Advances in Reliability, Safety and Security, Part 10 – Kołowrocki, Kosmowski (eds.)
 ISBN 978-83-68136-22-7 (printed), ISBN 978-83-68136-09-8 (electronic)

Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

XAI And Cyber Resilience: Sociotechnical Perspective

Dorthea Mathilde Kristin Vatn, Tor Olav Grøtan

SINTEF Digital, Trondheim, Norway
Norwegian University of Science and Technology, Trondheim, Norway

Abstract

As critical industries are increasingly dependent on digital technologies, new cyber-physical systems are created that are exposed to new forms of failure and malicious cyberattacks. The digitalization of these industries has given access to massive amounts of data that in combination with increased computational power has propelled the use of AI. While the first AI-models where easy to interpret, the emergence of more sophisticated and complex machine learning (ML) models based on deep neural networks have led to the emergence of the term "black-box"-models, to describe models producing decisions and predictions that are hard to explain and understand. Explainable AI (XAI) as a class of approaches that provide visibility into the decisions and predictions of an AI system has therefore gained increased interest. Within the field of cybersecurity XAI is predicted to have disruptive effects, both on operational cybersecurity models and on risk management approaches. Despite technological progress and more sophisticated measures and risk reduction strategies to counteract threats, possible cyberattacks and their related consequences can never be fully anticipated and modelled, pointing towards the need for organizations to be able to adapt to handle fundamental surprise. This makes cyber resilience highly relevant, incorporating an adaptive capacity enabling critical infrastructure organizations to handle fundamental shocks and surprises. In this paper we present a preliminary framework for how XAI might contribute to cyber resilience building on Grøtan et al.'s (2022) perspective on cyber resilience. Building on a sociotechnical perspective, we outline how XAI as a sociotechnical construct might relate to cyber-resilience through a bi-directional relationship. We first use examples from literature and practitioners press to discuss how XAI might contribute to cyber resilience through mitigation of bias, enhancing trust and improving decision-making. Then we secondly suggest that data from various sources of a critical infrastructure system could be modelled through XAI techniques and by this contribute to model real-time constitutions of adaptive capacity. We conclude by suggesting some future directions for research on the topic of XAI and cyber resilience.

Keywords: cyber resilience, adaptive Capacity, explainable AI, XAI, human-centred AI

1. Introduction

The industrialized world is increasingly dependent on digital technologies that are complicated, brittle, and fragile and could exhibit dynamics as well as failure modes beyond what they are designed for (Grøtan et al., 2022). While digital transformation is driven by the quest for efficiency gains and innovative operations modes, it also creates new cyber-physical systems that are exposed to new forms of failure and malicious cyberattacks, especially when digital transformation is implemented in critical infrastructure. The digitalization process across domains has given access to massive amounts of data and in combination with increased computational power this has propelled the use of AI. Within the field of cybersecurity AI is predicted to have disruptive effects, both on operational cybersecurity models and on risk management approaches (Brooks, 2023). While AI might be used to propel harmful cyber-attacks, there is also an optimism connected to the use of AI within the field of cybersecurity as AI can enable organizations to process and interpret vast amounts of data from various sources for early identification of anomalies. The U.S. Department of Energy Office of Cybersecurity, Energy Security and Emergency Response has established several projects that use AI to automate security vulnerability and patch management in energy delivery systems as well as to enhance the situational awareness of energy delivery systems to ensure uninterrupted flows of energy (DOE-CESER, 2020ab). AI can serve as a useful tool for several cybersecurity purposes (Aziz Al Kabir et al, 2023; Moustafa et al., 2023). AI can be used for threat

detection and analysis, malware detection as well as intrusion detection and prevention. By going through large datasets and as such being able to detect patterns and anomalies that can point towards known attack signatures, cyber-attacks might early be detected. Also, by monitoring network traffic in real time AI systems might be able to both recognize and react to potential security breaches through early detection of suspicious activity. Adding to this, by improving user authentication procedures, AI tools might make it more challenging for unauthorized users to access systems and confidential data. AI-tools can also be useful to identify vulnerabilities in software and systems by scanning and analysis of infrastructure, code, and configurations. On the threat intelligence level, AI tools can be used to analyze large volumes of data from various sources to spot threats, trends, and possible vulnerabilities. Possible interesting sources for analysis could be open forums on the internet and the dark web, and insight into these sources might enhance an organization's ability to undertake proactive defense measures. By AI systems' ability to continuously learn from new data, security systems based on AI might efficiently be able to detect and respond to new attack vectors, as well as use historical data to forecast security risks and trends (Ramya et al., 2023; Pradeesh, 2023).

