Safety Science 184 (2025) 106731

Contents lists available at ScienceDirect

Safety Science

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

Organising AI for safety: Identifying structural vulnerabilities to guide the design of AI-enhanced socio-technical systems

Alexandros Gazos [a,*], James Kahn [b,c], Isabel Kusche [d], Christian Büscher [a], Markus Götz [b,c]

[a] Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlsruhe, Germany
[b] Helmholtz AI, Karlsruhe, Germany
[c] Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany
[d] Department of Social Sciences, Economics and Business Administration, University of Bamberg, Bamberg, Germany

ARTICLE INFO

Keywords:
Socio-technical system
Organisation
Safety
Vulnerability
Microgrids
Artificial intelligence
Design

ABSTRACT

Artificial intelligence is increasingly being integrated into socio-technical systems. Existing design principles for ethical, safe and trustworthy AI tend to be highly abstract and focus on AI systems in isolation. They have rarely considered the adverse effects on safety that may emerge from interactions between AI and other technical components. Organisational theories of safety take such emergent outcomes of interactions between entities in socio-technical systems into account. They offer guidance on how to identify structural vulnerabilities in socio-technical systems enhanced by AI, and how to organise the design and operation of such systems for safety. In this paper, which is the result of a collaboration between sociologists and computer scientists (AI consultants), we conduct an analysis that can support the process of designing AI-enhanced autonomous systems in order to avoid structural vulnerabilities. It builds on organisational theories of safety and derives five key descriptors from them, the examination of which can guide the design of AI-enhanced systems. We demonstrate the utility of the descriptors by applying them to proposals for AI-enhanced critical functions in advanced microgrids. We discuss these proposals from the research literature on microgrids and review their effects on structural vulnerabilities. We then explore the implications that go beyond the example of advanced microgrids and propose steps for reviewing and reflecting on structural vulnerabilities that AI controllers may introduce into socio-technical systems.

1. Introduction

Artificial intelligence (AI) components are increasingly considered as enhancements to well-established technologies, ultimately for the creation of autonomous systems. In such systems, computer algorithms make operational decisions without human interference, for previously specified tasks. The non-deterministic nature of some AI systems, their complexity and computational power, which enable their autonomous adaptation to changing situations and new information, pose new challenges with regard to safety (Varshney and Alemzadeh, 2017; Johnson, 2022). This is especially true when critical infrastructures are enhanced by these systems, as a disruption of their services threatens to affect a "wide variety of social capacities" (Schulman and Roe, 2007).

In response to the new challenges posed by AI-enhanced systems making operational decisions autonomously, this paper proposes a new perspective on safe system design principles. It differs from the plethora of guidelines and principles for trustworthy or ethical AI (Hagendorff, 2020; Thiebes et al., 2020) by focusing not on the specificities, like the inherent opacity and lack of interpretability of some models (Burrell, 2016), and possible risks of AI models themselves (Amodei et al., 2016). Instead, the paper's major underlying assumption is that the *recombination* of technologies with AI adds complexity to seemingly well-understood systems to the extent that, apart from the expected benefits, unexpected new vulnerabilities may arise (Zio, 2016). When discussing AI we refer to *weak* AI only, i.e. AI algorithms trained on data to perform a specific, narrow set of tasks, in contrast to strong AI or Artificial General Intelligence, i.e. an AI system with human-like general intelligence across a broad range of tasks (Searle, 1980; Fjelland, 2020).

Although the notion of *weak* artificial intelligence has nothing to do with general artificial intelligence that matches human intelligence, the specific, narrow sets of tasks that weak AI is supposed to perform include capabilities that were once solely the domain of human operators in socio-technical systems. Yet, AI components and humans are dissimilar in many regards. AI components remain deterministic

* Corresponding author.
E-mail address: alexandros.gazos@kit.edu (A. Gazos).

https://doi.org/10.1016/j.ssci.2024.106731
Received 5 October 2023; Received in revised form 1 August 2024; Accepted 24 November 2024
Available online 12 December 2024

since their output is limited by the data the AI was trained with. In contrast, some capabilities of human operators extend beyond purely technical skills and can contribute to the safety of a system based on improvisation and novel actions — something beyond the reach of current AI technology. In this paper, we therefore aim to explore vulnerabilities of AI-enhanced socio-technical systems resulting from a design process that entails the decision to shift away from social actors to more technical components at the level of operation.

Analysing a socio-technical system in terms of its vulnerabilities implies two basic questions: "vulnerability of what?" and "vulnerability to what?" (Khazai et al., 2014). The first question demands a delineation of the system that is the focus of analysis. We are interested in understanding the consequences for vulnerability in systems where AI controllers take over operational decisions, interact with other technical components and social actors in the system. AI controllers (continuously) learn from data, which AI practitioners (Amershi et al., 2014; Vela et al., 2022) must provide. In principle, it would be possible to broaden the perspective and include the wider organisational context in which an AI-enhanced technology is deployed in the analysis. For the purpose of exploring vulnerabilities that AI may introduce when combined with previously used technologies, we will, however, focus on the social aspects of a socio-technical system primarily with regard to the interplay of AI (sub-)systems with the experts involved in their deployment for a particular operational task. This includes not only AI practitioners, but also, although otherwise rarely considered, domain experts (Amershi et al., 2014) and their knowledge about the socio-technical system

This focus leads to an answer to the second question. Gößling-Reisemann et al. (2013) differentiate between event-based vulnerability and structural vulnerability. Event-based vulnerabilities concern exposure and adaptive capacity in response to external perturbations, whereas structural vulnerabilities stem from potential failures independent of external perturbations and relative to the inherent adaptive capacity of a system (Gößling-Reisemann et al., 2013, p. 849). Our analysis aims at identifying structural vulnerabilities, as we look at vulnerabilities that are introduced by AI into systems independent of any threat event. The analysis thus provides an orientation towards closing gaps that a focus on AI systems in isolation (Johnson, 2022) leaves with regard to the safety of comprehensive systems incorporating AI.

Different domains (e.g. energy, traffic, medicine) define safety in specific ways, have their own design principles and regulations (Varshney and Alemzadeh, 2017). A re-combination of established technologies with an AI component does therefore not affect all domains uniformly. The more disruptive AI enhancements are for a domain, the more complex the necessary re-assessment of safety is likely to be as the wider organisational context will certainly be affected. Consequently, we develop our analysis of structural vulnerabilities based on the example of an established technology for which the introduction of AI components is an incremental and not a disruptive change.

This example is the AI-enhanced advanced microgrid. The U.S. Department of Energy's Microgrid Initiative defines microgrids as "a group of interconnected loads and distributed energy resources within clearly defined electrical boundaries that acts as a single controllable entity with respect to the grid. A microgrid can connect and disconnect from the grid to enable it to operate in both grid-connected or island-mode" (Ton and Smith, 2012, p. 84). Advanced microgrids have automatic functions beyond automated islanding and load-shedding. The shift towards more local grid operation is one possible solution to the challenges and vulnerabilities arising from the transition to clean energy in combination with the interdependencies of a large-scale electricity grid (Streck, 2021; Witsch, 2021). Multiple microgrids can isolate from the main system in the case of disturbances such as frequency drops or sudden overloads, (Rodrigues et al., 2020; Venkatanagaraju and Biswal, 2020) and thus increase the overall reliability of a critical infrastructure.

Controlling the interplay of multiple microgrids in a large-scale power grid is extremely demanding, which is why current research often suggests the use of artificial intelligence (AI) as part of the solution (Ali and Choi, 2020). AI controllers promise smooth adaptation to varying conditions on the production and consumption side by analysing vast amounts of real-time data, which are either translated into suggestions for human decision-makers or trigger specific decisions automatically. These data concern variables that are relatively well understood and limited in number (e.g. weather, load, consumption), compared to those relevant for many other technologies, for example autonomous cars. An exploration of the structural vulnerabilities that the integration of AI controllers into advanced microgrids may create can consequently offer an orientation that is relevant (although not necessarily sufficient) for other cases in which AI controllers are introduced into otherwise well-understood socio-technical systems. Since the use of AI in microgrids is still mostly at the research stage, the paper conducts this analysis based on a desktop study of research papers that address the design of microgrids and an iterative discussion process between the computer scientists and the social scientists who co-wrote the paper.

Section 2 introduces organisational theories of safety as the theoretical background of the paper and explains their relevance for understanding implications of AI for the safety and reliability of socio-technical systems. Section 3 draws on these theories to propose five descriptors of socio-technical systems that are affected when they include AI components. They capture key structural vulnerabilities that need to be considered at the design stage to ensure safety and reliability. Section 4 explains the empirical basis and methodological approach for applying the descriptors to the analysis of structural vulnerabilities in AI-enhanced advanced micgrogrids. Section 5 describes the basic structure of advanced microgrids and identifies critical functions for which AI applications are being considered and experimented with in current research. Section 6 then applies the descriptors proposed in Section 3 to the case of three critical functions in advanced microgrids. It explores structural vulnerabilities that result from introducing AI into microgrids, in particular from the altered interaction of technical components and from the role that AI practitioners and domain experts play in implementing such change. The section also indicates how and to what extent a mindful system design could avoid these vulnerabilities. Section 7 discusses the results of this analysis and the aspects that may be generalised beyond the example of microgrids. Finally, Section 8 offers a conclusion and discusses the limitations of our study.

