Proceedings of the DataNinja sAIOnARA 2024 Conference 1-9 DOI: 10.11576/dataninja-1156

Shaping Trustworthy AI: An Introduction to This Issue

Ulrike Kuhl UKUHL@TECHFAK.UNI-BIELEFELD.DE
Machine Learning Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany

1. Introduction

The rapid advancement of artificial intelligence (AI) has brought about a paradigm shift in various domains, from healthcare to finance, and from autonomous systems to natural language processing. As AI systems become increasingly integrated into our daily lives, ensuring their trustworthiness is paramount. The DataNinja sAIOnARA 2024 Conference, centered around the theme "Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches", brought together cutting-edge research aimed at addressing the multifaceted challenges of creating reliable and ethical AI systems. This collection of scientific abstracts represents a broad spectrum of innovative work that contributes to the overarching goal of trustworthy AI.

The contributions are grouped under three prevailing hot topics in XAI: fairness and ethics, interpretability and transparency, and reliability and robustness, highlighting the multifaceted approaches to developing AI systems that are both innovative and trustworthy.

2. Fairness and Ethics

A central theme related to trustworthy AI is in how far systems can be considered fair and ethical. As automated decision-making increasingly impacts individuals and communities directly, concerns about bias, equity, and transparency become critical.

Balestra [1] delves into the fundamental question of fairness in algorithmic rankings. Rankings are a ubiquitous feature in modern life, from search engines to personalized recommendations. They draw attention to the fact that-while fairness may not seem essential when ranking depersonalized items-it becomes deeply relevant when individuals are being ranked. In such cases, disparities in how people are represented or treated can have significant consequences. Highlighting the often conflicting relationships between group fairness, individual fairness, and diversity in rankings, Balestra [1] draws attention to the inherent trade-offs and complexities involved in attempting to optimize all these aspects simultaneously. Moreover, their exploration combines Shapley values [2], known for promoting individual fairness, with a diversity measure to ensure group fairness in rankings. This approach introduces a framework that balances individual contributions with the need for diverse representation, offering a potential pathway for future research to operationalize fairness in rankings.

The work presented in Hellwig and Maier [3] extends the conversation on fairness into the domain of workplace leadership, where AI may increasingly take over functions like rewarding employee performance or allocating resources. Drawing on the "Resource Theory of Social Exchange" [4], they present a study plan to explore whether certain resources, such as affiliation or emotional support, hold the same value when allocated by AI compared to human leaders. This raises profound ethical questions about the "humanness" of resource allocation and whether AI can truly fulfill the nuanced role of a leader in fostering relationships and providing more than just material rewards. This perspective introduces an important dimension to the discussion of trustworthy AI: while AI systems may be technically proficient at optimizing resources, their ability to consider the social and emotional impacts of their decisions remains questionable.

Shifting the focus from fairness in decision-making processes to the social dynamics of human-AI interactions, Arlinghaus and Maier [5] outline a research framework to explore these interactions in workplaces where humans and robots collaborate. They seek to explore how individuals experience social exclusion when working with robots compared to exclusion by human colleagues, highlighting a unique challenge in the development of AI systems: the human perception of social interaction with machines. The research will focus on addressing the psychological needs of belonging, control, and self-esteem, examining how these needs are impacted differently depending on whether the source of exclusion is a human or a robot. By questioning the validity of the "Computers Are Social Actors" theory [6], their work promises to open the door to a deeper understanding of how humans attribute social qualities to AI and robots and how these attributions influence the effectiveness and fairness of human-robot collaborations.

Taking a more technological perspective to the fairness and ethics discussion, Sanaullah et al. [7] focus on the issue of data privacy. They tackle the challenge of preserving privacy in machine learning models without sacrificing performance in the context of different encryption techniques, highlighting the trade-offs between model accuracy, memory usage, training time, and security. By demonstrating how different encryption methods impact both the robustness of the models and their interpretability, this work addresses a core concern in the development of trustworthy AI systems. Their findings provide crucial insights into how we can develop ML models that are both effective and respectful of individual privacy, ensuring that fairness is maintained even when data is securely encrypted.