While the first AI-models where easy to interpret, the emergence of more sophisticated and complex machine learning (ML) models based on deep neural networks have led to the emergence of the term "black-box"-models, to describe models producing decisions and predictions that are hard to explain and understand (Arrieta et al., 2020). There is often a trade-off between the effectiveness of a model and the degree of explainability possible to extract from the model. This has led to an increasing interest in explainable AI (XAI) as a class of approaches that provide visibility into the decisions and predictions of an AI system (Rai, 2020). XAI is described as a research discipline that proposes different ML techniques that produce explainable models while at the same time maintaining a high level of learning performance (Adadi & Berrada, 2018; Rai, 2020). Within the domain of cybersecurity XAI as opposed to black-box models could have several benefits, and recently Fowler (2023) outlined how XAI techniques might augment and enhance professionals working with cybersecurity and contribute to AI-driven autonomous responses that include the human-in-the-loop and as such ensure knowledge transfer between human and AI and enhance trust and improve decision-making. According to Fier (2022) breakthroughs in the work within the field of AI and cybersecurity will most likely not be purely technical in terms of advanced mathematical algorithms alone, but through enabling methods and processes that allow human users to comprehend and trust the results and decisions created by machine learning (Kuppa & Le-Khac, 2020). Explainability is also a critical key factor ensuring the ethical use of AI in cybersecurity, in line with work done by the high-level expert group on trustworthy AI in EU. This is also underscored in practitioners press, where opacity in AI-decisions might raise concerns about accountability in cybersecurity (Bansal, 2023). Given the possible positive impact XAI might have in the domain of cybersecurity, several survey articles have been published on the topic (e. g. Charmet et al., 2022; Zhang et al., 2022). However, a general limitation that becomes visible in these surveys is that they do not take into account the diversity of stakeholders involved in cybersecurity and their different explanation needs, and as such forget the sociotechnical aspect of XAI (Rjoub et al., 2023). This is an aspect that also can be seen in context of there being a general lack of organizational perspectives on XAI (Brasse et al., 2023).

Resilience, the ability to absorb, adapt, and effectively respond to change, is a concept that is increasingly applied and wanted in many areas, in cyber security, but also for services relying on cyber security. The term cyber resilience has been used to signify the shortcomings, and even the "death", of cyber security practices (Baukes, 2017), arguing that the latter relies on a “set it and forget it” model, for which "even the most seemingly impregnable of such barriers are laid down, hackers will be able, with time, to build a higher ladder". In contrast, it is argued that cyber resilience is the intelligent means of managing and mitigating cyber risk, requiring best practices to be followed every day. Grøtan et al. (2022) distinguishes between various origins of cyber resilience as a process, namely technology, risk-informed preparedness, and sociotechnical, adaptive capacity. While the two formers easily may be associated with a "set and forget" approach, the sociotechnical adaptive capacity implies continuous attention to daily operations and their reliance on IT, but, importantly, also beyond a "best practice" perspective. For cyber resilience as sociotechnical adaptive capacity, there is no "best practice" unless it also accommodates the intrinsic ability of instant adaptation and change in complex environments.

This paper aims therefore to explore how XAI as a sociotechnical concept relates to sociotechnical cyber resilience in organizations by using examples from practitioners press and current literature to suggest a preliminary framework for XAI and cyber resilience in organizations. The research question forming the basis of the paper is twofold:

1) *Through which mechanisms can XAI as a sociotechnical concept be related to cyber resilience in organizational contexts?*
2) *Can XAI techniques be used to support a specific facet of cyber resilience, namely adaptive capacity, in critical infrastructures?*

The paper is structured to first present a preliminary framework for how XAI relates to cyber resilience building on Grøtan et al.'s (2022) sociotechnical perspective on cyber resilience, in which adaptive capacity is a key aspect. Also, building on Leavitt's (1964) sociotechnical perspective we outline how XAI as a sociotechnical construct might relate to cyber-resilience through a bi-directional relationship. We use examples from literature and practitioners press to discuss how XAI might contribute to adaptive capacity through mitigation of bias, enhancing trust and improving decision-making. Then we also suggest that data from various sources of a critical infrastructure system could be modelled through XAI techniques and by this contribute to model real-time constitutions of adaptive capacity. We conclude by suggesting some future directions for research on the topic of XAI and cyber resilience.

2. A sociotechnical framework for XAI and cyber resilience

With increasing digitalization of critical infrastructures, development of protective measures is a constant and ongoing process to ensure cybersecurity. The optimism and increasing use of XAI as a protective measure in cybersecurity might be seen as part of this ongoing and never-ending effort to develop protective measures to withstand new and surprising threats. Technology introduced in critical systems needs to be combined with both old and new technologies, as well as align with given tasks, structures, and different actors within the system. This points towards the usefulness of looking at XAI through a sociotechnical lens in the context of cyber resilience. Figure 1, partially inspired by Lyytinen & Newman (2008, p. 594), illustrates how XAI treated as a sociotechnical subdimension within the field of AI might relate to sociotechnical cyber resilience. In a sociotechnical perspective, organizational systems can be seen as consisting of four aligned and interacting components encompassing technology, actors, tasks, and structures (Leavitt, 1964; Lyytinen & Newman, 2008). Framing XAI as a sociotechnical concept could mean that it could be placed in the technology box of the sociotechnical arrangement. However, underscoring that XAI technology would need to be aligned with already present technology in the system makes it natural to keep it as an independent construct in our framework. Keeping it as an independent construct also underscores the inherent challenge facing critical infrastructure where information technology (IT) and operational technology (OT) systems are blended in new ways. Bringing in a sociotechnical perspective will also embrace the actor and stakeholder perspective that is essential in the context of XAI, as the explainability part points towards a concrete recipient of information from the underlying AI-model. In the context of cybersecurity, the recipient of information of the model could be very different depending on the context. It could be a cybersecurity expert with deep technical expertise, or it could be a decision maker working at a greater distance from the technical system. Bringing in a sociotechnical perspective enables us to differentiate between the possible different explanation needs different actors might have.

