*Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference*
*Edited by* Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P4503-cd

# Cyber resilience as an organizational outcome in the age of AI – explainability as a means to foster adaptive capacity

Dorthea Mathilde Kristin Vatn

*Software Engineering, Safety and Security, SINTEF Digital, Norway. E-mail: dorthea.vatn@sintef.no*

Tor Olav Grøtan

*Software Engineering, Safety and Security, SINTEF Digital, Norway. E-mail: tor.o.grotan@sintef.no*

Explainable AI (XAI) has received growing attention the last years due to the growth of more sophisticated machine learning models. These modern models are often complex, and their internal workings are often challenging to explain, fuelling the interest in terms like explainability, interpretability and transparency. Although XAI is considered to be a highly multidisciplinary topic, and the explainability dimension of the concept implicitly point at the human recipient of the explanations, both the human and organizational perspective seems to be currently neglected in XAI research. This is highly problematic within the cyber security domain where cyber resilience is dependent on sociotechnical dimensions relating to both human as well as organizational aspects.
While AI is suggested to have big impacts on the cyber security domain, it is still undiscovered what role explainability of AI will have. This paper builds on a sociotechnical understanding of the XAI concept and draws on the theoretical perspective of adaptive capacity as fundamental in the understanding of cyber resilience. We combine these perspectives when we approach current literature on XAI in the context of cyber resilience, illustrating that XAI is mainly treated as a technical artefact, neglecting the human and organizational dimensions that are crucial to develop and foster adaptive capacity. We discuss the implications of this narrow view on XAI in the context of cyber resilience, suggesting future research avenues to be followed to fill the current gaps.

*Keywords*: Cyber resilience, adaptive capacity, XAI, Explainable AI, Sociotechnical theory

## 1. Introduction

Explainable Artificial Intelligence (XAI) has received growing attention in the last years due to the growth of more sophisticated machine learning models (Arrieta et al., 2020). These modern models are often complex, and their internal workings are often challenging to explain. Although XAI is considered to be a highly multidisciplinary topic, and the explainability dimension of the concept implicitly points at the human recipient of the explanations, both the human and organizational perspective on XAI seems to be neglected in current research (Brasse et al., 2023; Pumplun et al. 2022). This is highly problematic for application of AI within the cyber domain where cyber resilience is dependent on sociotechnical dimensions relating to both human as well as organizational aspects (Grøtan et al., 2022).

Within critical infrastructure the integration of Information Technology (IT) with Operational Technology (OT) is increasingly contributing to new cyber-physical systems (Pettersen & Grøtan, 2024). While there are several postulated advantages of integrated cyber-physical systems relating to efficiency gains as well innovative modes of operations, this digital transformation also brings in new failure modes making these modern cyber-physical systems brittle and fragile in new ways (Grøtan et al., 2022). The digital transformation of critical infrastructure also provides massive access to data. In combination with the increasing computational power that is available, adoption and use of AI is becoming increasingly relevant also within these cyber-physical systems, bringing in both new opportunities as well as challenges.

AI tools and techniques are being postulated to be important for several cyber security purposes (Moustafa et al., 2023). However, the nature of cyber security threats will seldom be only a technical matter (Vatn & Grøtan, 2024). This underscores the importance of considering sociotechnical issues relating to both the human actors present in the cyber-

physical systems as well as organizational matters. This is also implicitly visible when Mahboubi et al. (2024) point at three critical factors making cyber threats dynamic by nature. Firstly, malicious attackers will adapt and develop techniques, tactics and procedures, implicitly pointing at *human intentionality* as an important aspect with cyber threats. It is also reason to believe that who the attackers are, and their motives will change over time, contributing to further complexity. Secondly, modern IT environments are increasingly complex with cloud services, Internet of Things (IoT) devices, as well as remote access. Overall, this constitutes a significant attack surface that might be approached from several different angles. Thirdly, the combination of the increasingly complex nature of the data generated by the cyber physical systems and the fact that many organizations lack skilled expertise to analyze the data generated by security tools, makes it easy to overlook security threats (Mahboubi et al., 2024). In such a complex environment, organizations should expect to be surprised by unforeseen cyberattacks (Grøtan et al., 2022). These points elucidate that although AI tools are increasingly being developed for the cyber domain, the cyber resilience of a cyber physical system is not just a technical matter.