2. Theory: AI and the organisation of safety and reliability

Reflections on design principles for ethical, safe and trustworthy AI have proliferated in recent years and resulted in numerous guidelines (Hagendorff, 2020). They were triggered by the uptake of AI in various fields of application as well as concerns about the implications of a possible future "superintelligent AI" for humankind (Conn, 2015). By now, those reflections have started to find their way into legislation, in particular the AI Act of the European Union, which explicitly draws on the recommendations of a High-Level Expert Group on AI (2019) regarding ethics principles for trustworthy AI (Kusche, 2024). The notion of trustworthiness encompasses technical robustness and safety but also adherence to general ethical principles and respect for fundamental rights. Thiebes et al. (2020) compare different frameworks for trustworthy AI and find that most agree on the principles of beneficence, non-maleficence (which includes safety), human autonomy, justice and explainability. Significantly, these principles are both highly abstract and focus on AI systems and their providers. While AI certainly poses particular challenges with regard to ethics and fundamental rights and safety can be considered as one such right, AI also raises questions in terms of how to actually *organise* for safety. These questions go beyond technical measures of robustness or accuracy of AI (Hendrycks and Dietterich, 2019; Zhang et al., 2019). They concern the so far underdeveloped aspect of effects that emerge from interactions between

AI and technical components and the sociotechnical context of the design and redesign process. They cannot be answered without considering the difference that it makes for the safety of a socio-technical system when at least one of its parts behaves neither like a human nor a conventional technical component. Understanding the problem of integrating AI into a sociotechnical system as a matter of organisational design directs the attention to organisational theories, even though they did not incorporate the possibility of autonomous operation when they were developed.

Perrow's Normal Accident Theory (NAT) is a suitable starting point when exploring the implications of operational decision-making by AI. It heuristically categorises the types of interaction and couplings between system components. The theory posits that complex systems are prone to accidents if their components are tightly coupled, i.e. do not allow for elasticity and flexibility. As Hopkins summarises Perrow's distinction (Hopkins, 1999, p. 95): In tightly coupled systems, one thing rapidly follows another with little opportunity for intervention. These systems are usually highly automated. Power grids are, for Perrow, a prime example of tightly coupled systems (Perrow, 1999), as opposed to loosely coupled systems, where there is plenty of time and room for intervention before problems become serious. Yet, AI-based operations do not necessarily fall into Perrow's category of tight coupling. They are capable of adapting to environmental changes within the boundaries implied by their training data. The advent of artificial intelligence (AI) consequently suggests that the equation between tight coupling and automation may no longer be valid. Loose(r) couplings may increasingly be integrated into future socio-technical systems via smart, learning machines (Büscher, 2022).

High Reliability Theory provides another theoretical anchor point. Its proponents started out by studying the capacity of organisations to reliably cope with the challenges of complex and tightly coupled sociotechncial systems (La Porte and Consolini, 1991) and identified a set of abilities that describe High Reliability Organisations (Weick et al., 2008). Other authors treat the concept of High Reliability Organization (HRO) theory more as an ideal or goal for organisations and sociotechnical systems to strive for in order to become more reliable or safer (Cantu et al., 2020; Lekka, 2011).

With the advent of digitalisation, i.e. increasingly automated processes, the way reliability needs to be organised changes (Schwiderowski and Beck, 2023). Especially for fully autonomous systems, the design process itself becomes one of the central pillars of organising for reliability. By introducing AI into component interactions, the characteristics of these interactions changes as well. The capabilities of AI transcend traditional control, which was based on preprogrammed algorithms. On one hand, AI controllers are deterministic in the sense of being limited by their model features and the data they were trained on. On the other hand, AI practitioners have not determined ex ante what AI controllers are supposed to do with a particular input; instead, the respective algorithms learn and adapt based on the input data they are exposed to. How inputs are transformed into outputs during operation is therefore not fixed ex ante and, depending on the type of AI used, it can even be difficult to understand ex post. Consequently, the interactions within a socio-technical system become even more complex in the presence of AI controllers.

The interaction between AI controllers and technical components not only needs to be properly designed, but maintained and *organised*. For autonomously operating systems, the training process of AI controllers, their designers and the potential users of AI applications (e.g. companies operating a microgrid) constitute an organisation that can strive for high reliability. The principles of HROs help to shift the focus of AI design towards safety concerns. Mindful design principles that take into account interactions of AI with technical components and social actors are a key way to strive for high reliability.

In summary, the look at NAT suggests that the introduction of AI into a socio-technical system may imply a loosening of previously tight couplings, but also an increase in interactional complexity between components. HRO theory is a pointer to aspects that are important to consider already at the design stage when striving for reliability of a socio-technical system that incorporates AI (Schwiderowski and Beck, 2023). At the same time, none of these theories explicitly focuses on the design stage and considers the characteristics of technical components. Consequently, we also draw on Leveson's System-Theoretic Accident Model and Processes (STAMP) (Leveson et al., 2009) for language that can capture what it is that AI components do in a socio-technical system. Firstly, STAMP frames safety as a control problem at the level of the system and considers both human and automated controllers. Since AI in operational decision-making is all about controlling certain functions within a system, the notion of (lack of) control underlies the identification of structural vulnerabilities that AI may introduce. Secondly, STAMP employs the concept of process model. It refers to the model of the system being controlled that any controller, whether human or automated, must have. In the case of humans, this corresponds to mental models, for which HRO theory stresses flexibility and mindfulness as requirements (Weick et al., 2008). In the case of AI controllers, it is the AI method employed that provides a process model of what is supposed to be controlled. An AI process model is less flexible than the mental models of humans; yet it is still somewhat flexible due to the role that learning from data plays for AI.

3. Five descriptors for identifying structural vulnerabilities

The three theories introduced in the previous section emphasise different perspectives on socio-technical systems: NAT (Perrow, 1999) provides an overview of the risks, and in our case vulnerabilities, that are typical for complex systems. HRO theory (Weick et al., 2008) offers inspiration regarding abilities that organisations can strive for to operate highly reliable. STAMP (Leveson, 2012) is helpful for considering temporal elements of vulnerability and adapting organisational concepts to technical agents (e.g. employing the notion of process model). In combination, the three approaches suggest a number of interdependent characteristics that vary between socio-technical systems. These descriptors are affected when technical components operate autonomously based on AI, trained and retrained by humans according to a pre-specified design. Considering them can highlight structural vulnerabilities that the design of AI controllers and emergent effects of operational interactions with other technical components may create.

We propose the following five descriptors, subsequently explicated, to capture the structural vulnerability of a socio-technical system (see Table 1):

1. overall system behaviour (linear vs. complex)
2. mode of control of the system (tight vs. loose coupling)
3. points of control within the system (centralised vs. decentralised)
4. operational/structural system dynamics (low vs. high)
5. adaptive capacity of the system controllers (low vs. high)

The first three descriptors are linked to the basic idea of NAT that socio-technical systems with many complex interactions and tight coupling are more vulnerable than loosely coupled complex systems (Perrow, 1999). The implied causal link was challenged by HRO theory, but this challenge hinged on the importance of well-trained human operators and controllers. Since AI controllers differ from human ones, our analysis needs to treat overall system behaviour and coupling as variables without ex ante assuming a specific causal relation between them.

Complex as opposed to linear *overall system behaviour*, according to Perrow (Perrow, 1999, p. 72–78), means that system components interact with each other in unfamiliar or unexpected sequences, which are not observable or not immediately comprehensible. Such a behaviour is more likely when individual components are in close proximity to each other, while also being responsible for several functions in the system at once (common-mode connection). Furthermore, complex systems imply numerous monitoring and control parameters and even more numerous potential interactions between them and, concomitantly, a

Table 1
Synopsis of the five descriptors for identifying structural vulnerabilities.

Descriptors	Spectrum	Markers
System behaviour	linear ↔ complex	The more complex a system behaves, the closer the components are, the more functions they fulfil, the more interactions they exhibit, the more feedback loops and indirect information there are.
Mode of control	tight ↔ loose coupling	Strict protocol, without much delay and with little contingency in its actions ↔ flexible sequences, alternative courses of actions and more leeway for intervention and improvisation.
Points of control	centralised ↔ decentralised	System-wide awareness and coordination ↔ fine-grained local understanding and swift response.
System dynamics	low ↔ high	Operational dynamicity, i.e. fluctuations in key operational variables, and structural dynamicity, i.e. changes in key structural variables (e.g. component degradation, new technology and shifts in the environment).
Adaptive capacity	low ↔ high	High adaptive capacity is marked by a flexible control hierarchy, extensive training for edge cases and a continuously up-to-date process model of the system state.

multiplicity of new and unintended feedback loops (Perrow, 1999, p. 87). Operators, controllers, and by extension AI practitioners and domain experts, in systems with complex interactions tend to receive their information about the system in an indirect and derived form, resulting in an incomplete understanding of certain processes (Perrow, 1999, p. 84–88). A complex system consequently requires structural arrangements such as a more decentralised control structure "to cope with unplanned interactions of failures" (Hopkins, 1999, p. 98) in its components and the indirect information about their behaviour.

The distinction between tight and loose coupling refers to the timing and sequencing of operations in a socio-technical system (Perrow, 1999, p. 89–96), that is the *mode of control*. In tightly coupled systems, processes are time-sensitive and delays are not possible. The sequencing of operations is fixed, and buffers and redundancies are part of the design. In loosely coupled systems, time delays are possible, the order of process sequences can be changed and there is enough slack in the system to create buffers and redundancies not planned ahead and leeway for human intervention and improvisation.

As NAT suggested that complex systems require decentralised operations, it also argues that, in contrast, tightly coupled systems require rigid centralised control (Perrow, 1999, p. 331–333). The theory of HROs clarifies how this does not necessarily result in contradictory requirements for complex and tightly coupled systems, but that organisations can strike a balance with regard to the *points of control*: they can delegate authority to recognise and detect hazards at a local level for a fine-grained understanding and fast response times, while still being centrally aware of vulnerabilities, coordinating system-wide responses and initiating local learning processes (Weick et al., 2008, p. 60).