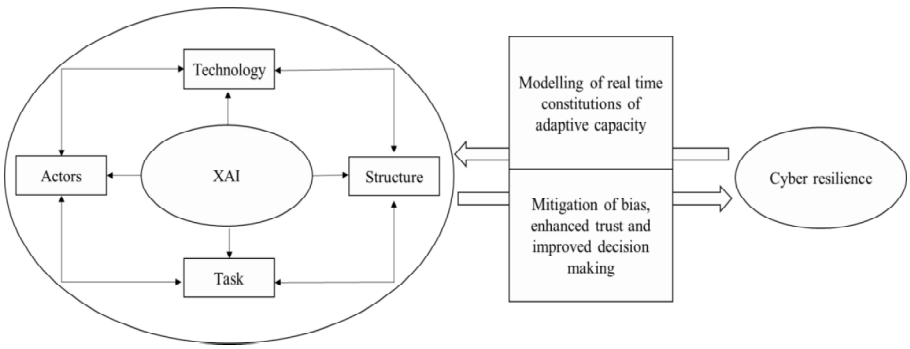

Fig 1. XAI as a sociotechnical construct and how it might relate to cyber resilience.

The sociotechnical perspective is useful not only when understanding XAI and its organizational impacts. Cybersecurity could also be framed as a sociotechnical challenge because the ability to withstand cyberattacks is not only a technical matter, but also a highly organizational matter of paying due attention to humans as resources (Grøtan et al., 2022). By bringing in a sociotechnical perspective on cybersecurity, the term resilience as the ability to adapt becomes highly relevant (Woods, 2018; Hollnagel, 2016). Possible cyber-attacks and their related consequences can never be fully anticipated and modelled, pointing towards the need for organizations to be able to adapt to handle fundamental surprise. As critical sectors increasingly adopt digital technologies several gains related to both safety and efficiency are important outcomes. However, the connection to the global

internet also implies increased threats from the whole world, as well as increased complexity stemming from the integration of IT and OT (Pettersen & Grøtan, 2024). The increased complexity and threat landscape have made scholars point towards a Strategic Agility Gap that organizations might suffer from (Woods & Alderson, 2021; Pettersen & Grøtan, 2024). This gap is arising because the increased complexity and new threat landscape challenges the ability of organizations to be prepared through traditional risk reduction strategies. This points towards the usefulness of an adaptive capacity enabling critical infrastructure organizations to handle fundamental shocks and surprises.

Building on Grøtan et al. (2022) theoretical concept of cyber resilience, we distinguish between three operationalisations of cyber resilience denoted as "Theory A", "Theory B" and "Theory C". Theory A is seen as an intrinsic part of a system that makes it fault-tolerant and can be depicted as "technological resilience". An example could be a system with a feedback control loop that balances gain and performance within a defined envelope of variability and resources. Use of AI to autonomously withstand certain cyber-attacks without a human in the loop could be regarded an example of Theory A. Theory B conceptualizes resilience as a repository of organised supportive resources designed to facilitate maintenance of function, in terms of robustness that absorbs or withstands disturbances, or the ability to rebound from a disturbance that has led to loss of function. Lastly, Theory C depicts resilience as underlying principles and conditions enabling organizations to adapt resiliently to situations that those designing procedures and allocating resources have not envisaged.

The distinction between Theory B and Theory C is illustrated in Figure 2, and the distinction is further nuanced by bringing in the term ***operational resilience*** versus ***organizational***[1] ***resilience***. This distinction indicates the difference in ***origin*** of resilient performance. The rules and procedures that are (legitimate to be) obeyed, modified and bent from a Theory B perspective originate from formal and institutionalized organizational processes of, e.g., risk management, emergency preparedness and business continuity. At the other side, when needed, the further bending and action beyond the rule originate from an adaptive capacity that depends on "real-time" situated experience and action, enabling what Woods (2018) denote ***graceful extensibility***. Adaptive capacity thus denotes the sustained ability to be poised for graceful extensibility. Hence, while adaptive capacity also depends on the same resources, priorities and policies allocated a priori through the organizational Theory B processes, adaptive capacity signifies the process of using and adapting these resources in a different manner, based on the situation at hand. In which, human traits like initiative, reciprocity, timing and rhythm across preconceived organizational structures and levels, steps into the foreground, and the formal organization moves to the background. This somewhat resembles the dynamic of High Reliability Organizations (HRO) which are able to shift to an alternative way of operation, when needed (Rosness et al., 2010). However, a difference is that in Woods' (2018) case, this shift is more precarious related to a more unpredicted situation. There is no "blueprint" for the alternative mode, it must emerge out of the specific situation. The combination of both an organizational and operational perspective is useful for defining cyber resilience. Such a view underscores the need for rules and procedures to accommodate cyber-threats as well as the importance of the ability to bend these rules and procedures when necessary for restoring and maintaining functionality. Also, this view encompasses that the surprising nature of possible cyber-attacks might demand acting beyond the rules to restore functionality, but knowing ***when*** to stop is at the core of operational resilience (Grøtan et al., 2022).