Our starting point is that cyber-physical systems ultimately should be regarded as complex sociotechnical systems, involving dynamics between technical aspects, human aspects, as well as organizational aspects. In light of this, the explainability dimension of AI systems developed for this domain might be regarded as a paramount "ingredient" for the overall cyber resilience developed for a cyber-physical system. This points at the usefulness of exploring how XAI is related to cyber resilience in current research literature and explore both if and how current literature address sociotechnical dimensions. The research questions serving as basis for this paper are: 1. *How is XAI related to cyber resilience in current literature?* 2. *How can a theoretically informed approach to sociotechnical cyber resilience guide the understanding of what XAI should encompass in the context of managing cyber risks towards critical cyber physical systems?*

To answer these questions, we first outline the theoretical background that has been guiding our work, building on the perspective of adaptive capacity (Woods, 2018) as fundamental in the understanding of cyber resilience (Grøtan et al., 2022). Next, we define XAI as a sociotechnical concept encompassing technical, human, as well as organizational dimensions, illustrating that addressing the explainability of AI in the context of cyber physical systems requires a careful understanding of the different stakeholders in the landscape surrounding it. Next, we present how current research literature approaches XAI and cyber resilience, clearly elucidating that there is a sociotechnical deficit in the current approaches to XAI, and thus on its contribution to sociotechnical cyber resilience. Lastly, building on the findings from our literature review, we propose a conceptual framework for utilizing XAI to provide both the transparency and accountability potentials that are needed to approach cyber resilience incorporating adaptive capacity towards surprise and disruption. These potentials will reside both in modelling approaches and in interactive processes adapted to user contexts.

## 2. Cyber resilience and adaptive capacity

Due to, inter alia, the geopolitical situation, hybrid threats and an increasing number of attack surfaces, traditional cyber security approaches struggle to keep pace with the cyber risks evolving from increasingly complex threat and vulnerability landscapes. Woods & Alderson (2021) coins this lagging pace of adaptation a Strategic Agility Gap (SAG) for critical infrastructures at large, and the cyber risk is no exception to this. Quite the contrary, surprises and shocks must be expected, and herein emerges the promise of "resilience".

The term "cyber resilience" is however polysemic and highly ambiguous, accommodating a variety of legacies, presumptions and pragmatic interests. Dupont et al. (2023) describes this situation as comprising four sensemaking tensions and five practices trying to resolve the tensions. In particular, they point to the challenge of combining prescriptive "playbooks" with adaptation to surprising events. Grøtan et al., (2022) address the same challenge in resilience terms through a simplified "Resilience ABC" taxonomy, comprising three different explanations or theories on the origin of resilient outcomes in the management of cyber risks. The theories

capture a variety in both the presumed outcome, and its origin. Put simply, Theory A attributes resilient outcomes of *absorption of* or *rebound from* anticipated disturbance to technical capabilities such as redundance, fault-tolerance and feedback control. Theory B attributes similar outcomes to existing practices of, e.g., risk management, emergency preparedness and business continuity planning with an emphasis on proactive adaptation and post-event learning. Theory C departs from the other two by focusing on sustained adaptive capacity at boundaries of preparedness as a distinct phenomenon and practice, grounded in complexity theory and thus recognizing the possibility of fundamental surprise and the possibility that single-minded robustification (Theory B) may imply an inherent fragility (Woods, 2018, Carlson and Doyle, 2000, p. 2529), and nurtured by a "silent/tacit" performance variability that remains unacknowledged by Theory A and B. Hence, Theory A is purely technical, while Theories B and C are both sociotechnical, but highly different regarding recognition of complexity and its implications, and on the value of performance variability (Hollnagel, 2002). Correspondingly, Theory C is needed to close the Strategic Agility Gap (Woods and Alderson, 2021). Pettersen and Grøtan (2024) addresses how this taxonomy can be used in the oil and gas industry context, and claims that Theory B and C must be combined, a position implying the dialectical challenge or implementing resilience in a context founded on an opposite organizational logic (Grøtan, 2015).

