne.fr/en/service/certification/certification-processes-ai.

[26] C. Novelli, F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi, AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act, Digital Society, vol. 3, no. 1, May 2024, doi: 10.1007/s44206-024-00095-1.

[27] ISO/IEC, AI Risk Assessments Under ISO/IEC 42001: A Practical Guide, Mar. 2025.

[28] Radu R. Steering the governance of artificial intelligence: national strategies in perspective. Policy Soc 2021;40(2):178–93. https://doi.org/10.1080/14494035.2021.1929728.

[29] H. Roberts, M. Ziosi, C. Osborne, and L. Saouma, A Comparative Framework for AI Regulatory Policy, Montreal, Feb. 2023.

[30] Roberts H, Cowls J, Morley J, Taddeo M, Wang V, Floridi L. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. AI & Soc Mar. 2021;36(1):59–77. https://doi.org/10.1007/s00146-020-00992-2.

[31] B. B. A. C. R. S. Chakravorti, Charting the Emerging Geography of AI, Harv. Bus. Rev., pp. 1–8, Dec. 2023.

[32] Schmid T, Hildesheim W, Holoyad T, Schumacher K. The AI methods, capabilities and criticality grid: a three-dimensional classification scheme for artificial intelligence applications. KI - Kunstliche Intelligenz Nov. 2021;35(3–4):425–40. https://doi.org/10.1007/s13218-021-00736-4.

[33] OECD, OECD Framework for the classification of AI systems, 2022. [Online]. Available: www.oecd.ai/wips.

[34] Miranda V, Proenga LM. Probabilistic choice vs. risk analysis - Conflicts and synthesis in power system planning. IEEE Trans Power Syst 1998;13(3):1038–43.

[35] Linares P. Multiple criteria decision-making and risk analysis as risk management tools for power systems planning. IEEE Trans Power Syst 2002;22(7):54. https://doi.org/10.1109/MPER.2002.4312382.

[36] Reiz C, Gouveia C, Bessa RJ, Lopes JP, Kezunovic M. Risk assessment of future power systems: Assuring resilience of electrification for decarbonization. Sustainable Energy Grids and Networks Sep. 2025;43. https://doi.org/10.1016/j.segan.2025.101849.

[37] Quest H, et al. A 3D indicator for guiding AI applications in the energy sector. Energy AI Aug. 2022;9. https://doi.org/10.1016/j.egyai.2022.100167.

[38] ISO/IEC, Artificial Intelligence (AI) — Assessment of the robustness of neural networks, Geneva (CH), Mar. 2021.

[39] European Commission, Better Regulation Toolbox, Jul. 2023.

[40] Mougan C, et al. The science and practice of proportionality in AI risk evaluations. Science Feb. 2026;391(6787):769–71. https://doi.org/10.1126/science.aea3835.

[41] F. Heymann, A. Marot, M. Subramanian, and M. Galus, Risk Evaluation of AI Systems in the Energy Sector - Three Case Studies from TSO Business, in 2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI), IEEE, Sep. 2024, pp. 172–177. doi: 10.1109/RTSI61910.2024.10761769.

[42] Martínez-Plumed F, Gómez E, Hernández-Orallo J. Futures of artificial intelligence through technology readiness levels. Telematics Inform May 2021;58. https://doi.org/10.1016/j.tele.2020.101525.

[43] J. Schuett, Risk Management in the Artificial Intelligence Act, European Journal of Risk Regulation, pp. 1–19, Mar. 2023, doi: 10.1111/rego.12094.

[44] MIT, The AI Risk Repository, https://airisk.mit.edu/.

[45] P. Slattery et al., The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence, Aug. 2024.

[46] Commission E. Proposal for a Regulation laying down harmonised rules on artificial intelligence (AI Act). Brussels: Belgium; 2021.

[47] Li G. Potential Impacts of European AI Regulation on the American Energy Sector. Florida (US): Gaineswille; 2020.

[48] F. Heymann, K. Parginos, A. Hariri, and G. Franco, Regulating Artificial Intelligence in the EU, United States and China-Implications for energy systems, in IEEE ISGT Conference, Grenoble, 2023, pp. 1–6. [Online]. Available: https://hal.science/hal-04167091.