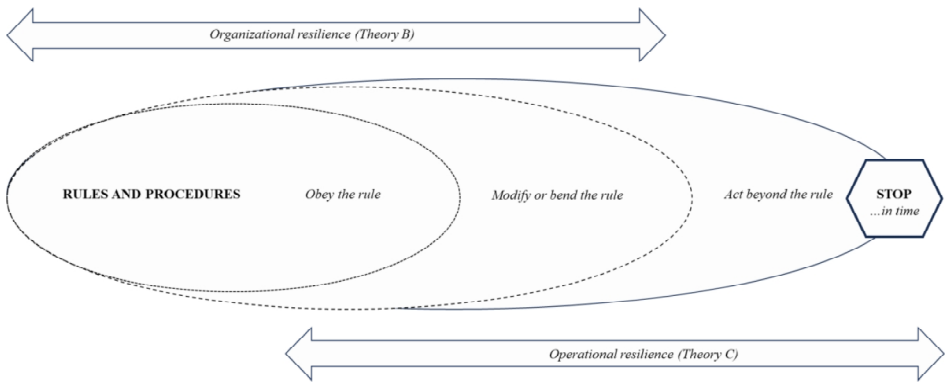

Fig 2. Organizational vs operational cyber resilience (adapted version of Grøtan et al. 2022)

[1] We are aware that the term may be associated with the resilience of the organization as an outcome. However, in this paper, "organizational" points to the (formal, institutionalized) organization as an origin and context for the process producing cyber resilience.

As highlighted by Pettersen and Grøtan (2024), Theory B and Theory C must be recognized a s fundamentally different, but at the same time inextricably intertwined. Seeing them separately, it is much more straightforward to conceive a role for XAI in association with Theory B, than for Theory C. A "recipe" for robustness and rebound towards explicated disturbances is something that AI can learn, given the data available. Adaptive capacity towards the unknown is something completely different, embedding a dimension of novelty and uniqueness, including the inherent, but unknown limits to this capacity. Hence, the AI must also "learn" about limiting concepts, e.g., what Woods and Branlat (2011) denote "adaptive traps".

As the point of departure for this paper is that AI cannot support decisions on cyber resilience without human intervention, we will now explore how XAI can support cyber resilience as a combination of Theory A, B and C.

Figure 1 illustrates how XAI might relate to cyber resilience through two different processes. First, we suggest that XAI might relate to cyber resilience through mitigating bias, enhancing trust, and improving decision-making. This means that in the context of cybersecurity, AI-tools that are able to provide explanations for its decisions and actions to relevant stakeholders might be beneficial compared to black-box models to foster cyber resilience in a Theory B perspective. Second, we propose the idea that sensor data collected from critical infrastructure could be used to train an AI-model that through XAI techniques could provide insight into real time constitutions of adaptive capacity. Next, we elaborate on these two suggested relationships between XAI and cyber resilience by using examples from current literature.

2.1. Cyber resilience fostered through XAI that mitigates bias, enhances trust and improves decision-making

By pointing towards how XAI contribute to cyber resilience through mitigation of bias, enhancement of trust and improvement of decision-making, we provide suggestions on how XAI as opposed to black-box algorithms contribute to cyber resilience through a Theory B perspective. Ensuring explainability in technological tools (Theory A) might be considered an integrated aspect of organizational cyber resilience, Theory B.

In recent literature the usefulness of research on detection algorithms that enable systems to both detect several types of large-scale attacks as well as build trust in the model's performance through XAI approaches is underscored by several (Muna et al., 2023; Zolanvari et al., 2021; Moustafa et al., 2023). AI is often used as an important part of these Intrusion Detection Systems (IDS) used to identify malicious network activity before it compromises information, availability, integrity or confidentiality (Mahbooba et al., 2021). The primary goal of developing such systems is to detect suspicious network activity that a standard firewall cannot detect (Sivamohan & Sridhar, 2022). These IDS systems could be based on complex ML-models based on deep learning, giving them the characteristics of black boxes making it hard to understand them. However, by including XAI techniques in these systems network traffic and telemetry data of IoT devices might be analyzed, potential threats might be identified, and the humans in the system might be given an explanation of the system's findings. This could be seen as a prerequisite for enabling strategic moves to take action effectively and make moves that reduce the negative impact of possible cyber-attacks. Lack of confidence in the decision-making process of automated deep learning-based IDS might negatively impact the root cause analyses of detected cyberattacks and as such prevent important learning processes that could make the organization better equipped to withstand future cyber-attacks (Moustafa et al., 2023). This points towards the usefulness of incorporating XAI techniques that could benefit both non-technical decision-makers in addition to cyber security experts.

XAI techniques could support human decision making by fostering trust in the output of AI. Being able to make sound judgements and take responsibility for decisions, especially within critical sectors, requires trust in the output of an AI tool. XAI might give a better understanding of attack behaviors also for non-technical experts and as such facilitate discussion and dialogue between decision makers and security experts within an organization. In a resilience perspective, based on Theory B, decision makers are important providers of terms and premises through the framework conditions they create. These framework conditions however also enable or inhibit the intended principles and conditions for sustained resilient capabilities, underlying operational cyber resilience in our suggested framework. If decision makers through XAI can understand why certain events are flagged as malicious it is also easier to enable dialogue between cyber security experts and decision makers, and as such facilitate that cyber defense solutions efficiently are built into the cybersecurity strategies of organizations. Also, for cybersecurity experts, XAI might help to identify the specific networks, features and security policies that are compromised by attackers. By this, proper action might be taken, whether it means debugging the IDS model itself, or applying new security policies to prevent similar attacks in the future (Mahbooba et al., 2021).