## 3. XAI as a sociotechnical concept

XAI is a complex and diverse research topic (Weber et al., 2024). Departing from the sociotechnical HTO-standpoint (Karltun et al., 2017), Vatn & Mikalef (2024) argues that XAI might be regarded a multidisciplinary, layered, and dynamic concept encompassing technical, human, as well as organizational dimensions. This approach acknowledges the dualistic view on current XAI research presented by Weber et al. (2024) as either narrowly focusing on developing techniques and methods extracting information about the inner workings of modern machine learning models or as focusing on all kinds of AI models and how these provide explanations about their decisions to different types of users. While the first view is mainly focusing on developing techniques and methods providing explanations that technical experts can benefit from, the latter view explicitly focuses on the human as the user receiving the explanations from AI. As such, "*XAI targets all types of users thus entailing social and psychological aspects*" (Weber et al., 2024, p. 302).

In our approach to XAI, we however depart from an understanding of it as a sociotechnical concept, and that addressing explainability of AI systems requires a careful stakeholder focus. This is an understanding in line with highly cited definitions of XAI. For instance, Arrieta et al. (2020) state that "*Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand*" (p. 85). This definition implicitly points at what is explainable to one audience, might not be that for another audience. By connecting XAI to responsible implementation of AI methods in practical business applications, also the organizational perspective is made evident (Arrieta et al., 2020).

Addressing XAI in the context of cyber resilience would first of all require knowledge of the stakeholders surrounding a cyber physical system using AI. Second, "XAI" may imply very different pragmatic foci and "meaning" for the different stakeholders. Figure 1 illustrates a stratified stakeholder perspective, from which different "XAI" interests may be derived. Here, each strata signifies a community of practice that will share a pragmatic interest in use of XAI, encapsulated in specific ways of asking "what", "why", "when" and so on. By using the sociological notion of "stratification", we also hint beyond the mere shared (communicative) pragmatic interest, to inherent power relations. In which, it is not obvious that power is exerted only from the regulator to the liable user, to the technologists. Quite the contrary, the AI specialist may possess an unprecedented "model power" (Bråten, 1973) if XAI is not applied in a careful manner. For instance, at the Core AI level(s), data scientists would be interested in understanding how the model explains its decisions in terms of internal technical aspects, such as which weight the model give to different features, how/when the model hallucinates, how the model can suddenly break down, how vulnerable it will be towards data tampering,

manipulation and data poisoning. This would however be only partially interesting for the system engineer that builds AI into their existing or new AI-enabled systems. Both the system engineer/designer and the presumed users would be more concerned about the system's safety in use; that it is reliable related to a presumed use and application and does not lead the users astray or betray them. Moreover, an organization or company that is liable for operation of a critical infrastructure, would need assurances that the key operational and managerial processes, e.g., risk management to avoid major accidents in high-risk industries, are not disturbed. If not, they may experience objections from a regulator, which in turn must address the relation between XAI and the critical processes from a governance/policy perspective, which demands other, more high-level explainability capabilities.

The above exemplifications are only tentative and will require extensive elaboration in different regulatory contexts. However, the overall, fundamental challenge is a rapidly developing AI technology that runs an unlimited appetite for solving "everything", without regard for existing knowledge, competence and practice in organizational contexts. This is a general challenge also at the societal level. Maybe most urgently, a high-risk industry running "proven" risk management processes cannot allow itself to experiment with AI-support, without running a thorough risk management on that issue itself, now matter how much the AI technologists advocate their model advances that make sense at their level. In principle, the regulator would not allow that to happen. However, given the severeness of the emerging cyber risk and the strategic agility gap associated with current cyber security approaches, both liable companies, regulators – and society – may be tempted to "buy" the unlimited promises from AI as technology (hype). We argue that this should not happen, and XAI should be developed through both AI model features, and continuously interactive sociotechnical processes of building and maintaining trust.