[49] MINT, India will regulate AI to ensure user protection, https://www.livemint.com/ai/artificial-intelligence/india-will-regulate-ai-to-ensure-user-protection-11686318485631.html.

[50] The Register, India set to regulate AI, Big Tech, with sweeping Digital Act, https://www.theregister.com/2023/05/26/india_digital_act_draft_june/.

[51] Heymann F, Parginos K, Bessa RJ, Galus M. Operating AI systems in the electricity sector under European's AI Act – Insights on compliance costs, profitability frontiers and extraterritorial effects. Energy Rep Nov. 2023;10:4538–55. https://doi.org/10.1016/j.egyr.2023.11.020.

[52] OECD, Explanatory memorandum on the updated OECD definition of an AI system, Mar. 2024. [Online]. Available: http://www.oecd.org/termsandconditions.

[53] Catapult E. Prospects for reinforcement learning - a guide for energy sector applications. UK: Birmingham; 2023.

[54] EU JRC, AI Watch Defining Artificial Intelligence 2.0, Publications Office of the European Union, pp. 1–125, 2021, doi: 10.2760/019901.

[55] D. Piorkowski, M. Hind, and J. Richards, Quantitative AI Risk Assessments: Opportunities and Challenges, pp. 1–16, Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.06317.

[56] Figueira J, Roy B. Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. Eur J Oper Res Jun. 2002;139(2):317–26. https://doi.org/10.1016/S0377-2217(01)00370-8.

[57] Chatzivasileiadis S, Venzke A, Stiasny J, Misyris G. Machine learning in power systems: is it time to trust it? IEEE Power Energ Mag 2022;20(3):32–41. https://doi.org/10.1109/MPE.2022.3150810.

[58] R. Machlev et al., Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities, Aug. 01, 2022, Elsevier B.V. doi: 10.1016/j.egyai.2022.100169.

[59] L. Koessler and J. Schuett, Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries, Jul. 2023.

[60] Brans JP, De Smet Y. PROMETHEE methods. International Series in Operations Research and Management Science 2016;233:187–219. https://doi.org/10.1007/978-1-4939-3094-4_6.

[61] Behzadian M, Kazemzadeh RB, Albadvi A, Aghdasi M. PROMETHEE: a comprehensive literature review on methodologies and applications. Eur J Oper Res Jan. 2010;200(1):198–215. https://doi.org/10.1016/J.EJOR.2009.01.021.

[62] ENTSO-E, Demand forecasting methodology, 2019.

[63] Lumbreras S, Ramos A. The new challenges to transmission expansion planning. Survey of recent practice and literature review. Electr Pow Syst Res 2016;134:19–29. https://doi.org/10.1016/j.epsr.2015.10.013.

[64] Heymann F, Melo J, Martínez PD, Soares F, Miranda V. On the emerging role of spatial load forecasting in transmission / distribution grid planning. In: 11th Mediterranean Conference on Power Generation. Transmission: Distribution and Energy Conversion (MEDPOWER; 2018. p. 2018.

[65] Melo JD, Zambrano-Asanza S, Padilha-Feltrin A. A local search algorithm to allocate loads predicted by spatial load forecasting studies. Electr Pow Syst Res 2017;146:206–17. https://doi.org/10.1016/j.epsr.2017.01.020.

[66] Heymann F, vom Scheidt F, Soares FJ, Duenas P, Miranda V. Forecasting energy technology diffusion in space and time: model design, parameter choice and calibration. IEEE Trans Sustain Energy Apr. 2021;12(2):802–9. https://doi.org/10.1109/TSTE.2020.3020426.