In addition to improving the preconditions for decision-making, XAI might contribute to the removal of ***bias*** from models that constitute the IDS. XAI techniques might present the patterns constructed by the model, which helps data scientists and developers to analyze and remove irrelevant patterns in datasets. Also, XAI techniques

might contribute to counteract automation bias, defined as the tendency to use automated cues as a heuristic replacement for vigilant information seeking and information processing (Schemmer et al., 2022). In the domain of cybersecurity placing all trust in what an AI-tool might be able to detect might cause one to miss important signs of an ongoing attack that the AI tool has not been able to detect. Systems built on XAI that are able to engage operators to do their own assessment in addition to the ones that are done by an AI-tool might in this perspective be beneficial. Designing systems based on XAI could by this enable human actors to compensate for technological errors (Jussupow et al., 2021). Also, XAI-techniques might contribute to more precise detecting intrusion compared to other methods and as such reduce both complexity and enhance detection accuracy in the learning predictions (Sivamohan & Sridhar, 2023). Increased precision in IDS might also be beneficial for security operators that often might be overwhelmed with vast amounts of security alerts every day, and as such reduce alert fatigue (Charmet et al., 2022). Also, moving from explainability in systems designed for cybersecurity specialists, there are some examples of studies that investigate the explainability of AI-based cybersecurity tools more from a user-perspective of the general organizational member. Several large-scale cyber-attacks have begun with phishing e-mails to an unexpected victim. XAI-based solutions to classify phishing emails might increase the effectiveness of warning dialogs for guarding users against phishing attacks. The increased effectiveness is achieved through anti-phishing warnings that give the users ability to understand the reasons why the system considers a message or website suspicious (Greco et al., 2023).

We have until now been looking at resilience from a structural organizational perspective, through the lens of Theory B. We have illustrated how use of XAI as opposed to black-box models could contribute as part of the repository of organized, supportive resources designed to uphold function when there is loss of functionality due to a cyber-attack. However, this way of framing resilience limits resilience to what one in risk management can anticipate, deduce, and operationalize based on existing knowledge, past experience and external input. Therefore, it is also useful to look at resilience from an *operational perspective* in terms of underlying principles and conditions for sustained resilient capabilities as this perspective addresses an adaptive capacity. Therefore, we now move on to suggest how XAI techniques could be used to model adaptive capacity in a critical system.

2.2. Cyber resilience modelled through XAI providing insights into the real time constitutions of adaptive capacity

A mantra of the Resilience Engineering approach since its inception is to pay attention to and learn from why operations go right, rather than from only their flip side – why they go wrong (Hollnagel et al, 2006). This is easily said, but harder done, as the ratio of successful operations outperforms the ratio of failure with huge scales. AI outperforms human capacity to analyze big amounts of data, and in principle, we cannot rule out the possibility that AI and deep learning may reveal significant patterns that humans are not able to see. Nevertheless, there are numerous reasons why we cannot presume that data availability – or relevance – is unlimited. Moreover, for our purpose, we must be able to express "adaptive capacity" as a learning target for AI, before we can address the explainability of the models generated, which in any case will be a matter of translation from the data scientist to the organizational actor accountable for the decisions made. For the latter, preunderstanding of adaptive capacity will inevitably influence interpretations and decisions. Hence, it is necessary to make some initial delimitations regarding data capture, based on presumptions derived from conceptualizations of adaptive capacity.

Adaptive capacity is something that is sustained over time (Woods, 2018). A singular, standalone adaptation is therefore merely an indication of adaptive capacity in the moment, but not sufficient evidence of its sustained presence. Woods' (2018) notion of being "poised to adapt" is a more forward-looking, supplementary indication of sustained adaptive capacity. What we should aim to reveal is therefore the "stories of success" hidden in the data, not only on singular sociotechnical adaptations, but also on the continuity of successful adaptations, and the predictive value for being poised to adapt, from a series of successful adaptations. Such stories can be conceived as a series of situated adjustments that are not foreseen or do deviate from plans and procedure, and that convey a sense of surprise, uncertainty, unpredictability, or of urgency. This may encompass performance variability in many forms, such as approximate adjustments due to time pressure (Hollnagel, 2016), deliberate adaptations of own practice due to circumstance, adoptions of others' adaptations (as for software patches in urgent response to zero-days), or graceful extensibility in which several practices are extended beyond their presumed boundaries, separately or in coordination (Woods, 2018). The data capture should also include the event of a breakdown within a series of successful adaptations, to capture, e.g., breakdown due to functional resonances in which performance variabilities amplify each other (Hollnagel, 2016), or other breaking points. By combining the knowledge of prior successes with the knowledge about the arrival of the breaking point, key information about the sustainability of adaptive capacity may be captured. This insight can in turn be combined with existing models of adaptive traps (Woods & Branlat, 2011), or be used to build additional models.

So far, the discussion has limited our attention to explicit actions related to concrete situations demanding adjustments or adaptations from a safety or security point of view. Presumably, there is a potential for using data related to these actions for building models of adaptive capacity by generative AI, and to interpret XAI models by means of the concepts mentioned above. However, we do not know the full story why people choose their actions. Actions on cyber systems are also a result of enactment towards intangibles. According to Weick (1988), the term enactment means that when people act under pressure, they bring structures and events into existence and set them in action. The process of structures and events brought into existence and subsequently acted upon is beyond the reach of information audited from cyber events, hence we must search for proxies. This point of view also justifies a warning: it might be that an XAI-generated model of adaptive capacity drawn from explicit, audited action might have a bias, a blind spot in terms of background organizational dynamics and dynamisms in which enactment is a main part. Subsequently, we must ask how the presence of XAI-models of adaptive capacity will relate to this organizational dynamism. Is the latter "doomed" to remain a blind spot for which the XAI-models will never catch up? As stated above, our best hope may be to identify some relevant proxies that can supplement the generative modelling.