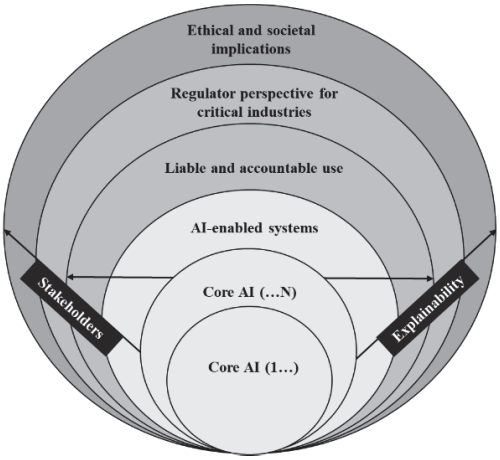


Fig. 1 A stratified stakeholder perspective.

## 4. Literature review

To identify literature, the search string "((XAI OR "explainable AI" OR "Explainable Artificial Intelligence") AND "cyber resilience")" was applied in Scopus, Web of Science, and IEEEXplore the 1st of October 2024. Duplicates were removed and 138 hits were screened on title and abstract level. Of these, 37 addressed XAI and cyber in an organizational context and were included for further full text analysis in EndNote. In the full text analysis, we chose to include only journal articles that explicitly addressed both XAI and cyber resilience, leaving us with a final paper pool of 10 papers.

In the analysis of the papers, we applied a thematic content analysis (TCA) presented by Anderson (2007) to explore and identify themes relating to XAI and cyber resilience across the papers. TCA is a stepwise qualitative method that might be summarized in four steps encompassing (1) reading the papers, (2) provide a code to aspects relevant for the research question, (3) categorize the codes that are similar, and (4) provide a coherent label to the different categories of codes. We developed a concept matrix in Excel to help out in this stage, enabling us to extract excerpts from the papers that addressed XAI and cyber resilience, respectively. In the write up we added a fifth step aimed at bringing the different themes to a level where we could discuss how they coherently provide an answer to the research question serving as basis for our work, namely how XAI is related to cyber resilience.

## 4. XAI in the context of cyber resilience – current conceptualizations

The thematic content analysis was performed with the two questions "how is XAI conceptualized" and "how is cyber resilience addressed" in mind. This gave rise to four themes that are presented in the following.

### 4.1. *XAI as a technical solution that can unleash the potential of complex AI models*

Most of the papers address XAI as a technical solution, encompassing a collection of techniques and methods (Zkik et al., 2024; Sadeghi et al., 2024; Dari et al., 2023; Sarker et al., 2024). The usefulness of these techniques and methods are argued to stem from their potential to unleash the potential of complex AI models to be used for cyber purposes. For instance, Zkik et al. (2024) develops a resilience strategy based on explainable deep learning technologies and a blockchain-based strategy. In this context, XAI is understood as the overarching collection of techniques that can facilitate the use of deep learning models, allowing users to understand how the model arrived at its output and why certain decisions were made. Also, Sarker et al. (2024) underscore the important role XAI methods might have in solving complex cybersecurity issues facilitating that AI methods can be used in a way that ensures human-understandable explanations and trustworthiness. According to Masud et al. (2024), XAI might ensure that AI systems have high learning performance understood as prediction accuracy, while at the same time provide explainable models.

### 4.2. *XAI as means to facilitate trust*

While most of the papers have a technical entrance to XAI, some of the papers also have an entrance to XAI as a means to facilitate trust (Chinu & Bansal, 2024) by ensuring accommodation of ethical, legal and regulatory concerns that apply to nonexplainable AI (Hoenig et al., 2024; Masud et al., 2024; Halgamuge et al., 2024; Jeffrey et al., 2023). For instance, XAI is seen as the means that might ensure that the "*vital needs for control, fairness, accountability, safety, cyber-resilience, and cybersecurity*" (Hoenig et al., 2024, p. 73113) are met. The perspective on XAI is also extended beyond merely a technical one, as the importance of "translating" the outputs of XAI methods to meaningful visualizations and customized explanations for different user groups is underscored (Hoenig et al., 2024). Also, Masud et al. (2024) and Chinu and Bansal (2024) have as a starting point that lack of interpretability and transparency from AI systems might challenge human users' trust. XAI is therefore seen as the means that might ensure that AI systems might be implemented and used while at the same time taking a responsibility dimension into account (Masud et al., 2024). By ensuring that human stakeholders are able to comprehend, trust, and manage AI systems, XAI seeks to be able to provide an answer to the question of who is responsible if things go wrong. This could be seen in connection with the use of AI/ML in the context of industrial IoT networks in industrial domains where operations might have significant economic or life safety considerations (Jeffrey et al., 2023). In contexts where there is an assumption that humans need to make interventions to ensure successful operations, it is paramount that these humans understand the predictive decisions made by AI/ML systems (Jeffrey et al., 2023).