In the domain of process segmentation in operational systems, the work presented by Norouzifar and van der Aalst [8] highlights six potential research questions related to leveraging information not just from desirable, but also from undesirable events in process mining tasks. This research has significant implications for fairness in operational decision-making, as the method provides a more nuanced view of process data to help organizations ensure that cases are treated equitably based on their complexity and risk. Thus, the work presented by Norouzifar and van der Aalst [8] highlights the pathway to fairer outcomes in process management that will eventually support organizations in delivering more balanced and transparent decisions.

The contributions in this section collectively highlight the complexity and multifaceted nature of fairness and ethics in AI. Taken together, they illustrate the breadth of ethical challenges we face as AI systems take on increasingly important roles in society: from ensuring fair representation in rankings to maintaining ethical considerations in resource allocation and leadership; from fostering inclusive social dynamics in AI-human collaboration to protecting privacy while maintaining performance. They call for ongoing research, thoughtful design, and ethical foresight to ensure that AI systems not only perform their tasks efficiently but also align with broader human values of fairness, equity, and trust.

3. Interpretability and Transparency

While the relationship between interpretability and trust is more nuanced than often assumed, a prevailing notion remains that systems that are more interpretable and transparent are generally easier to trust [9]. The contributions in this section explore various approaches to making AI systems more interpretable and transparent, ensuring that these systems can be trusted not just for their performance but also for their ability to provide insight into their inner workings. The research presented here investigates how feature importance, logical constraints, and inherently interpretable models may contribute to this goal.

The work by Kolpaczki [10] addresses one of the most pressing issues in interpretability: understanding which features contribute most to a model's prediction. In many machine learning models, particularly those involving high-dimensional data, it is critical to assign importance scores to features to understand the model's behavior. These importance scores are not only crucial for interpretability but also serve practical purposes, such as feature selection, which can reduce the complexity of data and models. Kolpaczki [10] addresses the challenge of computational complexity and performance. Via empirical evaluation, they demonstrate an advantage of stratification methods, which may be attributed to the influence of feature subset size on overall correlation. The significance of this work lies in its practical application: by improving how we quantify feature importance efficiently, Kolpaczki [10] helps to make machine learning models more transparent.

A key challenge to interpretability is the "black-box" nature of deep neural networks. Despite their impressive performance across numerous domains, the complexity of neural networks often renders them opaque, making it difficult to interpret or trust their predictions [11]. This opacity poses a barrier when deploying these models in high-stakes environments such as healthcare, finance, or autonomous systems, where understanding the rationale behind a decision is just as important as the decision itself. Lutz and Neider [12] propose a framework that uses Linear Temporal Logic to create inherently interpretable machine learning models. By its very nature, Linear Temporal Logic provides a transparent and structured way of expressing temporal relationships and dependencies within data. Corresponding formulas are intuitive for human experts and can be easily validated against known rules or domain knowledge, ensuring that the models appear trustworthy in practical settings. At the same time, the approach presented by Lutz and Neider [12] offers remarkable flexibility, as it can be applied in both supervised and unsupervised learning settings, and it can be adapted to handle noisy data and incorporate expert knowledge. Thus, their work demonstrates how inherently interpretable models can bridge the gap between high-performing AI and trustworthiness.

In a similar vein, the research plan proposed by Katzke et al. [13] explores how deep learning models can benefit from the incorporation of logical constraints. They outline a strategy for extending existing frameworks to integrate constraints grounded in foundational prior knowledge. Ultimately, this approach may enable the model to operate with formal guarantees regarding its behavior. Thus, by embedding logical rules within the model, their approach aims to not only enhance model performance but also ensure that essential properties are preserved throughout the inference process. The significance of this work lies in its ability to provide formal guarantees, which is an essential feature in applications where the reliability and trustworthiness of AI systems are critical.

While logical constraints provide a structured way to enhance the reliability of deep learning models, another promising approach to interpretability lies in concept extraction methods. Traditionally applied to image models, Holzapfel et al. [14][1] extend these techniques to time series models and demonstrate the practical insights these methods offer. By analyzing the extracted concepts, their work reveals which features the model relies on for its predictions, as well as those contributing to errors. This level of analysis provides users with valuable insights into the biases within the model or dataset. The preliminary results presented by Holzapfel et al. [14] confirm that the adapted algorithm can successfully identify meaningful features in time series data, making these concepts critical for enhancing the interpretability of the model. Thus, their work promises to directly advance trustworthy AI by allowing users to identify and address potential biases or misleading features in critical fields like healthcare, finance, or any domain that relies on complex time series data.