A possible route to such proxies may be inspired by Power's (2016) coining of the term "Riskwork". Riskwork points beyond the frameworks and designs to avoid risk, towards the negotiation of risk objects and values, conflict, emotion and practice, and the micro-sociology of risk management. In other words, even if "risk" does not materialize, it is worked upon. Power (2016) argues that "risk is expressed by contingencies of future possibilities which have yet not crystallized into events". As such, risks do not exist, but are yet "seen" by being re-presented and processed, in what he denotes as "various apparatuses for their management". Moreover, Power (2016) argues that representation of risk not only is a philosophical necessity, but also has has a sociological correlate, as "practices are littered with artefacts, many of which contain and inscribe representations of risk". Risk is, by implication, therefore inscribed in artefacts used in organizational work, such as models, maps and metrics. This resonates with a paradoxical statement ascribed to Karl E. Weick: "when nothing happens, a lot is happening". An overall objective with Riskwork, with special significance for our purpose, is thus to move attention "*from the formal front stage of practice – the frameworks and protocols which characterize risk management texts and training manuals – to the backstage of negotiation, network building, emotional commitment, entanglement with representational devices, and the inevitable everyday conflicts of organizational life*" (Power 2016, p. 17).

This way, the Riskwork perspective inspires us to point out another empirical arena from which data for a proxy can be gathered for the purpose of generating models of adaptive capacity. This opportunity is not about access to people's mind, but about the traces they leave behind in what Power (2016) coins "artefacts and representational devices". In our view, backstage "artefacts and representational devices" in daily use will mirror people's enactments no less than their observable adherence, adjustments or violations of frontstage rules and protocols. Put more modestly, they may at least provide an equally suitable basis for calculation of correlations, which is the engine of generative AI. Or, put more bluntly, food for the "stochastic parrots" (Bender et al., 2021).

Hence, if artefacts inscribed with risk are necessary for the sustained existence of an organization that is able to manage the risks it presumably is exposed to as argued by Power (2016), a similar argument may be raised for adaptive capacity. That is, also adaptive capacity will need some form of representation to be addressed continuously through organizational processes. Hence, as soon as XAI models incorporating "explanations" of adaptive capacity become available, they will inevitably be positioned "frontstage", while the sociotechnical processes of adaptation will have a parallel, backstage arena as soon as the very idea of adaptive capacity is inscribed into other representational devices in daily use. Importantly, artefacts are not neutral mirrors of risk, nor passive intermediaries of information. Power (2016) argues that they must be studied as potentially powerful mediators and actors in their own right, because their embeddedness with known human agency shapes organizational and individual attention to risk, on a daily basis. With the XAI perspective, the balance between human and technological agency is shifted or displaced, it does not disappear, but takes new forms. For governance of adaptive capacity, actors must also be aware that the frontstage/backstage distinction is not only a matter of bias in data, influencing the model, but also that the XAI-derived "explanations" plays directly into the ongoing dialectic between frontstage and backstage, populated by humans in different sociotechnical and organizational contexts. Using Power's (2016) terms, an(y) AI-explanation/interpretation will in itself be an "artefact" or "representational device".

3. Conclusions and outlook

By presenting a sociotechnical framework for XAI and cyber resilience, we have suggested two processes by which XAI might relate to cyber resilience, using a theoretical conceptualization of cyber resilience

distinguishing between the three operationalizations of cyber resilience depicted as Theory A, Theory B and Theory C (Grøtan et al., 2022). We have first through examples from current literature and practitioners press illustrated how XAI might relate to cyber resilience in a technological (Theory A) and structural, organizational (Theory B) perspective. XAI-models opposed to black-box models could contribute to mitigate bias, enhance trust and improve decision-making, and as such be part of the repository of organized, supportive resources designed to uphold function when there is a loss of functionality due to a cyber-attack. As the way cyber resilience is framed in a Theory B perspective is limited to what one in risk management can anticipate and operationalize on existing knowledge, we have also discussed how adaptive capacity (Theory C) might relate to XAI through a discussion of how sensor data collected from critical infrastructure could be used to train an AI-model that through XAI-techniques could reflect real time constitution of adaptive capacity. However, as the further discussion based on the concepts of enactment and Riskwork illustrates, there is, from a theoretically perspective, limits to how much is gained without expanding the data capture beyond the directly observable merits of adaptation. We argue that a natural expansion is to encompass the conceptual twin of Power's (2016) notion of the ***backstage***, at which activities related to enactment of adaptive capacity in daily activities is more visible. Through the identification of some relevant proxies for adaptive capacity it ***might*** be possible that XAI will support a fruitful dialectic between, paraphrasing Power (2016), the frontstage and backstage activities constituting cyber resilience, encompassing adaptive capacity, not as a "set and forget" phenomenon, but a goal that must be pursued through daily work. In that respect, our work is an elaboration of Baukes (2017) argument on the difference between cyber security and cyber resilience but drawing on other organizational perspectives.