### 4.3. *Cyber resilience as a technical matter*

By looking into how the papers address cyber resilience, we see that both cyber resilience in a Theory A and a Theory B perspective is covered. From a Theory A perspective, cyber resilience will encompass using technology as a means to achieve cyber resilience. Several of the papers bring in AI as a possible technological means to achieve cyber resilience (Hoenig et al., 2024; Dari et al., 2023; Alevizos & Dekker, 2024; Masud et al., 2024). For instance, Hoenig et al. (2024) underscores how cyber resilience should be regarded as an intrinsic part of a cyber physical system, ensuring that the system might continue to work during cyberattacks. In this context, XAI might be used as technical means to detect and activate methods allowing the system to continue functioning during the attack, contributing to cyber resilience in a Theory A perspective. Also, the underscoring of AI's ability to perform continuous monitoring and improvement (Alevizos & Dekker, 2024), and how the integration of sophisticated AI architectures such as neural networks is a

potential strategy to improve cyber resilience (Dari et al., 2023) are examples of cyber resilience as a technical matter.

### 4.4. *Cyber resilience as something "more"*

From a Theory B perspective, several of the papers underscore how the explainability dimension contributes to resilience implicitly by making it possible to identify potential vulnerabilities and threats that could be missed, and as such "*help analysts develop more effective security strategies and prioritize their resources more efficiently*" (Sarker et al., 2024, p. 954). Also, Zkik et al. (2024) make a note on how explainability might facilitate human-AI collaboration, contributing to more informed decision-making.

Some of the papers combine a cyber resilience approach encompassing several of the perspectives in Grøtan et al. (2022) approach to cyber resilience. Masud et al. (2024) make a note on how cyber resilience is a broader notion that the term cyber security, and that cyber resilience can help provide answers to three underlying cyber security questions. The first question relates to how large-scale security flaws in IoT devices can be avoided, that in our theoretical understanding of cyber resilience would point at a Theory A perspective. The second question relates to how the functioning of a network can be kept at a tolerable level of performance while absorbing the impact of an attack, that through our theoretical starting point could be relating to a Theory A and Theory B perspective combined. The third question relates to which resources and assets must be available to facilitate that a system can adapt to cyber threats. This question could both be seen in the Theory B perspective if one narrowly sees the assets and resources as plans and structures that are organized beforehand to tackle cyberattacks. However, if one thinks that these assets and resources also could encompass the ability of an organization to perform gracefully extensibility (Woods, 2018) beyond given rules and plans, we are touching upon cyber resilience in a Theory C perspective. However, the notion of gracefully extensibility is not explicitly or theoretically addressed in the papers, and therefore it remains largely speculative what is meant by adaptability in the papers. The most reasonable interpretation is a dynamic or agile planning capability, which is essentially Theory B.

A few papers make a note on the importance of flexibility and agility in the context of cyber resilience as attackers will always find new tactics and approaches and make changes in the cyberattack strategies used due to technological advancements (Dari et al., 2023; Sadeghi et al., 2024). For instance, Dari et al. (2023) underscore the importance of having a comprehensive and flexible cyber security plan as attackers at all times will make changes in their approaches and tactics. In a similar vein, Sadeghi et al. (2024) bring in the importance of agility when meeting cyber-attacks in an everchanging threat landscape. This agility is underscored to be vital in decision-making to ensure that new and changed information and knowledge can be taken into account. To foster agile decision-making, clear and understandable information is necessary and seen as a prerequisite, explicitly pointing at AI transparency as important. Although both flexibility and agility are terms that might point at cyber resilience in a Theory C perspective, it is visible that cyber resilience seems to be mainly approached as something that heavily relies on the ability of decision support systems to provide access to predefined information and business continuity plans. A clear distinction between cyber resilience in a Theory B and Theory C perspective is that the latter focus on a capability that extends beyond what can be foreseen and written down in plans projected from current practices.