In addition to the three basic descriptors overall system behaviour, mode of control and points of control, the temporality of a socio-technical system as a whole is relevant to its structural vulnerabilities. Systems change over time in ways beyond their control and cannot be assumed to be static throughout their lifetime (Leveson, 2012, p. 175–176). Their *structural dynamicity* includes the degradation of system components, the introduction of new technology, changes in the goals and structure of the organisation operating the system and changes to the regulatory environment. In the case of AI, model degradation (Vela et al., 2022), changes in terms of functional expectations, a change in the valuation of key variables (or different key variables altogether) and unexpected events all contribute to structural dynamicity. They require a retraining of the AI models that should be based on deliberations of AI practitioners with domain experts, Amershi et al. (2014) in order to include their experiences and knowledge about the socio-technical system. At the same time, advanced microgrids and AI are both intended to be technologies that can reliably and quickly cope with *operational dynamicity*, meaning fluctuations in key variables during operation (Section 1).

Finally, the *adaptive capacity* of a system affects its structural vulnerability (Gößling-Reisemann et al., 2013). The research on HROs demonstrated multiple ways of how operators and system controllers could foster this capacity. In essence, the cases identified by this research (Weick et al., 2008; La Porte and Consolini, 1991) as HROs pointed to a flexible authority or decision-making structure as a key to adaptive capacity (Sawyerr and Harrison, 2019). Training to "increase people's response repertoire enlarge[s] the range of issues that they notice and can deal with" Weick and Sutcliffe (2015, p. 110). Adaptive capacity also requires mindfulness and an adaptive cognitive model, i.e. the ability to process new information about the system generated in the course of adjusting to unexpected events and to retrieve this information when needed (Weick et al., 2008). Furthermore, an emergent quality of a system's modus operandi was assumed in cases where operators continuously "combine fragments of old routines with novel actions into a unique response to deal with a unique input" (Weick et al., 2008, p. 55). AI controllers differ from human operators and cannot be expected to contribute to adaptive capacity in identical ways. In particular, they are unable to come up with truly novel actions because they rely entirely on the *history* of a system's process states. Nevertheless, it is possible to draw some analogies with regard to other underlying principles of fostering adaptive capacity.

AI controllers cannot spontaneously delegate authority laterally or to new controllers outside predetermined pathways. Rather, authority is given to sub- or superordinate controllers according to the system's predetermined hierarchy. However, if AI controllers are able to shift control between higher and lower levels, they can form a flexible control hierarchy within the boundaries of their predetermined control logic. If AI controllers face an event that leaves these boundaries, they would have to delegate control to human operators, i.e. knowledgeable personnel, or the designers, so they can retrain the model based on the unexpected event.

With regard to training the response repertoire, an analogy in the training of an AI controller, which is otherwise obviously very different from the training of human operators, would be the training on examples at the extreme ends of their operational boundaries. Since an AI's boundaries of possible actions result from the limits of the data it was trained on, these so-called edge cases in the data are a prerequisite for it being prepared for rare events.

Similar to human operators, AI controllers need to be sensitive to data and able to indicate changes in the system state to allow operators and designers to retake control, if necessary. Whenever the actual process state in a given socio-technical system diverges, e.g. due to emergent phenomena or hazards, from the process model that controllers, whether human or AI, rely on accidents become more likely (Leveson et al., 2009, p. 243). Although AI controllers are not able to process new information and adjust their control actions on the fly, akin to an adaptive cognitive model or mindfulness, they can at least somewhat

compensate for this. They can improve on their capacity to adapt to the full range of system state changes that have already occurred, as they have been trained on data that is supposed to represent the entire history of possible system changes. Depending on the data they were trained on, AI controllers have potentially access to a wider range of possible system states and their process model can encompass more variables, compared to human operators. Consequently, mindfulness must primarily be located in the design process and requires AI practitioners and domain experts to be aware of the limitations of the data used for training AI models. Another limitation of AI models that they have to be aware of is the possibility of catastrophic forgetting, which has no parallel in human operators or conventional technical components. It can occur if AI models are retrained on a new dataset, which under certain conditions risks erasing what they were previously trained (French, 1999).

In sum, the five descriptors should be able to capture key structural vulnerabilities that need to be considered when organising for the safe integration of AI components into a socio-technical system.

4. Method

Due to the state of research on AI-enhanced microgrids, we conducted our analysis as an exploratory desk study of peer-reviewed original research and review articles on applications and trends of AI for deployment in (advanced) microgrids. The starting point of the study was a formal definition for advanced microgrids, published in a White Paper by the U.S. Department of Energy (Bower et al., 2014) and summarised in the next section, as well as the most cited comprehensive review of the state of the art for artificial intelligence techniques in smart grids (Ali and Choi, 2020) at the time the project started in January 2021. Based on the latter, we determined the scope of our analysis and restricted it to AI use for components of microgrids that are critical to system services in the sense of controlling and maintaining their functionality. We thus identified three critical functions (see Section 5). We searched for peer-reviewed articles and reviews that deal with the integration of AI-based methods for these critical functions into microgrids and/or compared the impact to traditional methods. Using Google Scholar, we combined keywords denoting the subject of microgrids (“microgrids” OR “smart grids”) and AI (“artificial intelligence” OR “machine learning”) with keywords related to the three critical functions (“load forecasting” OR “fault detection” OR “energy management system”). After this initial search, we also added literature referenced in the identified reviews and articles. We included articles published between 2015 (earliest relevant publication) and July 2022 (cut-off date).

We analysed the respective articles employing the descriptors of structural vulnerability developed from the theory. The results of this analysis are presented in Section 6. The collation, comparison, and analysis of the literature was integrated with a continuous dialogue between the three sociologists and the two computer scientists and AI consultants on the research team. Although presented in the previous section, this dialogue also informed our specification of the descriptors. Consequently, the development of descriptors, based on theories that focus on organising for safety, and their application to the example of advanced microgrids overlapped. Theory and empirical example thus fed back into each other to make full use of the interdisciplinary composition of our research team. Such a recursive approach, and the dialectic between empirical case and theoretical perspective it involves, is common in qualitative research in the social sciences (Timmermans and Tavory, 2012). For our paper, this approach was essential for gradually developing a shared vocabulary — manifested through the descriptors — that would resonate with both computer engineers and social scientists, thereby bridging the gap between the purely technical literature on AI and advanced microgrids and a social science perspective.

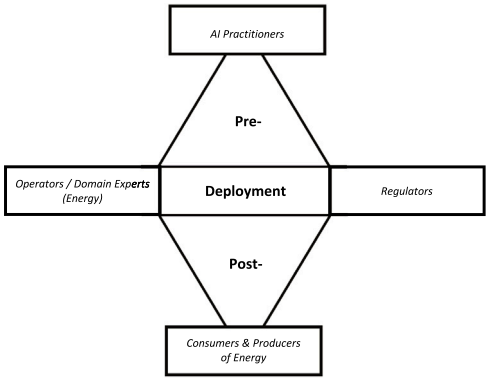

Fig. 1. A conceptual model of the social actors involved in the (pre-/post-) deployment of a microgrid.

Research on AI-enhanced microgrids is largely at the development stage, which includes simulations and experiments but only exemplary applications in real-world environments. We propose that a socio-technical perspective at this stage is particularly useful to identify potential structural vulnerabilities that stem from the combination of a well-understood technology, such as microgrids, and AI components. Their identification can inform the future real-world implementation of AI-enhanced microgrids. Moreover, the analysis, developed by drawing on the example of microgrids, aims at a level of generalisation that makes it possible to apply it to other cases where an established technology is supposed to be enhanced by AI.

5. Social context and design principles of advanced microgrids and the role of AI

Before we can explore the structural vulnerabilities of an AI-enhanced microgrid where operational decisions are transferred from social actors to AI controllers, a conceptual model of the social actors involved helps to identify the nature of the shift away from human operators. The autonomous operation of an AI enhanced microgrid is preceded by a predeployment phase and succeeded by a postdeployment phase (see Fig. 1).

Predeployment involves the three organisationally embedded groups of AI practitioners (computer engineers and programmers), operators/domain experts (e.g. controllers and electrical engineers) and regulators (e.g. government, standard-setting bodies). AI practitioners select the data and train the AI model according to the functional expectations of the operators/domain experts. The former operators are no longer involved in the actual process of operating the deployed system. As their domain expertise guides the data selection and training process, they rather take part in the design process for a more autonomous, AI-enhanced microgrid. AI practitioners and domain experts are the social actors on whose interplay with the AI and other technical components our subsequent analysis will focus. Their actions are however also influenced by the third group that forms part of the microgrid as a socio-technical system: Regulators set limits on the possible specifications the design can take. Their regulations can take the form of legislation, standards and certificates, crucially affecting the safety requirements of the organisations behind the AI practitioners and operators/domain experts (Leveson et al., 2009). Moreover, organisational embeddedness means that the AI practitioners and operators/domain experts are also affected by company, project, manufacturing, and operations management decisions made in other parts of their respective organisations (ibid.).

Postdeployment involves regulators and operators/domain experts as well, but also includes consumers and producers. The latter consist of residential communities, commercial or industrial users that can

consume and produce electricity as well as share energy with other users (Ali and Choi, 2020). For our design focus they come into play in the form of energy demand or load fluctuations and when they add or remove decentralised energy resources or energy storage systems. The load is not only used to provide the information to the deployed AI controller, on which it can enact control decisions, but the history of the previously installed grid and its load fluctuations are used to train the AI model during the predeployment phase. The relationship between pre- and postdeployment is therefore cyclical. Structural dynamics (see 3) like shifting operational goals, the addition or removal of energy resources or storage systems as well as new regulations would also necessitate a move from postdeployment back to predeployment and redesign.

Component-wise, advanced microgrids contain all the key elements of a regular utility grid. They are connected to the main utility grid at one point of common coupling, through which they exchange power and information with the main grid in real time. Bower et al. (2014) outline the defining characteristics of advanced microgrids. They are geographically delimited, connected to the main/host utility grid at one point of common coupling, and fed from a single substation.

They automatically transition to/from islanded mode, and include distributed energy resources, an EMS and power and information exchanges on both sides and across the point of common coupling in real time.