[67] Doostan M, Chowdhury BH. Power distribution system fault cause analysis by using association rule mining. Electr Pow Syst Res 2017;152:140–7. https://doi.org/10.1016/j.epsr.2017.07.005.
[68] Wang F, et al. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. Energy Convers Manag 2018;171(April):839–54. https://doi.org/10.1109/PerCom.2014.6813940.
[69] Heymann F, Bessa R, Liebensteiner M, Parginos K, Hinojar JCM, Duenas P. Scarcity events analysis in adequacy studies using CN2 rule mining. Energy AI May 2022;8. https://doi.org/10.1016/j.egyai.2022.100154.
[70] Marot A, et al. The impact of the growing use of machine learning/artificial intelligence in the operation and control of power networks from an operational perspective. Paris: France; Nov. 2024.
[71] A. Marot, Personal communication, Feb. 12, 2025, Paris, France.
[72] Mari A, et al. Real-time estimates of Swiss electricity savings using streamed smart meter data. Appl Energy Jan. 2025;377. https://doi.org/10.1016/j.apenergy.2024.124537.
[73] SFOS, Competence Network for Artificial Intelligence (CNAI) - project database, Neufchatel (CH), 2023.
[74] Marot A, et al. Learning to run a power network challenge for training topology controllers. Electr Pow Syst Res Dec. 2020;189:106635. https://doi.org/10.1016/j.epsr.2020.106635.
[75] V. Biagini, M. Subasic, A. Oudalov, and J. Kreusel, The autonomous grid: Automation, intelligence and the future of power systems, Jul. 01, 2020, Elsevier Ltd. doi: 10.1016/j.erss.2020.101460.
[76] B. Kroposki et al., Autonomous Energy Grids, in Hawaii International Conference on System Sciences Waikoloa, Hawaii, 2018, pp. 1–11. [Online]. Available: http://www.osti.gov/scitech.
[77] H. Huang and P. Burgherr, MCDA Calculator: A Streamlined Decision Support System for Multi-Criteria Decision Analysis, in Decision Support Systems XIV. Human-Centric Group Decision, Negotiation and Decision Support Systems for Societal Transitions, vol. 506, Cham: Springer, 2024, pp. 28–42. [Online]. Available: https://mcda-calculator.psi.ch.
[78] A. Marot, A. Rozier, M. Dussartre, L. Crochepierre, B. Donnot Rte, and A. Lab, Towards an AI Assistant for Power Grid Operators, ArXiv, pp. 1–19, 2021.
[79] CEER, Cybersecurity Report on Europe's Electricity and Gas Sectors, Brussels, Belgium, 2018.
[80] Kachirayil F, Huckebrink D, Bertsch V, McKenna R. Trade-offs between system cost and supply security in municipal energy system design: an analysis considering spatio-temporal disparities in the value of lost load. Appl Energy Mar. 2025;381. https://doi.org/10.1016/j.apenergy.2024.124896.
[81] Institute for AI Policy and Strategy, AI integrity: Defending against backdoors and secret loyalties, Jan. 2026.
[82] Heymann F, Küfeoğlu S, Galus M. Digitalization, autonomy and the future of energy policy. Energy Res Soc Sci Sep. 2025;127. https://doi.org/10.1016/j.erss.2025.104167.
[83] THEMA Consulting Group, Data Exchange in Electric Power Systems: European State of Play and Perspectives, Oslo, 2017. [Online]. Available: https://www.entsoe.eu/Documents/News/THEMA_Report_2017-03_web.pdf.
[84] Booshehri M, et al. Introducing the Open Energy Ontology: enhancing data interpretation and interfacing in energy systems analysis. Energy AI Sep. 2021;5. https://doi.org/10.1016/j.egyai.2021.100074.
[85] M. Goumas and V. Lygerou, An extension of the PROMETHEE method for decision making in fuzzy environment: Ranking of alternative energy exploitation projects, Eur. J. Oper. Res., no. 123, pp. 606–613, Dec. 1998, [Online]. Available: www.elsevier.com/locate/dsw.
[86] Huang R, Kadziński M, Figueira JR, Corrente S, Siskos E, Burgherr P. A modular Simos-Roy-Figueira framework for tailored weight elicitation in multi-criteria decision aiding. Expert Syst Appl May 2026;311. https://doi.org/10.1016/j.eswa.2026.131315.
[87] Niet I, van Est R, Veraart F. Governing AI in electricity systems: reflections on the EU artificial intelligence bill. Front Artif Intell 2021;4(July):1–7. https://doi.org/10.3389/frai.2021.690237.
[88] National Institute of Standards and Technology, Challenges to the monitoring of deployed AI systems, Mar. 2026. doi: 10.6028/NIST.AI.800-4.