## 5. Discussion

Based on the analysis of the current literature we suggest a theoretical model illustrating that several important questions related to XAI, and cyber resilience, is currently neglected and can only be approached by departing from a sociotechnical perspective. Using the simplified "Resilience ABC" model, a major aspect is that the junction between Theory B and Theory C signifies a major divide, with huge implications for the potential utility of XAI. Theory B represents the most prevalent conceptualization of resilience related to the recognized need to ensure controlled recovery and rebound from incidents that elude risk management efforts. In which, resilience as a practice is commensurate

with (and therefore merely an epiphenomenon of) practices of risk management, emergency planning and business continuity planning. This is accompanied by the (hopeful/pious) presumption that such preparedness aimed at anticipated events projected from a current understanding of how the system works, also will be useful in the event of fundamental surprise (or "black swans").

Theory C however represents a radically different approach. It is rooted in complexity theory in which fundamental surprise is inherent, common basic assumptions of order, rational choice and intent in organizational decision-support and strategy are challenged (e.g., Kurtz and Snowden, 2003). It does not aim for "solutions" in the classical engineering sense, but for an adaptive capacity that enables organizations to constantly revise their models, anticipations and response repertoire. This way, they may avoid becoming stuck and stale when encountering complexity and deviations from the "normal". By embracing change and being poised to adapt through pursuing resilience as a distinct practice in which practice is constantly revised, organizations will avoid becoming "robust, yet fragile", and instead enable what Woods (2018) denote "graceful extensibility. A main point for our purpose here, is that Theory C carries an inherent presumption of continued **novelty** of resilient performance, and that repetition and thus commensurability across organizations will be lower than what can be presumed through Theory A and B.

This divide has major implications for what we may expect from XAI. We should expect that AI is suited for mapping - and detection of deviation from - repetitive patterns of both attack and defence. This indicates that XAI has a potential for supporting Theory A and Theory B, although with different pragmatic needs across different strata (Fig. 1). On the other hand, there are clear limits related to the ability of AI to support a Theory C perspective, in which processes of human imagination and sensemaking are presumed to produce novel patterns of practice, beyond mere projections of current or past practice. However, if AI models are trained to recognize dysfunctional or successful adaptative patterns, a conceptually potential solution is to use AI to support a process of introspection of the adaptive history, to imagine a precarious present, and to envisage a resilient future. However, also in this case, "XAI" need to be adapted to the different strata. The "lower" the strata, more can be expected from interpreting AI model performance. The "higher" the strata, more emphasis must be put on sociotechnical interactivity and trust. The exact boundaries and balances between these approaches at different strata requires much more empirical support. As XAI currently is an immature field (Weber et al. 2024), it will also require thorough theoretical attention and scientific rigour beyond the state of the art.

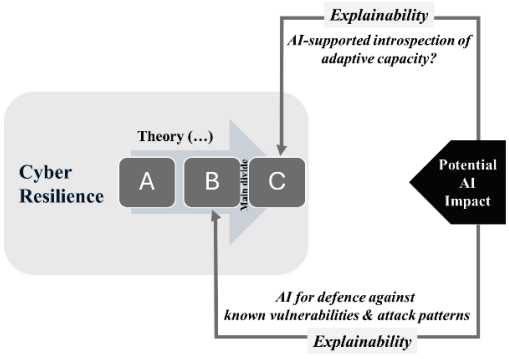


Fig. 2. XAI and cyber resilience.

**Acknowledgement**

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**

Alevizos, L., & Dekker, M. (2024). Towards an AI-Enhanced Cyber Threat Intelligence Processing Pipeline. *Electronics (Switzerland), 13*(11).

Anderson, R. M. (2007). *Thematic Content Analysis (TCA)*. Retrieved from: https://rosemarieanderson.com/wp-content/uploads/2014/08/ThematicContentAnalysis.pdf (16.12.2024)

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82-115.

Carlson, J. M., & Doyle, J. (2000). Highly optimized tolerance: Robustness and design in complex systems. *Physical review letters*, *84*(11), 2529.

Chinu, & Bansal, U. (2024). Fair XIDS: Ensuring fairness and transparency in intrusion detection models. *Concurrency and Computation: Practice and Experience*, *36*(25), e8268.