The physical components of a microgrid belong to five categories (Yoldas et al., 2017): the transmission system, the distribution system connecting all microgrid components, the distributed generation of energy, the energy storage systems and the load. The transmission system is only active during grid-connected mode, and its feeder lines maintain the bidirectional power flow through the point of common coupling. If the frequency or load leaves a predetermined range, then the microgrid switches to islanded mode. The remaining four categories of physical components continue to be active in islanded mode.

As an advanced microgrid cannot be continuously controlled via physical on-site operations, the system requires accurate state estimation at each point of control. State estimation needs accurate, high-resolution information from the physical components to provide controllers with an up-to-date *process model* of the actual system state. Process modelling at scale is enabled by the advanced metering infrastructure. Data from the microgrid components is also collected by smart sensors and measurement devices, such as intelligent electronic devices and remote terminal units.

As soon as automated controllers receive the information necessary to determine the operational state of the advanced microgrid, they perform appropriate control actions. The levels within an advanced microgrid at which control is exercised can be subdivided according to their timescale requirements (Ray and Biswal, 2020): Primary or local-level controllers are used for fast response (less than 1s) control functions, involving no communications and using only local measurements (e.g. used for frequency/voltage control, power-sharing, etc.). The secondary or upper-level microgrid central controller considers power, loads, and storage measurements (less than 1 h time scales) to manage the microgrid as a whole (including island detection, power dispatching, etc.). At the tertiary level, the distribution management system controls the microgrid central controller as part of the host or main grid. It ensures that the microgrid meets the overall grid demands and operates on longer (greater than 1 h) time scales (including forecast generation, demand response, etc.).

Whether the control itself is performed centrally, in a distributed manner or via a hybrid[1] approach depends on the microgrids' configuration, size, and components used. Centralised control works by delegating decision-making power to the microgrid central controller, which makes optimal decisions based on aggregated data from local controllers. With decentralised control, on the other hand, local level controllers manage power without the need for control commands from a microgrid central controller. The hybrid approach combines both centralised and decentralised controllers to optimise grid operation according to local and system-level demands.[2]

When discussing the consequences of enhancing microgrids with AI, we distinguish between the two broad categories of components that serve critical functions and those that serve non-critical ones. A non-critical function is, for instance, the optimisation of production costs; critical functions maintain the essential "services that the network and systems provide" (Young and Leveson, 2014, p. 32). Our analysis focuses on components that are critical to system services in the sense of controlling and maintaining their functionality. In the comprehensive review of AI applications by Ali and Choi (2020) we identified three critical functions, based on the aforementioned definition:

Forecasting of loads, generation capabilities, grid status, etc. at various time scales.

Fault detection for identification of locations and various causes of line faults, e.g. lightning strikes.

Energy management for economical use of grid components and coordination and optimisation of e.g. distributed energy resources and energy storage systems. This includes the integration of new systems.

The descriptors of structural vulnerability developed in Section 3 can be applied to these critical functions from two complementary perspectives in order to analyse the consequences of introducing AI. On the one hand, the AI-based solutions proposed for each of the critical functions directly concern one descriptor each, corresponding to the function itself. As the following sections will clarify, forecasting is directly linked to the dynamics of the overall microgrid system (Section 6.1), fault detection directly concerns the mode of control in the system (Section 6.2), and energy management is primarily about its points of control (Section 6.3). On the other hand, the descriptors of structural vulnerability are interdependent, which means that the introduction of an AI-based solution will have specific implications for all of them.

6. Analysis: Descriptor interaction in critical functions

6.1. Forecasting and system dynamics

The overall goal of forecasting functions is to ensure the balance between the amount of power being generated and the amount being used (the load). This helps to keep the system running smoothly, reduces energy wastage and maximises the use of renewable energy sources. Forecasting is thus directly concerned with the *operational dynamics* of a microgrid system, which include weather changes as well as fluctuations in people's energy usage. Forecasting uses time-series data to predict microgrid states and capacities at various time scales (Khan et al., 2016): these range from short-term daily operations (up to one week) to medium-term operational planning (weekly up to yearly) and long-term expansion planning (longer than a year). Automated control, and the potential use of AI for it, is most relevant for short-term forecasting.

For forecasting in microgrids and their advanced variant in particular, STLF is the most critical function since many operating decisions are based on it (Chitsaz et al., 2015, p. 50). In contrast to a regular

[1] Hybrid control is also sometimes referred to as hierarchical; we use the term hybrid to avoid confusion with the flexible control hierarchy we discuss in this work when referring to a system's adaptive capacity.

[2] A detailed review of hybrid microgrid control, with a focus on the explicit implementation of building microgrids, can be found in Yamashita et al. (2020).

power system, a microgrid is smaller, which means that it does not have as much built-in stability. In other words, it has *low inertia* due to more load fluctuations (Chitsaz et al., 2015, p. 50). The decentralised energy sources as well as power usage fluctuate, which is harder to compensate within a small-scale system as it lacks buffers present in the main grid (Dag and Mirafzal, 2016). Consequently, operations are *tightly coupled* and thus structurally vulnerable to high operational dynamicity. That makes accurately forecasting the load, i.e. power usage or demand, essential to reliable operation.

AI models in forecasting support estimates of how the process state of a microgrid will develop. Artificial neural networks are most commonly used (Gerwig, 2015). Different from statistical methods, AI-based methods are designed to handle the non-linear data resulting from the many components of an advanced microgrid, like the load, storage units and energy supply. They also have to deal with data from the many devices used to monitor and control these components. By including non-linear variables, artificial neural networks provide further information and take additional, complex relationships in the data (Zor et al., 2017) into account. They are therefore expected to significantly improve the accuracy of short-term load forecasting.

A more accurate Short-Term Load Forecasting (STLF) gives controllers that depend on its forecasts time to react ahead of any upcoming fluctuations or disruptions (e.g. due to a rapid and unexpected change in energy demand). In terms of the descriptors, this amounts to *loosening the mode of control* by mitigating the adverse effects of the low inertia in advanced microgrids. Human operators or other automated controllers can use the predictions of an AI-enhanced STLF to change the order of sequences and consider alternatives for the most appropriate response at their disposal.

As a common-mode function STLF contributes to the *complex behaviour* of the whole socio-technical system since several other functions depend on it. Should the STLF be disrupted or fail, the ensuing dysfunctional interactions would adversely affect the other critical functions, for example in the Energy Management System (see Section 6.3), and could thus trigger cascading failures in the entire system. Human operators and automated controllers depend on STLF to construct and constantly update their process model, which serves as a cognitive or computational model of how the system is currently operating. It allows them to quickly respond if the system goes beyond safe limits for its operation, for example the 50 Hz frequency of currents. AI-enhancement of STLF can help to reduce complexity for and improve response times of other, dependent controllers, in particular under conditions of high operational dynamicity. However, this will only work if the designers trained the artificial neural network on data that includes all the relevant variables. AI designers can hardly make such an assessment on their own but have to rely on domain experts and their experiential knowledge, Domain experts can in effect assist in reducing overall system complexity for human operators and automated controllers by providing AI Designers with criteria to assess the relevant variables and data for the training process of an AI-enhanced STLF.

If domain experts guide the training process, the AI enhancement of STLF can provide a more accurate process model to dependent operators and automated controllers, which would mean an increase in adaptive capacity. There is, however, a trade-off. All neural networks import a structural vulnerability into any system in which they are used, namely the black box problem and the associated lack of interpretability (Burrell, 2016). Operators, designers, and domain experts are unable to understand why a neural network gives a particular output. Furthermore, at least at this stage of research, neural networks cannot provide reliable prediction uncertainties, which means that they cannot (reliably) convey the degree of confidence in their outputs (Chua et al., 2023). In load forecasting, for example, unreliable predictions and a false confidence in them, like a prediction that underestimates the future load requirements, can cause the entire power system network to fail (Khwaja et al., 2017).

Consequently, research is focused on probabilistic load forecasting approaches that lead to more reliable prediction uncertainties (Wang et al., 2019a; Yang et al., 2019; Afrasiabi et al., 2020; Brusaferri et al., 2022). Reliable estimates about how uncertain a prediction is permits human operators and automated controllers to trade some degrees of optimisation against reliability of operation in a rational way. They can thereby err on the side of caution. Moreover, such estimates are important not just for day-to-day operations, but also for knowing when an AI model might start making less reliable decisions due to model degeneration or concept drift (Vela et al., 2022) and deviate from what is considered safe or reliable. If prediction uncertainty increases, designers and domain experts could initiate a retraining of the AI mode preemptively.

6.2. Fault detection and mode of control

Fault Detection and Classification (FDC) is another function in advanced microgrids that is potentially suitable for enhancement by AI (Ali and Choi, 2020).

In general, fault detection involves three steps (Wei et al., 2018):

1. registering that a fault has occurred (detection)
2. classifying the type and location of the fault (classification)
3. taking an appropriate control action to isolate the fault in an effort to protect the rest of the grid (protection)

This means that fault detection directly concerns the *mode of control*. In a tightly coupled system, faults are potentially dangerous since they can easily propagate within the system. FDC intermittently loosens the mode of control when it isolates faulty components while ensuring the functionality of the overall grid.

Fault detection in microgrids is particularly challenging, compared to conventional power systems. Fahim et al. (2020), Bansal and Sodhi (2018), Beheshtaein et al. (2019). It has to deal with both internal faults, which have to be located and corrected, and internal faults, which require a decision on whether to island the microgrid. The distributed generation of energy means that the system is prone to problems like bidirectional currents (Bansal and Sodhi, 2018; Beheshtaein et al., 2019). Conventional, model-based FDC methods monitor the microgrid state and ensure that it is within acceptable limits, predefined by the model (Bansal and Sodhi, 2018).

AI-based FDC methods promise a more fine-grained fault detection. They cannot only detect faults but also classify them into types, based on patterns they have learned to identify in training data from the system. Such a classification would allow human operators and automated controllers (see Section 6.3) to refine their process model of the physical system state, in particular with regard to the types of faults that may occur in it. In this way, it is possible to detect anomalies that model-based methods would ignore (Bansal and Sodhi, 2018). AI practitioners and domain experts can thus improve their understanding of possible system states, with potential benefits for adaptive capacity.