Collectively, the presented contributions highlight the growing importance of interpretability and transparency in AI systems, particularly as these systems become more complex and integral to decision-making in high-stakes environments. They underscore the necessity of building AI systems that can explain their decisions, adhere to logical principles, and be scrutinized by human users. Interpretability and transparency are not just desirable features—they are essential for building trust in AI systems, especially as AI continues to evolve and take on more significant roles in society.

4. Reliability and Robustness

It is essential for a system to maintain consistent performance across varying conditions. AI systems specifically must be resilient enough to manage unexpected inputs, noisy data, and novel scenarios without failure. The contributions in this section are dedicated to designing AI models and systems that not only deliver high performance reliably, but are also robust enough to handle the complexities and uncertainties of real-world environments. The presented research approaches reliability and robustness from different perspectives, including active learning, physical sensor integration, decentralized robotic control, and bias correction in neural networks.

Zelba et al. [15] introduce a system designed to monitor technical processes and detect anomalies with minimal training data. Their COMETH system leverages active learning, a technique where the system efficiently queries the most informative data points to improve model performance, significantly reducing the amount of data needed for training. This approach is particularly important in real-world industrial applications, such as heating, ventilation, or air conditioning systems, and industrial machinery, where collecting large volumes of labeled training data is often impractical. A key strength of COMETH lies in its capacity to integrate feedback into its learning process, thereby enhancing the reliability and robustness of anomaly detection. Moreover, Zelba et al. [15] introduce an intriguing extension by integrating large language models (LLMs), adding another layer of robustness by incorporating context-aware insights that allow the system to provide more specific and actionable recommendations to users. This fusion of machine learning techniques

[1] Holzapfel et al. [14] were awarded with the 1st Place Best Poster Award at the DataNinja sAIOnARA 2024 Conference, Bielefeld.

underscores the importance of adaptability and feedback in creating AI systems that are both reliable and trustworthy.

While anomaly detection as presented by Zelba et al. [15] tackles issues of system reliability in real-time monitoring, Krebs et al. [16] shift the focus to long-term process improvement in battery manufacturing, using virtual measurements to balance efficiency and quality assurance. Virtual measurements reduce testing efforts while maintaining high standards of quality. In the presented contribution, Krebs et al. [16] identify the main requirements to facilitate trustworthy virtual measurements even for complex process chains found in battery manufacturing, with the goal of making virtual data measurements as robust as physical tests.

Shifting from industrial applications to healthcare, Grimmelsmann et al. [17] and Vieth [18] both focus on improving predictive accuracy and sensor-based systems for enhancing human physical abilities. Grimmelsmann et al. [17] expand the use of AI in biomechanics through the development of exoskeletons and the prediction of limb movements. Relying on surface electromyography signals, their study trains virtual sensors to predict the movement of deep muscles, providing an intuitive method for controlling exoskeletons in rehabilitation and physical enhancement settings. Thus, this approach paves the way for exoskeletons that can function effectively across a range of scenarios, making them more resilient to variations in muscle activity and improving the overall stability and reliability of AI-driven biomechanical systems. Vieth [18] focuses on improving the placement of pressure sensors in a smart shoe insole by exploring nonlinear modeling techniques. In previous work [19], they used a linear model to predict weight distribution on the foot and leg based on data from pressure sensors. While the linear model was effective, the current study demonstrates that the number of sensors can be reduced with nonlinear modeling techniques while maintaining robustness. Thus, their work reflects core aspects of robustness: maintaining reliable performance with fewer resources and demonstrating resilience in the face of reduced sensor input. In applications like smart insoles for healthcare, where consistent and accurate monitoring of weight distribution is crucial for diagnosing and treating mobility issues, this enhancement is critical.