A way to bring it all together is by including a stakeholder perspective that deals with the different audiences data scientists will have to support in the work with XAI-models in the cyber domain. The ***operational*** cyber resilience audience will presumably be interested in successful adaptive patterns including adaptive traps, and successful interactions, including timing and interactional rhythms. Our guess is that this audience will be more inclined to value a backstage bias for the generative model of adaptive capacity. The ***organizational*** cyber resilience audience will presumably be more interested in the allocation of critical resources for keeping the organization poised to adapt, and the potential conflicts with objectives related to allocation of the same resources for more traditional, risk-informed contingency planning. Our guess is thus that they will be more inclined to value a front-stage bias. In line with the presumption that effective cyber resilience demands a composite of approaches (Grøtan et al., 2022), it is crucial that the organization is aware that XAI support for governing adaptive capacity plays directly into the ongoing dialectic between frontstage and backstage, and that an(y) AI-explanation/interpretation will in itself not be a substitute, but an "artefact" or "representational device" inserted into an ongoing process of cyber resilience governance. This will require an understanding of (limitations of) generative AI, and how the various "explanations" will be intertwined with the dynamism of the frontstage/backstage interplay.

Although our forecast "tempers" the most optimistic and instrumental expectations of XAI fueling adaptive capacity, we also acknowledge the significant potential for advance. For instance, Grøtan et al. (2022) argue that successful adaptive patterns emerging out of a Theory C perspective as a discovery of a new protocol or procedure, can be routinized and solidified as a Theory B construct for later use. Such a move will have to be thoroughly judged, but it is beyond doubt that properly implemented, XAI can enrich the process and improve such judgements and processes substantially. On the other hand, leaving such judgements to a non-mediated AI model would be dangerous. Future work on the topic of XAI and cyber resilience could take several paths. Our point of departure is that adaptive capacity (Theory C) should not be approached without being seen in conjunction with organizational cyber resilience (Theory B). Starting with the latter is therefore a viable choice if not addressed properly before. However, even when employing such a limited focus, it is not recommended to ignore the frontstage/backstage distinction (Power, 2016), which originates from the risk management domain, from which Theory B is intrinsically linked. However, when expanding the scope to include adaptive capacity (Theory C), it also becomes urgent to keep in mind that the artefacts and representational devices at play will be of an even more intangible and ephemeral nature, and that the target for our attention may be constantly changing. Outmanoeuvring the complexity organizations encounter cannot be done at distance, the adaptive steps taken will also impact the adaptive capacity. The most important takeaway is that research on XAI and cyber resilience should always encompass questions related to ensuring that an XAI model used for cyber resilience purposes actually is explainable for the user group (audience) it is addressing. The data analyst understanding what the machine understands at a sub-symbolic level is an intermediary and a translator towards an audience that will be accountable for their decisions in an organizational and sociotechnical context. This requires researchers to always look into XAI also from a non-technical design and human factors perspective.

Acknowledgements

This research is funded by the project Theoretical Advances of Cyber Resilience – Practice, Governance and Culture of Digitalization (TECNOCRACI), funded by the Research Council of Norway, grant no. 303489.

References

Adadi, A. & Berrada, M. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

Aziz Al Kabir, M., Elmedany, W. & Sharif, M. S. 2023. Securing IoT Devices Against Emerging Security Threats: Challenges and Mitigation Techniques. Journal of Cyber Security Technology, 7(4), 199-223.

Bansal, A. 2023. The Permanence Of AI In Cybersecurity: Why This Trend Is Here To Stay. https://www.forbes.com/sites/forbestechcouncil/2023/11/09/the-permanence-of-ai-in-cybersecurity-why-this-trend-is-here-to-stay/ (visited November 24, 2023).

Baukes. M. 2017. Cybersecurity Is Dead. https://www.forbes.com/sites/forbestechcouncil/2017/06/06/cybersecurity-is-dead/?sh=180418f04012 (visited December 19, 2023)

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 610-623.

Brasse, J., Broder, H. R., Förster, M., Klier, M. & Sigler, I. 2023. Explainable artificial intelligence in information systems: A review of the status quo and future research directions. Electronic Markets, 33(1), 26.

Brooks, C. 2023. A Primer On Artificial Intelligence And Cybersecurity. https://www.forbes.com/sites/chuckbrooks/2023/09/26/a-primer-on-artificial-intelligence-and-cybersecurity/ (visited November 24, 2023).

Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P. F., Han, Y., Jmila, H. & Zhang, Z. 2022. Explainable artificial intelligence for cybersecurity: a literature survey. Annals of Telecommunications, 77(11-12), 789-812.

Fier, J. 2022. The future of cyber security: 2022 predictions from Darktrace. https://darktrace.com/blog/the-future-of-cyber-security-2022-predictions-from-darktrace (visited November 24, 2023).

Fowler, M. 2023. Smart Policy And Sophisticated Technology: Augmenting The U.S. National Cyber Workforce. https://www.forbes.com/sites/forbestechcouncil/2023/09/20/smart-policy-and-sophisticated-technology-augmenting-the-us-national-cyber-workforce/ (visited November 24, 2023).

Greco, F., Desolda, G. & Esposito, A. 2023. Explaining phishing attacks: An XAI approach to enhance user awareness and trust. In Proc. of the Italian Conference on CyberSecurity (ITASEC '23).

Grøtan, T. O., Antonsen, S. & Haavik, T.K. 2022. Cyber resilience: a pre-understanding for an abductive research agenda. Matos, F, Paulo Maurício Selig, P. M., Henriqson, E. (Ed.) In Resilience in a Digital Age: Global Challenges in Organisations and Society, Springer International Publishing, Cham, 205-229.

Ho, S. K. 2020. 91% of all cyber attacks begin with a phishing email to an unexpected victim. https://www2.deloitte.com/my/en/pages/risk/articles/91-percent-of-all-cyber-attacks-begin-with-a-phishing-email-to-an-unexpected-victim.html (visited November 24, 2023).