However, the topology of a microgrid changes whenever a distributed energy resource is added or removed, and this *structural dynamicity* creates practical problems for the use of AI in FDC (Bansal and Sodhi, 2018). AI models require large amounts of data, both for training and the validation of predictions. These data have to be collected from grid measurements or simulated, for example, by creating a digital twin of the microgrid. AI practitioners and domain experts face a considerable challenge in this regard. They have to collect and choose the appropriate amount of data, knowing that the AI model can only deal with faults it has encountered in the training data. Training and validation based on simulations runs the risk of deviations between simulated process states and actual ones. Moreover, each change to the topology of a microgrid requires a retraining of the AI model. A high frequency of re-trainings adds *complexity* to the whole socio-technical system, since AI practitioners and domain experts need to be mindful

each time of possible errors during the (re-)training process and their potential consequences.

The high *operational dynamicity* that distributed energy generation creates adds to the practical problems with AI. Processing times increase with the complexity of the trained models, Bansal and Sodhi (2018) ultimately resulting in longer response times. Yet, short response times are crucial because of the fluctuations of distributed energy generation and the low inertia of microgrids. A better understanding of the system in terms of faults and anomalies would therefore come at the cost of slowing down a tightly coupled process. In other words, the attempt to improve adaptive capacity by creating a more fine-grained process model could end up increasing the structural vulnerability of the whole system. A central AI controller, which could improve protection against faults based on its sensitivity to different types of faults and anomalies in the overall system, would consequently need to be counterbalanced by simpler, decentralised FDC devices with faster response times. This would add more devices and interactions between them, increasing the overall complexity of the system. The trade-off with regard to the points of control currently limits the practical use of AI for FDC in microgrids and requires design choices by AI practitioners and domain experts. Neglecting this balance or shifting it too much to one side of the spectrum exposes advanced microgrids to structural vulnerabilities.

To design less vulnerable advanced microgrids in the future and actualise the potential increase in adaptive capacity resulting from AI-based methods, research suggests a prospective development path. AI-based methods could feasibly complement rule-based models in a hybrid control scheme, if technologies like 5G improve on the process of real-time data collection by intelligent electronic devices and the subsequent communication to a central controller (Gutierrez-Rojas et al., 2021). There, the input could be processed by a neural network and the protection setting updated with the help of another AI method (e.g. support vector machine) (Lin et al., 2019).

6.3. Energy management and points of control

The Energy Management System fulfils another critical function for the microgrid. It manages its power flow by coordinating distributed energy resources and the loads based on operational goals (Zia et al., 2018). These operational goals include the minimisation of outages and stable control of renewable energy sources, but also economic and ecological aspects. The number of and balance between operational goals can change over time since they typically entail trade-offs. AI practitioners and domain experts need to be mindful of such changes and trade-offs. They need to make shifts in the relative weight of specific goals explicit, since the goals have to be implemented in the process of (re-)training AI models.

The EMS in advanced microgrids typically consists of decision-making modules, which encompass data monitoring and analytics, forecasting, optimisation and real-time control geared towards the operational goals. Because of its coordination function, the design of an EMS directly concerns the *points of control* in the system. It can prioritise different loads by categorising them, for example into critical loads, which are energy demands that have to be met at all times, and controllable loads, which are more flexible in terms of demands (Wang et al., 2019b; Bagherian and Tafreshi, 2009). Accordingly, several loads could be flexibly curtailed during disruptions that require time-critical decisions (Zia et al., 2018). While designing AI modules as Energy Management System (EMS)-controllers, AI practitioners and domain experts have to consider two points in particular: Firstly, the criteria for these categories may change over time (Alahmed and Al-Muhaini, 2020). Secondly, such categorisations are not just technical but social. They imply a prioritising of certain energy users over others during disruptions. An AI-enhanced EMS will derive its categories for prioritisation from the historical data of the system on which the model is trained. It can therefore reproduce historical biases present in the data, and, for example, continue to categorise certain energy demands as critical, although actual requirements have changed as part of the structural dynamics of the overall system.

The literature on EMS control distinguishes between centralised and decentralised control architectures, with the latter varying in the extent of decentralisation (Meng et al., 2016; Zia et al., 2018). Central control is mostly suitable for small-scale installations, microgrids where security or privacy is a priority, and microgrids whose setup is not expected to change significantly (Meng et al., 2016), i.e. setups with low structural dynamicity. The limits to the use of AI in centralised EMS control are similar to those observed for fault detection, such as longer computing times and problems to adapting the system when decentralised energy resources are added or removed. Current research suggests additional limitations: uncertainties that result from the fluctuations of distributed energy resources (Espín-Sarzosa et al., 2020) and the lack of stability, since the failure of a centralised EMS would result in the breakdown of the whole system (Zia et al., 2018; Meng et al., 2016).

Zia et al. (2018) state in their extensive literature review that the current research focus shifts from centralised control schemes towards decentralised control in more recent research, due to the aforementioned constraints. Decentralised control can take the form of a multi-agent system in which multiple control agents interact with each other (Jimeno et al., 2011) and the environment to achieve both local and global objectives. Research has in particular proposed AI approaches based on multi-agent systems (Tazi et al., 2020; Bourakadi et al., 2020; Harrold et al., 2022).

As information is processed locally, decentralisation limits the complexity and the operational system dynamics that any single controller faces. Compared to a centralised EMS, the mode of control is looser by avoiding a single point of failure and the adaptive capacity is higher as the local controllers maintain operational flexibility during disruptions of their central EMS. Decentralised control can also better deal with structural system dynamics, as it is possible to add components without major revisions to the overall design. However, the decentralised points of control and their local process models require reliable synchronisation. Consequently, the communication network becomes a crucial factor for the system's safety and stability (Zia et al., 2018). Proposals for hybrid architectures, which introduce a supervisor agent that observes the whole microgrid but does not interfere with local agents (Tazi et al., 2020), indicate that, similar to fault detection, the appropriate balance between centralised and decentralised control is a key challenge for AI design and implementation. AI practitioners and domain experts cannot design such a balance and assume it to be continuously appropriate for the system operation as a whole. As the EMS fulfils the critical function of supervisory control, it is necessary to monitor any changes that may require a different balance in the points of control.

What is not discussed in the literature are consequences of the interaction between an AI-enhanced EMS and other AI-enhanced critical functions, such as forecasting and fault detection, in the microgrid. Although the use of AI for each function is supposed to increase the adaptive capacity of the whole system, this result is far from guaranteed when several AI-based components are combined. The layering of AI-based components in an advanced microgrid increases complexity and tightens the mode of control. An EMS relies on proper output from forecasting and fault detection, while these subordinate functions depend on a reliable EMS. If some or all of these modules employ more or less opaque AI models, the result could be cascading failures throughout the system once one of the modules fails to work reliably. Model degradation and concept drift decrease the reliability of AI models over time (Vela et al., 2022). AI practitioners and domain experts would need to keep attention to the specific re-training needs of each AI-enhanced module and how their interaction plays out in operation, which is a demanding task. Moreover, even if all modules work reliably throughout their lifetime, hazardous system states could build up from unforeseen feedback loops, as Leveson et al. (2009) have shown for complex systems in general.

7. Discussion

The promise of AI for advanced microgrids lies in managing the complex interactions between energy resources, storage and energy use, as well as the high dynamicity of the overall system in operation. Compared to the capabilities of human operators, AI models are tightly coupled due to their underlying deterministic set-up. However, the mode of control is a question that pertains to the whole system and not only some of its components, and it is a matter of degree (Perrow, 1999). As the example of advanced microgrids has shown, the integration of AI does not amount to a tightly coupled mode of control for the whole system. On one hand, the critical functions for which its use is foreseen, namely short-term load forecasting, fault detection and energy management, tend to loosen the mode of control in the microgrid. They provide time to adequately respond in advance to change the order of sequences and make use of alternative methods (STLF), decouple faulty components from the microgrid and decouple the microgrid from a failing main grid (FDC) and they enable the prioritisation of loads (EMS). The overall mode of control in the socio-technical system is therefore a result of various design choices, made by AI practitioners and domain experts, that concern the interaction of AI-enhanced components with other parts of the system.

The analysis of this example has, however, identified three main challenges that so far hinder the use of AI in microgrids but also have implications beyond it. The *first challenge* concerns the structural dynamicity of the overall system. When energy resources, for example residential solar installations, are to be added to or removed from an advanced microgrid, the components for forecasting or fault detection must be able to accommodate these changes. However, AI, especially in the case of the commonly used artificial neural networks, are not intrinsically adaptive. This problem with structural dynamicity reaches beyond the case of microgrids, since it concerns a feature of AI methods. Artificial neural networks are generally very sensitive to small changes in input data distributions, an effect known as concept drift (Widmer and Kubat, 1996; Tsymbal, 2004). They also have issues with producing reliable uncertainties and cannot reliably create information about whether uncertainties are due to model imperfections or noisy data, or due to never having encountered the input data before (Chua et al., 2023). Artificial neural networks consequently often do not register when a system state is beyond their capacities to deal with it. In combination with the sensitivity to changes in input data distribution, this makes for a scenario where it is not possible to know from the AI's output alone when it must be retrained for new, shifted input data. This lack of reflexivity regarding the distinction between operational and structural dynamicity is something in which AI controllers fundamentally differ from human operators. This has implications for the notion of adaptive capacity, adopted from research on HROs (Section 3). For AI enhancement of socio-technical systems, there are two different forms of process modelling and adaptation:

Adaptive operation is characterised by responses to events within the scope of system design.