Building on the theme of physical movement, we turn from human biomechanics to robotic locomotion. Hermes et al. [20] address the challenges of legged locomotion in robots, which is inherently more complex than wheeled or tracked movement due to the intricate coordination required between legs. Inspired by the biological coordination seen in insects, Hermes et al. [20] propose a decentralized control system for hexapod robots, using a Graph Neural Network to model inter-leg coordination. The decentralization of control mimics biological systems, where different parts of the body communicate and adapt locally, ensuring robustness in navigating difficult terrains. This decentralized approach enhances the robustness of the robot by allowing it to make local adjustments to its leg movements based on the specific challenges of the terrain. Preliminary results demonstrate how this gives rise to a stable tripod gait, highlighting how decentralized and flexible coordination, inspired by biology, produces robust and reliable solutions, particularly in autonomous robotics.

Just as Hermes et al. [20] seek to optimize control and coordination in robots, Posada-Moreno and Trimpe [21] focus on optimizing model performance by correcting biases and aligning predictions with expert knowledge. Concept extraction is a useful approach in explainable AI to identify model biases that may affect transparency and fairness (see also [14], this issue, for an extension to time-series data). Posada-Moreno and Trimpe [21] extend this idea by introducing Concept Regularization, a method that goes beyond simply identifying biases by embedding a regularization term during the retraining process, adjusting the model's sensitivities based on feedback from human experts. Importantly, the proposed Concept Regularization method addresses the robustness of AI systems by ensuring that identified biases do not compromise model performance. Thus, Posada-Moreno and Trimpe [21] demonstrate the critical role of feedback loops in enhancing the robustness and reliability of AI systems, particularly when addressing complex issues like bias.

From this extension of concept extraction, we move to an extension of the Multi-Armed Bandit framework in reinforcement learning, i.e., the Dueling Bandit problem. While traditional Dueling Bandit algorithms assume immediate feedback on which option performs better after the learner chooses two options, Brandt et al. [22] introduce a strategy that can start a new duel even if the feedback from the previous duel has not yet been observed. They demonstrate that this approach significantly improves the time efficiency of the algorithm by balancing the expected information gain and feedback delay. Considering that feedback delays are common on many dynamic environments, the contribution by Brandt et al. [22] offers a reliable solution for real-time applications, such as online recommendation systems and adaptive learning environments.

Finally, Moriz et al. [23] address the robustness of deep learning models trained on synthetic data and seeks to identify and mitigate factors contributing to performance gaps in real-world applications. With the increasing use of synthetic data for model training, the sim-to-real gap remains a significant challenge for broader adoption. Focusing on the impact of texture variation in synthetic validation sets for object detection, the work presented by Moriz et al. [23] concludes that texture properties alone do not fully account for the observed performance gap between synthetic and real datasets. This finding suggests that other factors, such as object size or illumination, may play a more critical role and warrant further investigation. By improving the reliability of synthetic training data, this research enhances the robustness of AI systems when deployed in real environments.

Overall, the contributions in this section illustrate the importance of building AI systems that are both reliable and robust in the face of real-world challenges. They offer solutions on how AI systems may be made more reliable, robust, and thus adaptable even in the face of real-world challenges

5. Bridging Themes

The development of trustworthy AI encompasses various domains, from performance optimization to the ethical considerations of human-AI interactions. However, achieving truly trustworthy AI systems requires integrating multiple perspectives and methodologies, building connections between diverse fields such as machine learning, human-computer interaction, and real-world applications. The research in this section highlight these interdisciplinary connections, offering innovative approaches that bridge the gap between technical advancements, user-centric designs, and transparency.

The contribution presented by Fischer and Bunse [24] lays the foundation for assessing the reliability and accountability of AI systems by presenting a "Sustainable and Trustworthy Reporting" (STREP) framework. The STREP framework presents a structured method for reporting performance indicators of AI systems, combining empirical data, theoretical algorithmic properties, and evaluation context. While AI systems are often evaluated in controlled environments, real-world applications introduce variables that can influence performance and trustworthiness. STREP contributed to understanding and communicating these nuances, thus enabling stakeholders to assess AI systems in a more comprehensive and transparent manner. This connection between data, knowledge, and context reflects the need for multidisciplinary approaches to create AI systems that are both technically robust and socially responsible.

Improving the performance and explainability of reinforcement learning by incorporating cognitive-inspired methods, Lange et al. [25] introduce techniques inspired by human cognition, such as enhanced state representations and causal reasoning. While traditional reinforcement learning systems rely heavily on trial-and-error learning, this extension provides an intriguing approach that may allow agents to reason about the past and future, explore hypothetical options, or learn from mistakes. Thus, the work presented by Lange et al. [25] demonstrates how AI systems can be designed to align better with human cognitive processes, creating models that are not only efficient but also more understandable and trustworthy.