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.

Braten, S. (1973). Model monopoly and communication: Systems theoretical notes on democratization. *Acta Sociologica*, *16*(2), 98-107.

Dari, S. S., Thool, K. U., Deshpande, Y. D., Aush, M. G., Patil, V. D., & Bendale, S. P. (2023). Neural Networks and Cyber Resilience: Deep Insights into AI Architectures for Robust Security Framework. *Journal of Electrical Systems, 19*(3), 78-95. https://doi.org/doi:10.52783/jes.653

Dupont, B., Shearing, C., Bernier, M., & Leukfeldt, R. (2023). The tensions of cyber-resilience: From sensemaking to practice. *Computers & security*, *132*, 103372.

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.

Halgamuge, M. N. (2024). Leveraging Deep Learning to Strengthen the Cyber-Resilience of Renewable Energy Supply Chains: A Survey. *IEEE Communications Surveys and Tutorials, 26*(3), 2146-2175.

Hoenig, A., Roy, K., Acquaah, Y. T., Yi, S., & Desai, S. S. (2024). Explainable AI for Cyber-Physical Systems: Issues and Challenges. *IEEE ACCESS, 12*, 73113-73140.

Hollnagel, E. (2002, September). Understanding accidents-from root causes to performance variability. In *Proceedings of the IEEE 7th conference on human factors and power plants* (pp. 1-1). IEEE.

Jeffrey, N., Tan, Q., & Villar, J. R. (2023). A Review of Anomaly Detection Strategies to Detect Threats to Cyber-Physical Systems. *Electronics (Switzerland), 12*(15).

Karltun, A., Karltun, J., Berglund, M., & Eklund, J. (2017). HTO–A complementary ergonomics approach. *Applied ergonomics, 59*, 182-190.

Mahboubi, A., Luong, K., Aboutorab, H., Bui, H. T., Jarrad, G., Bahutair, M., Camtepe, S., Pogrebna, G., Ahmed, E., Barry, B., & Gately, H. (2024). Evolving techniques in cyber threat hunting: A systematic review. *Journal of Network and Computer Applications*, 232.

Masud, M. T., Keshk, M., Moustafa, N., Linkov, I., & Emge, D. K. (2024). Explainable Artificial Intelligence for Resilient Security Applications in the Internet of Things. *IEEE Open Journal of the Communications Society*, 1-1.

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 25* (3), 1775 – 1807.

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

Pumplun, L., Peters, F., Gawlitza, J. F., & Buxmann, P. (2023). Bringing machine learning systems into clinical practice: a design science approach to explainable machine learning-based clinical decision support systems. *Journal of the Association for Information Systems, 24*(4), 953-979.

Sadeghi R., K., Ojha, D., Kaur, P., Mahto, R. V., & Dhir, A. (2024). Explainable artificial intelligence and agile decision-making in supply chain cyber resilience. *Decision Support Systems, 180*.

Sarker, I. H., Janicke, H., Mohsin, A., Gill, A., & Maglaras, L. (2024). Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. *ICT Express, 10*(4), 935-958.

Vatn, D. M. K., & Mikalef, P. (2024). Theorizing XAI-a layered concept being multidisciplinary at its core. *In: AMCIS 2024 Proceedings*, Salt Lake City.

Vatn, D. M. K. V. & Grøtan, T. O. (2024). XAI and cyber resilience – a sociotechnical perspective. *ESREL 2024, Krakow*.

Weber, R. O., Johs, A. J., Goel, P., & Silva, J. M. (2024). XAI is in trouble. AI Magazine, 45(3), 300-316.

Woods, D. (2018). The theory of graceful extensibility: Basic rules that govern adaptive systems. *Environment Systems and Decisions, 38*(5), 433-457.

Woods, D. D., Alderson, D. L., (2021). Progress toward resilient infrastructures: are we falling behind the pace of events and changing threats? *Journal of Critical Infrastructure Polic*y 2 (2), 5–18.

Zkik, K., Belhadi, A., Kamble, S., Venkatesh, M., Oudani, M., & Sebbar, A. (2024). Cyber resilience framework for online retail using explainable deep learning approaches and blockchain-based consensus protocol. *Decision Support Systems, 182*.