Adaptive restructuring is characterised by responses to events that transcend the boundaries of adaptive operation. Consequently, AI controllers must be retrained and/or the overall system must be redesigned (e.g. conventional methods or human control)

Although AI facilitates adaptive operation thanks to more detailed process models of variable operations, adaptive restructuring poses a considerable challenge. The latter continues to be the purview of AI practitioners and domain experts, who have to retrain the AI models. Adaptive restructuring in response to changes in the topology of the overall system is still faster, compared to models relying on explicit configuration by engineers. It can be particularly fast if a simulation or digital twin of the system exists. However, this acceleration requires mindfulness (Weick and Sutcliffe, 2006) on the part of AI practitioners and domain experts, as constant retraining runs the risk of errors during every iteration. Moreover, they need to keep in mind that simulations can deviate from actual system states.

Provided AI practitioners and domain experts have an extensive data pool available to them and mindfully select relevant variables, the scope of adaptive operations of AI enhanced systems could be widened, thereby potentially reducing the necessary frequency of retraining. For instance, by first training an AI-based forecasting model on mindfully selected data from regions with similar conditions and potentially higher weather extremes, the model can be primed to handle a high operational system dynamicity before being fine-tuned to the given microgrid. Such a procedure, known as transfer learning, has long been established within AI research and Zhuang et al. (2020). Another approach to facilitate adaptive restructuring would presuppose a solution to the problems with reliable uncertainties in ANNs. Then AI practitioners and domain experts could retrain an artificial neural network right before the input data drifts out of acceptable operational limits.

The *second* challenge for AI enhancement concerns the points of control in the system. In contrast to human operators, any flexibility of a system's decision hierarchy has to be designed in advance in the case of AI controllers. Our analysis suggests that choices regarding points of control affect adaptive operations at different hierarchical control levels differently.

Upper-level adaptive operation involves system-wide process modelling and coordination of control responses. It is improved by centralised points of control.

Lower-level adaptive operation involves fine-grained local process models and rapid control responses. It is improved by decentralised points of control.

A trade-off between upper-level and lower-level adaptive operations is consequently inevitable. Proposals for hybrid control schemes in fault detection and energy management, where central and decentralised AI controllers are combined, indicate that it is possible to strike a balance and implement a flexible control hierarchy, including lower level autonomous control. Yet, determining the right balance for a socio-technical system will continue to pose problems for AI practitioners and domain experts.

The *third* challenge is the ensuing complexity when several AI components are to be integrated into the socio-technical system. Our analysis encountered it as an open question with regard to the energy management system in advanced microgrids. Proposals for AI enhancement address it in isolation. They do not consider the possibility that functions subordinate to energy management, such as forecasting and fault detection, could be enhanced by AI concurrently. In the layering of AI modules lies a source of fatal structural vulnerabilities if no corresponding countermeasures are implemented. These countermeasures would amount to a mindful design (Salovaara et al., 2019) that takes into account the descriptors of structural vulnerability in the following ways:

- Map and continuously monitor any interdependencies between the AI modules and other system components to prepare for complex interactions and unforeseen feedback loops.
- Establish points of control that strike an appropriate balance between a coordinating central controller with system-wide awareness and decentralised control devices for swift responses based on a granular process model (corresponding with the second challenge). This would also enable access to more direct information, counteracting the adverse effects of complexity.
- Design for redundancies and substitutions. Superordinate controllers in particular would require such measures to maintain their critical function. This could mean that a conventionally programmed controller, not based on AI, takes over control or the function is temporarily delegated to all decentralised controllers (e.g. in a multi-agent system).

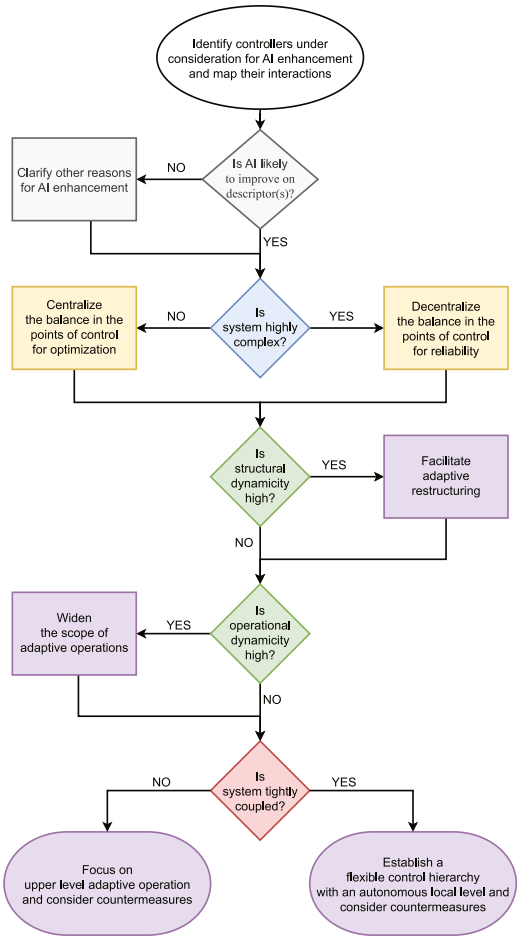

Fig. 2. Template for mindfully designing AI enhancement in order to avoid structural vulnerabilities. Colours indicate the different descriptors.

- Implement a safe fail mechanism to allow for operations to stop in a reasonable manner, initiating an examination and intervention by human operators (Varshney and Alemzadeh, 2017). The results of this examination should inform AI practitioners and domain experts preparing a retraining, ensuring that they are mindful of the component interaction that led to the safe fail.
- Reflect on initial operational goals and on how they have changed until the moment of retraining.

Although these countermeasures are most crucial when several AI modules interact within a socio-technical system, they also offer guidance for reviewing and reflecting design choices about implementing AI in systems more generally. Such a review is particularly helpful given the widespread enthusiasm for AI solutions. It may crowd out reflection on how desirable such solutions actually are for the system in question. Based on our analysis, we propose the following steps to encourage AI practitioners and domain experts to systematically reflect on the consequences of redesigning a socio-technical system for more autonomous operations by integrating AI controllers:

1. Identify the controllers that are under consideration for AI enhancement, map their known interactions with other system elements.
2. Check whether AI enhancement actually entails a plausible promise to improve the system with regard to at least one descriptor. If not, clarify for what other reasons, not related to reliability and safety, AI enhancement is considered (e.g. expected cost-savings).
3. Apply the descriptors to characterise the degree of complexity, points of control, system dynamics, mode of control and adaptive capacity (see Table 1).
4. Reflect on which descriptors you can influence to avoid structural vulnerabilities due to AI.

Fig. 2 provides a template for these steps for a mindful design process. Step 1 entails a check on how controllers interact with technical components, but also with the social actors depicted in Fig. 1. Step 2 is a reminder to clarify the expectations regarding the benefits of introducing AI into the system. Step 3 and 4 directly incorporate the results of the analysis presented in Section 6 and of the previous discussion: As the introduction of AI tends to increase complexity, due to not executing a fixed programme, its implementation is the more challenging the more complex the overall system already is. For a less complex system, centralisation in the points of control allows for optimising the fulfilment of operational goals. If the system is rather complex, centralised control is not feasible due to reliability problems and decentralised points of control are preferable. The dynamicity of the system has to be considered both in operation and over longer time horizons. High operational dynamicity requires more and more varied data for the training of AI to widen the boundaries of safe operation. High structural dynamicity raises the question of how to facilitate the restructuring of the overall system. Finally, if the system is tightly coupled it is crucial to consider the aforementioned countermeasures and establish a flexible control hierarchy between the upper and lower levels of adaptive operations, ensuring that the lower levels are able to operate autonomously when necessary. In contrast, for loosely coupled systems the focus can remain on improving system wide-process modelling based on a central controller, i.e. upper-level adaptive operations and considering countermeasures.

The template offers orientation, but not a fixed rulebook on how to integrate AI-enhanced components into a socio-technical system. It is the result of an analysis indicating the importance of organising for safety when using AI, and a shift towards a mindful design process as a key path towards this goal. It can be used to reflect in advance on the consequences that the use of AI may have for a given system, and to check whether planned adjustments of the system can be expected to fully mitigate the system-wide effects of AI controllers.

8. Conclusion and outlook

In this work, we have presented an analysis to identify structural vulnerabilities resulting from the introduction of AI controllers into socio-technical systems and used the example of advanced microgrids. From this analysis, we have derived design principles for the AI enhancement of socio-technical systems that help avoid such structural vulnerabilities. Our perspective is informed by sociological approaches to organisational theory (NAT and HRO) and insights from an engineering point of view (STAMP). Building on these approaches, we have identified five descriptors, and their respective markers, that help to analyse a system with regard to its potential structural vulnerabilities. By accounting for the interaction between the descriptors, social actors, such as AI practitioners and domain experts, can design, maintain and organise a safer interplay between AI controllers, alternative solutions, e.g. non-AI-based methods, and other technical components. Our analysis reveals latent or unexpected trade-offs that can occur when systems are supposed to be enhanced by AI to operate autonomously. The implications for design processes include new perspectives on system dynamics and dimensions of adaptive capacity. As a result, the analysis identifies alternative design choices as well as potential research directions that suggest how certain structural vulnerabilities could be mitigated in the future.

Our study has limitations. In terms of technology, there are AI research directions that we have not extensively covered that could improve the interaction of descriptors in autonomously operating systems: If the decisions of the AI controllers are explainable or easier to

interpret, operators, AI practitioners and domain experts could better identify the boundary violation that led to a malfunction or disaster. Consequently, they could retrain the AI for a more appropriate process model, effectively extending the boundary conditions of adaptive operations.

In view of the entanglement of technology and social organisation, we have refrained from addressing event-based vulnerabilities as it would affect dynamics of a wider organisational context beyond the scope of design, implementation, and operation. In principle, it would be possible to include event-based vulnerabilities in the analysis, for instance by relating descriptors to disruptive events. Nevertheless, it is likely that the descriptors would need to be extended or modified for this purpose. Most importantly, we have not considered how the social design and technical control of AI-enabled autonomous systems are embedded in larger socio-technical systems. This would require the reflection of wider organisational processes as well as different stakeholders (e.g. consumers, users etc.). Research on interactive machine learning between AI practitioners and users (Amershi et al., 2014) presents a potential avenue for a more holistic design framework, involving an iterative design process where descriptor interactions could be applied, reviewed and restructured.