The nuanced discussions in Belosevic and Buschmeier [26] and Schmidt and Cimiano [27] underscore the role of AI in sensitive settings, highlighting the need for reliable and understandable AI applications. Belosevic and Buschmeier [26] take a linguistic-based approach to understanding trust in interactions between humans and LLM-based chatbots, focusing on trust calibration, i.e., the process by which users adjust their trust in AI systems based on their interactions. Addressing the need for enhancing AI literacy and preventing overtrust or misuse of chatbot-provided information, particularly in educational contexts, their research offers critical insights into how users can be better supported in managing their trust levels. As a first step, Belosevic and Buschmeier [26] present a case study on the route to conversational agents capable of delivering proactive dialogue actions that assist students in controlling their trust in chatbot responses. This study begins by identifying the linguistic units perceived as trust cues in human-chatbot interactions, providing a foundation for recommendations on designing effective communication strategies that promote trustworthy AI. Schmidt and Cimiano [27] demonstrate how AI systems can leverage real-world data to impact healthcare outcomes.

Their study focuses on extracting information from online healthcare forums to automate the process of answering quality-of-life questionnaires for cancer patients. The results of their ongoing work could significantly reduce the burden on both patients and healthcare professionals, making it easier to assess patient well-being in real-time.

Bridging the gap between mathematical modeling and real-world applications, the contribution by Besginow et al. [28][2] leverages Gaussian processes for equation discovery in dynamical systems. Their method aims to uncover the differential equations that govern the physical processes observed in time series data. In contrast to traditional time series analysis with Gaussian processes, their approach seeks to identify the most frequently occurring differential equations within the data, offering a more refined understanding of the underlying system dynamics. Thus, their work provides a deeper understanding of complex, multi-component systems, contributing to the development of robust AI systems that can more accurately model and predict the behaviors of dynamical systems.

Finally, Sanaullah et al. [29][3] offer a comprehensive review of the evolution and optimization of artificial neural networks. Their paper examines improvements in network architecture, training algorithms, optimization techniques, and hardware acceleration, all of which have significantly enhanced the capabilities of neural networks. By analyzing the progression of deep learning models, such as convolutional neural networks and spiking neural networks, the review highlights their impact on critical areas like natural language processing and computer vision. Their contribution categorizes neural networks into distinct generations, emphasizing key milestones that have improved performance, scalability, and transparency—essential factors for building AI systems that are not only powerful but also reliable, interpretable, and aligned with human values, thus contributing to the broader goal of trustworthy AI.

6. Conclusions

The development of trustworthy AI is a multidimensional challenge, requiring attention to fairness, interpretability, reliability, and the integration of interdisciplinary approaches. Throughout this collection, various contributions highlight these essential facets, showcasing the importance of creating AI systems that are not only high-performing but also aligned with human values.

In the realm of fairness and ethics, research that explores fairness in rankings [1] or AI-led resource allocation in leadership roles [3] emphasizes the need for ethical decision-making frameworks in AI. These contributions underscore the importance of ensuring equitable outcomes and maintaining user trust in AI-driven processes. Interpretability and transparency are considered to be central to fostering trust [9]. The analysis of feature importance using Shapley values [10] and the development of inherently interpretable models [12] pave the way to empower users to make educated and well-informed evaluations of AI systems, particularly in high-stakes scenarios. Without transparency, even the most technically advanced AI systems risk losing credibility. Reliability and robustness are critical for AI systems operating in dynamic and unpredictable real-world environments. Advances in anomaly detection [15] or sensor optimization [18] highlight the necessity of building systems that maintain consistent performance under varying conditions, ensuring that AI systems can handle complexity without sacrificing reliability. Finally, it is important to recognize that achieving trustworthy AI requires more than just technical innovation: It is equally critical to integrating diverse perspectives and interdisciplinary approaches throughout the development, evaluation, and deployment of AI systems.

To conclude, the collective contributions to the DataNinja 2024 sAIOnARA Conference illustrate the multifaceted nature of building trustworthy AI. Technical performance must be complemented by ethical considerations, user transparency, and robust system design to ensure AI systems meet the needs of human users and society.