Hollnagel, E. 2016. Resilience Engineering. https://erikhollnagel.com/ideas/resilience-engineering.html (visited November 24, 2023)

Hollnagel, E. 2016. The Functional Resonance Analysis Method. https://functionalresonance.com/basic-principles.html (visited December 19, 2023).

Hollnagel, E., Woods, D.D. & Levenson, N. 2006. Resilience Engineering. Concepts and Precepts. CRC Press

Jussupow, E., Spohrer, K., Heinzl, A. & Gawlitza, J. 2021. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Information Systems Research, 32(3), 713-735.

Kuppa, A. & Le-Khac, N. A. 2020. Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In 2020 International Joint Conference on neural networks (IJCNN) (pp. 1-8). IEEE.

Leavitt, H. J. 1964. Applied organization change in industry: structural, technical, and human approaches. Cooper, S., Leavitt, H., Shelly, K. (Ed). In New Perspectives in Organizational Research, Wiley, Chichester, 55-71

Lyytinen, K. & Newman, M. 2008. Explaining information systems change: a punctuated socio-technical change model. European Journal of Information Systems, 17, 589-613.

Mahbooba, B., Timilsina, M., Sahal, R. & Serrano, M. 2021. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity, 2021, 1-11.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. 2022. Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 1-66.

Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A. Y., & Tari, Z. 2023. Explainable Intrusion Detection for Cyber Defences in the Internet of Things: Opportunities and Solutions. IEEE Communications Surveys & Tutorials.

Muna, R. K., Hossain, M. I., Alam, M. G. R., Hassan, M. M., Ianni, M., & Fortino, G. 2023. Demystifying machine learning models of massive IoT attack detection with Explainable AI for sustainable and secure future smart cities. Internet of Things, 24, 100919.

Pettersen, S., & Grøtan, T. O. 2024. Exploring the grounds for cyber resilience in the hyper-connected oil and gas industry. Safety Science, 171, 106384.

Power, M. (Ed.). 2016. Riskwork: Essays on the organizational life of risk management. Oxford University Press.

Pradeesh, J. 2023. Artificial Intelligence In Cybersecurity: Unlocking Benefits And Confronting Challenges. https://www.forbes.com/sites/forbestechcouncil/2023/08/25/artificial-intelligence-in-cybersecurity-unlocking-benefits-and-confronting-challenges/ (visited November 24, 2023)

Rai, A. 2020. Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science, 48, 137-141.

Rath, R. C., Baral, S. K., & Goel, R. 2022. Role of Artificial Intelligence on Cybersecurity and Its Control. In Cross-Industry Applications of Cyber Security Frameworks (pp. 15-35). IGI Global.

Rjoub, G., Bentahar, J., Wahab, O. A., Mizouni, R., Song, A., Cohen, R., ... & Mourad, A. 2023. A Survey on Explainable Artificial Intelligence for Cybersecurity. IEEE Transactions on Network and Service Management.

U.S. Department of Energy. 2021. ADDSec: Artificial Diversity and Defense Security. https://www.energy.gov/sites/default/files/2021-11/ADDSec%20Artificial%20Diversity%20and%20Defense%20Security%20-%20SNL.pdf (visited November 24, 2023).
U.S. Department of Energy. 2021. Security Patch Automated Remediation Tool Analyzing the NVD (SPARTAN). https://www.energy.gov/sites/default/files/2021-08/Security%20Patch%20Automated%20Remediation%20Tool%20Analyzing%20the% 20NVD%20%28SPARTAN%29%20-%20SEEDS_508.pdf (visited November 24, 2023).
Ramya, P., Babu, S. V., & Venkatesan, G. 2023. Advancing Cybersecurity with Explainable Artificial Intelligence: A Review of the Latest Research. In 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1351-1357). IEEE
Rosness, R., Grøtan, T. O., Guttormsen, G., Herrera, I. A., Steiro, T., Størseth, F., Tinmannsvik, R. K., & Wærø, I. 2010. Organisational Accidents and Resilient Organisations: Six Perspectives. SINTEF-report https://www.sintef.no/globalassets/upload/teknologi_og_samfunn/sikkerhet-og-palitelighet/rapporter/sintef-a17034-organisational-accidents-and-resilience-organisations-six-perspectives.-revision-2.pdf (visited November 30, 2023)
Schemmer, M., Kühl, N., Benz, C., & Satzger, G. 2022. On the influence of explainable AI on automation bias. arXiv preprint arXiv:2204.08859.
Sivamohan, S., & Sridhar, S. S. 2023. An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework. Neural Computing and Applications, 35(15), 11459-11475.
Weick, K. E. 1988. Enacted sensemaking in crisis situations. Journal of Management Studies, 24(4), 305-317
Woods, D. D., Alderson, D. L., 2021. Progress toward resilient infrastructures: are we falling behind the pace of events and changing threats? J. Crit. Infrastruct. Pol. 2 (2), 5–18
Woods, D. 2018. The theory of graceful extensibility: Basic rules that govern adaptive systems. Enviroment Systems and Decisions, 38(5), 433-457.
Woods, D. D., & Branlat, M. 2011. Basic patterns in how adaptive systems fail. Hollnagel, E. (Ed.) In Resilience engineering in practice, CRC Press, 127-143.
Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. 2022. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access.
Zolanvari, M., Yang, Z., Khan, K., Jain, R., & Meskin, N. 2021. Trust xai: Model-agnostic explanations for ai with a case study on iiot security. IEEE internet of things journal.