Overall, our analysis builds on the premises of organisational theories to present five key descriptors that capture the emergent effects on structural vulnerability of component interactions when a system is transformed by AI. Our approach can be used to guide the design of systems enhanced by AI controllers in order to organise the process for safety. Importantly, it also has the potential to open a conversation between multiple stakeholders with different backgrounds, allowing them to reflect on whether AI should be introduced into a seemingly well-understood system. Similarly to the interdisciplinary cooperation that resulted in this paper, we suggest that the inclusion of social science perspectives on reliability and safety would be helpful in this regard.

CRediT authorship contribution statement

Alexandros Gazos: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **James Kahn:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Isabel Kusche:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Christian Büscher:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Markus Götz:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Funding

The authors thank the funding agencies. This work was partially funded by the Helmholtz Program "Engineering Digital Futures", Germany.
Alexandros Gazos, Isabel Kusche, and Christian Büscher's work was supported by grants from the German Federal Ministry of Education and Research within the framework of the program "Research for Civil Security" of the Federal Government, Germany and the German Federal Ministry of the Interior, Germany (grant no. MOTRA-13N15218). James Kahn and Markus Götz's work was supported by the Helmholtz Association Initiative and Networking Fund under the Helmholtz AI platform grant, Germany.

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.

Appendix. Table of acronyms

Acronyms

EMS Energy Management System

FDC Fault Detection and Classification

HRO High Reliability Organization

NAT Normal Accident Theory

STAMP System-Theoretic Accident Model and Processes

STLF Short-Term Load Forecasting

References

Afrasiabi, M., Mohammadi, M., Rastegar, M., Stankovic, L., Afrasiabi, S., Khazaei, M., 2020. Deep-based conditional probability density function forecasting of residential loads. IEEE Trans. Smart Grid 11 (4), 3646–3657. http://dx.doi.org/10.1109/TSG.2020.2972513.

Alahmed, A.S., Al-Muhaini, M.M., 2020. An intelligent load priority list–based integrated energy management system in microgrids. Electr. Power Syst. Res. 185, 106404. http://dx.doi.org/10.1016/j.epsr.2020.106404.

Ali, S.S., Choi, B.J., 2020. State-of-the-art artificial intelligence techniques for distributed smart grids: A review. Electronics 9 (6), 1030. http://dx.doi.org/10.3390/electronics9061030.

Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T., 2014. Power to the people: The role of humans in interactive machine learning. AI Mag. 35 (4), 105–120. http://dx.doi.org/10.1609/aimag.v35i4.2513, Number: 4.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D., 2016. Concrete problems in AI safety. arXiv:1606.06565 [cs], arXiv:1606.06565.

Bagherian, A., Tafreshi, S.M., 2009. A developed energy management system for a microgrid in the competitive electricity market. In: 2009 IEEE Bucharest PowerTech. pp. 1–6. http://dx.doi.org/10.1109/PTC.2009.5281784.

Bansal, Y., Sodhi, R., 2018. Microgrid fault detection methods: reviews, issues and future trends. In: 2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia). pp. 401–406. http://dx.doi.org/10.1109/ISGT-Asia.2018.8467938.

Beheshtaein, S., Cuzner, R., Savaghebi, M., Guerrero, J.M., 2019. Review on microgrids protection. IET Generation Trans. Distrib. 13 (6), 743–759. http://dx.doi.org/10.1049/iet-gtd.2018.5212.

Bourakadi, D.E., Yahyaouy, A., Boumhidi, J., 2020. Multi-agent system based on the extreme learning machine and fuzzy control for intelligent energy management in microgrid. J. Intell. Syst. 29 (1), 877–893. http://dx.doi.org/10.1515/jisys-2018-0125.

Bower, W.I., Ton, D.T., Guttromson, R., Glover, S.F., Stamp, J.E., Bhatnagar, D., Reilly, J., 2014. The Advanced Microgrid. Integration and Interoperability. Technical Report SAND2014-1535, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), http://dx.doi.org/10.2172/1204100, URL https://www.osti.gov/biblio/1204100.

Brusaferri, A., Matteucci, M., Spinelli, S., Vitali, A., 2022. Probabilistic electric load forecasting through bayesian mixture density networks. Appl. Energy 309, 118341. http://dx.doi.org/10.1016/j.apenergy.2021.118341.

Burrell, J., 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc. 3 (1), http://dx.doi.org/10.1177/2053951715622512.

Büscher, C., 2022. The problem of observing sociotechnical entities in social science and humanities energy transition research. Front. Sociol. 6, http://dx.doi.org/10.3389/fsoc.2021.699362.

Cantu, J., Tolk, J., Fritts, S., Gharehyakheh, A., 2020. High reliability organization (HRO) systematic literature review: Discovery of culture as a foundational hallmark. J. Conting. Crisis Manag. 28 (4), 399–410. http://dx.doi.org/10.1111/1468-5973.12293.

Chitsaz, H., Shaker, H., Zareipour, H., Wood, D., Amjady, N., 2015. Short-term electricity load forecasting of buildings in microgrids. Energy Build. 99, 50–60. http://dx.doi.org/10.1016/j.enbuild.2015.04.011.

Chua, M., Kim, D., Choi, J., Lee, N.G., Deshpande, V., Schwab, J., Lev, M.H., Gonzalez, R.G., Gee, M.S., Do, S., 2023. Tackling prediction uncertainty in machine learning for healthcare. Nat. Biomed. Eng. 7 (6), 711–718. http://dx.doi.org/10.1038/s41551-022-00988-x.

Conn, A., 2015. Benefits & risks of artificial intelligence. URL https://futureoflife.org/ai/benefits-risks-of-artificial-intelligence/,

Dag, O., Mirafzal, B., 2016. On stability of islanded low-inertia microgrids. In: 2016 Clemson University Power Systems Conference. PSC, IEEE, Clemson, SC, USA, pp. 1–7. http://dx.doi.org/10.1109/PSC.2016.7462854.

Espín-Sarzosa, D., Palma-Behnke, R., Núñez Mata, O., 2020. Energy management systems for microgrids: main existing trends in centralized control architectures. Energies 13 (3), 547. http://dx.doi.org/10.3390/en13030547.

Fahim, S.R., Sarker, S., Muyeen, S.M., Sheikh, M., Das, S., 2020. Microgrid fault detection and classification: machine learning-based approach, comparison, and reviews. Energies http://dx.doi.org/10.3390/en13133460.

Fjelland, R., 2020. Why general artificial intelligence will not be realized. Humanit. Soc. Sci. Commun. 7 (1), http://dx.doi.org/10.1057/s41599-020-0494-4, Place: London, United Kingdom Publisher: Palgrave Macmillan.

French, R.M., 1999. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences 3 (4), 128–135. http://dx.doi.org/10.1016/S1364-6613(99)01294-2.

Gerwig, C., 2015. Short term load forecasting for residential buildings—An extensive literature review. In: Neves-Silva, R., Jain, L.C., Howlett, R.J. (Eds.), Intelligent Decision Technologies. In: Smart Innovation, Systems and Technologies, Springer International Publishing, Cham, pp. 181–193. http://dx.doi.org/10.1007/978-3-319-19857-6_17.

Gößling-Reisemann, S., Wachsmuth, J., Stührmann, S., Gleich, A.v., 2013. Climate change and structural vulnerability of a metropolitan energy system. The case of Bremen-Oldenburg in northwest Germany. J. Ind. Ecol. 17 (6), 846–858. http://dx.doi.org/10.1111/jiec.12061.

Gutierrez-Rojas, D., Nardelli, P.H.J., Mendes, G., Popovski, P., 2021. Review of the state of the art on adaptive protection for microgrids based on communications. IEEE Trans. Indus. Inform. 17 (3), 1539–1552. http://dx.doi.org/10.1109/TII.2020.3006845.

Hagendorff, T., 2020. The ethics of AI ethics: An evaluation of guidelines. Minds Mach. 30 (1), 99–120. http://dx.doi.org/10.1007/s11023-020-09517-8.

Harrold, D.J.B., Cao, J., Fan, Z., 2022. Renewable energy integration and microgrid energy trading using multi-agent deep reinforcement learning. Appl. Energy 318, 119151. http://dx.doi.org/10.1016/j.apenergy.2022.119151.

Hendrycks, D., Dietterich, T., 2019. Benchmarking neural network robustness to common corruptions and perturbations. http://dx.doi.org/10.48550/arXiv.1903.12261, URL http://arxiv.org/abs/1903.12261, arXiv:1903.12261 [cs, stat].

Hopkins, A., 1999. The limits of normal accident theory. Saf. Sci. 32 (2–3), 93–102. http://dx.doi.org/10.1016/S0925-7535(99)00015-6.

Jimeno, J., Anduaga, J., Oyarzabal, J., de Muro, A.G., 2011. Architecture of a microgrid energy management system. Eur. Trans. Electr. Power 21 (2), 1142–1158. http://dx.doi.org/10.1002/etep.443.

Johnson, B., 2022. Metacognition for artificial intelligence system safety – An approach to safe and desired behavior. Saf. Sci. 151, 105743. http://dx.doi.org/10.1016/j.ssci.2022.105743.

Khan, A.R., Mahmood, A., Safdar, A., Khan, Z.A., Khan, N.A., 2016. Load forecasting, dynamic pricing and DSM in smart grid: A review. Renew. Sustain. Energy Rev. 54, 1311–1322. http://dx.doi.org/10.1016/j.rser.2015.10.117.

Khazai, B., Kunz-Plapp, T., Büscher, C., Wegner, A., 2014. VuWiki: An ontology-based semantic wiki for vulnerability assessments. Int. J. Disaster Risk Sci. 5 (1), 55–73. http://dx.doi.org/10.1007/s13753-014-0010-9.

Khwaja, A., Zhang, X., Anpalagan, A., Venkatesh, B., 2017. Boosted neural networks for improved short-term electric load forecasting. Electr. Power Syst. Res. 143, 431–437. http://dx.doi.org/10.1016/j.epsr.2016.10.067.

Kusche, I., 2024. Possible harms of artificial intelligence and the eu ai act: fundamental rights and risk. J. Risk Res. 1–14. http://dx.doi.org/10.1080/13669877.2024.2350720.

La Porte, T.R., Consolini, P., 1991. Working in practice but not in theory: Theoretical challenges of "high-reliability organizations". J. Public Admin. Res. Theory http://dx.doi.org/10.1093/oxfordjournals.jpart.a037070.

Lekka, C., 2011. High Reliability Organisations - A Review of the Literature. Technical Report, Health and Safety Executive, URL https://api.semanticscholar.org/CorpusID:2214218.

Leveson, N.G., 2012. Engineering a safer world: Systems thinking applied to safety. In: Engineering Systems, MIT Press, Cambridge, MA, USA, http://dx.doi.org/10.7551/mitpress/8179.001.0001.

Leveson, N., Dulac, N., Marais, K., Carroll, J., 2009. Moving beyond normal accidents and high reliability organizations: A systems approach to safety in complex systems. Organ. Stud. 30 (02&03), 227–249. http://dx.doi.org/10.1177/0170840608101478.

Lin, H., Sun, K., Tan, Z., Liu, C., Guerrero, J.M., Vasquez, J.C., 2019. Adaptive protection combined with machine learning for microgrids. IET Generation Trans. Distrib. 13 (6), 770–779. http://dx.doi.org/10.1049/iet-gtd.2018.6230.

Meng, L., Sanseverino, E.R., Luna, A., Dragicevic, T., Vasquez, J.C., Guerrero, J.M., 2016. Microgrid supervisory controllers and energy management systems: a literature review. Renew. Sustain. Energy Rev. 60, 1263–1273. http://dx.doi.org/10.1016/j.rser.2016.03.003.

Perrow, C., 1999. Normal Accidents: Living with High Risk Technologies - Updated Edition, REV - Revised Princeton University Press, http://dx.doi.org/10.2307/j.ctt7srgf.

Ray, P., Biswal, M. (Eds.), 2020. Microgrid: Operation, control, monitoring and protection. In: Lecture Notes in Electrical Engineering, vol. 625, Springer, Singapore, http://dx.doi.org/10.1007/978-981-15-1781-5, URL http://link.springer.com/10.1007/978-981-15-1781-5.

Rodrigues, Y., Monteiro, M., Abdelaziz, M., Wang, L., de Souza, A.Z., Ribeiro, P., 2020. Improving the autonomy of islanded microgrids through frequency regulation. Int. J. Electr. Power Energy Syst. 115, 105499. http://dx.doi.org/10.1016/j.ijepes.2019.105499.

Salovaara, A., Lyytinen, K., Penttinen, E., 2019. High reliability in digital organizing: mindlessness, the frame problem, and digital operations. MIS Quarterly 43 (2), 555–578. http://dx.doi.org/10.25300/MISQ/2019/14577.

Sawyerr, E., Harrison, C., 2019. Developing resilient supply chains: lessons from high-reliability organisations. Supply Chain Manag.: Int. J. 25 (1), 77–100. http://dx.doi.org/10.1108/SCM-09-2018-0329.

Schulman, P.R., Roe, E., 2007. Designing infrastructures: Dilemmas of design and the reliability of critical infrastructures. J. Conting. Crisis Manag. 15 (1), 42–49. http://dx.doi.org/10.1111/j.1468-5973.2007.00503.x.

Schwiderowski, J., Beck, R., 2023. Mindful design and operation for high reliability autonomous systems. In: ECIS 2023 Research Papers.

Searle, J.R., 1980. Minds, brains, and programs. Behav. Brain Sci. 3 (3), 417–424. http://dx.doi.org/10.1017/S0140525X00005756.

Streck, R., 2021. Europa ist am blackout vorbeigeschrammt. URL https://www.heise.de/tp/features/Europa-ist-am-Blackout-vorbeigeschrammt-5028090.html.

Tazi, K., Abbou, F., Abdi, F., 2020. Multi-agent system for microgrids: design, optimization and performance. Artif. Intell. Rev. http://dx.doi.org/10.1007/s10462-019-09695-7.

Thiebes, S., Lins, S., Sunyaev, A., 2020. Trustworthy artificial intelligence. Electron. Markets http://dx.doi.org/10.1007/s12525-020-00441-4.

Timmermans, S., Tavory, I., 2012. Theory construction in qualitative research: From grounded theory to abductive analysis. Sociol. Theory 30 (3), 167–186. http://dx.doi.org/10.1177/0735275112457914, Publisher: SAGE Publications Inc.

Ton, D.T., Smith, M.A., 2012. The U.S. department of energy's microgrid initiative. Electr. J. 25 (8), 84–94. http://dx.doi.org/10.1016/j.tej.2012.09.013.

Tsymbal, A., 2004. The problem of concept drift: definitions and related work. URL https://api.semanticscholar.org/CorpusID:8335940.

Varshney, K.R., Alemzadeh, H., 2017. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. Big Data 5 (3), http://dx.doi.org/10.1089/big.2016.0051, Publisher: Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.

Vela, D., Sharp, A., Zhang, R., Nguyen, T., Hoang, A., Pianykh, O.S., 2022. Temporal quality degradation in AI models. Sci. Rep. 12 (1), 11654. http://dx.doi.org/10.1038/s41598-022-15245-z, Publisher: Nature Publishing Group.

Venkatanagaraju, K., Biswal, M., 2020. Mitigation of power system blackout with microgrid system. In: Ray, P., Biswal, M. (Eds.), In: Microgrid: Operation, Control, Monitoring and Protection, vol. 625, Springer Singapore, Singapore, pp. 307–332.

Wang, Y., Gan, D., Sun, M., Zhang, N., Kang, C., Zongxiang, l., 2019a. Probabilistic individual load forecasting using pinball loss guided lstm. Appl. Energy 235, 10–20. http://dx.doi.org/10.1016/j.apenergy.2018.10.078.

Wang, T., He, X., Deng, T., 2019b. Neural networks for power management optimal strategy in hybrid microgrid. Neural Comput. Appl. 31, http://dx.doi.org/10.1007/s00521-017-3219-x.

Wei, X., Xiangning, X., Pengwei, C., 2018. Overview of key microgrid technologies. Int. Trans. Electr. Energy Syst. 28 (7), e2566. http://dx.doi.org/10.1002/etep.2566.

Weick, K.E., Sutcliffe, K.M., 2006. Mindfulness and the quality of organizational attention. Organiz. Sci. 17 (4), 514–524. http://dx.doi.org/10.1287/orsc.1060.0196.

Weick, K.E., Sutcliffe, K.M., 2015. Managing the Unexpected: Sustained Performance in a Complex World, third ed. Wiley, Hoboken (N.J.).

Weick, K.E., Sutcliffe, K.M., Obstfeld, D., 2008. Organizing for high reliability: Processes of collective mindfulness. In: Crisis Management. In: Sage library in business & management, SAGE Publications, Los Angeles, pp. 31–66.

Widmer, G., Kubat, M., 1996. Learning in the presence of concept drift and hidden contexts. Mach. Learn. 23 (1), 69–101. http://dx.doi.org/10.1023/A:1018046501280.

Witsch, K., 2021. Handelsblatt energie-gipfel: Kurz vor blackout: Europas stromnetz wäre im januar fast zusammengebrochen. URL https://www.handelsblatt.com/unternehmen/energie/handelsblatt-energie-gipfel-kurz-vor-blackout-europas-stromnetz-waere-im-januar-fast-zusammengebrochen/26820168.html.

Yamashita, D.Y., Vechiu, I., Gaubert, J.-P., 2020. A review of hierarchical control for building microgrids. Renew. Sustain. Energy Rev. 118, 109523. http://dx.doi.org/10.1016/j.rser.2019.109523.

Yang, Y., Hong, W., Li, S., 2019. Deep ensemble learning based probabilistic load forecasting in smart grids. Energy 189, 116324. http://dx.doi.org/10.1016/j.energy.2019.116324.

Yoldas, Y., Önen, A., Muyeen, S.M., Vasilakos, A.V., Alan, I., 2017. Enhancing smart grid with microgrids: Challenges and opportunities. Renew. Sustain. Energy Rev. 72, 205–214. http://dx.doi.org/10.1016/j.rser.2017.01.064.

Young, W., Leveson, N.G., 2014. An integrated approach to safety and security based on systems theory. Commun. ACM 57 (2), 31–35. http://dx.doi.org/10.1145/2556938.

Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I., 2019. Theoretically principled trade-off between robustness and accuracy. http://dx.doi.org/10.48550/arXiv.1901.08573, URL https://arxiv.org/abs/1901.08573v3.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proc. IEEE PP, 1–34. http://dx.doi.org/10.1109/JPROC.2020.3004555.

Zia, M.F., Elbouchikhi, E., Benbouzid, M., 2018. Microgrids energy management systems: a critical review on methods, solutions, and prospects. Appl. Energy 222, 1033–1055. http://dx.doi.org/10.1016/j.apenergy.2018.04.103.

Zio, E., 2016. Challenges in the vulnerability and risk analysis of critical infrastructures. Reliab. Eng. Syst. Saf. 152, 137–150. http://dx.doi.org/10.1016/j.ress.2016.02.009.

Zor, K., Timur, O., Teke, A., 2017. A state-of-the-art review of artificial intelligence techniques for short-term electric load forecasting. In: 2017 6th International Youth Conference on Energy. IYCE, IEEE, Budapest, Hungary, pp. 1–7. http://dx.doi.org/10.1109/IYCE.2017.8003734, URL http://ieeexplore.ieee.org/document/8003734/.