Acknowledgments

This research was supported by the research training group "Dataninja" (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins

[2] Besginow et al. [28] were awarded with the 3rd Place Best Poster Award at the DataNinja sAIOnARA 2024 Conference, Bielefeld.

[3] Sanaullah et al. [29] were awarded with the 2nd Place Best Poster Award at the DataNinja sAIOnARA 2024 Conference, Bielefeld.

Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

References

[1] Chiara Balestra. Is it possible to characterize group fairness in rankings in terms of individual fairness and diversity? In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 10–12, 2024. doi: 10.11576/dataninja-1157.

[2] Lloyd S Shapley. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

[3] Paul Hellwig and Günter W. Maier. Distributive justice of resource allocation through artificial intelligence. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 75–77, 2024. doi: 10.11576/dataninja-1177.

[4] Edna B Foa and Uriel G Foa. Resource theory of social exchange. *Handbook of social resource theory: Theoretical extensions, empirical insights, and social applications*, pages 15–32, 2012.

[5] Clarissa Sabrina Arlinghaus and Günter W. Maier. Feeling socially excluded when working with robots. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 36–38, 2024. doi: 10.11576/dataninja-1165.

[6] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

[7] Sanaullah, Hasina Attaullah, and Thorsten Jungeblut. Trade-offs between privacy and performance in encrypted dataset using machine learning models. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 39–42, 2024. doi: 10.11576/dataninja-1166.

[8] Ali Norouzifar and Wil van der Aalst. Leveraging desirable and undesirable event logs in process mining tasks. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 32–35, 2024. doi: 10.11576/dataninja-1164.

[9] Umang Bhatt, Pradeep Ravikumar, et al. Building human-machine trust via interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9919–9920, 2019.

[10] Patrick Kolpaczki. Comparing shapley value approximation methods for unsupervised feature importance. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 13–15, 2024. doi: 10.11576/dataninja-1158.

[11] Feng-Lei Fan, Jinjun Xiong, Mengzhou Li, and Ge Wang. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6):741–760, 2021.

[12] Simon Lutz and Daniel Neider. Interpretable machine learning via linear temporal logic. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 72–74, 2024. doi: 10.11576/dataninja-1176.

[13] Tim Katzke, Simon Lutz, Emmanuel Müller, and Daniel Neider. Provable guarantees for deep learning-based anomaly detection through logical constraints. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 65–68, 2024. doi: 10.11576/dataninja-1174.

[14] Antonia Holzapfel, Andres Felipe Posada-Moreno, and Sebastian Trimpe. Concept extraction for time series with eclad. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 78–80, 2024. doi: 10.11576/dataninja-1178.

[15] Franziska Zelba, Stefanie Hittmeyer, and Gesa Benndorf. Cometh—an active learning approach enhanced with large language models. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and*

Achievements for Reliable Approaches, pages 23–25, 2024. doi: 10.11576/dataninja-1161.

[16] Lukas Krebs, Tobias Müller, and Robert H. Schmitt. Trustworthy virtual measurements in battery manufacturing. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 81–83, 2024. doi: 10.11576/dataninja-1179.

[17] Nils Grimmelsmann, Malte Mechtenberg, Markus Vieth, Barbara Hammer, and Axel Schneider. Prediction of intermuscular cocontraction based on the semg of only one muscle with the same biomechanical direction of action. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 47–49, 2024. doi: 10.11576/dataninja-1168.

[18] Markus Vieth. Nonlinear prediction in a smart shoe insole. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 50–52, 2024. doi: 10.11576/dataninja-1169.

[19] Markus Vieth, Nils Grimmelsmann, Axel Schneider, and Barbara Hammer. Efficient sensor selection for individualized prediction based on biosignals. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 326–337. Springer, 2022.

[20] Luca Hermes, Barbara Hammer, and Malte Schilling. Bioinspired decentralized hexapod control with a graph neural network. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 53–55, 2024. doi: 10.11576/dataninja-1170.

[21] Andres Felipe Posada-Moreno and Sebastian Trimpe. Closing the loop with concept regularization. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 62–64, 2024. doi: 10.11576/dataninja-1173.

[22] Jasmin Brandt, Björn Haddenhorst, Viktor Bengs, and Eyke Hüllermeier. Dueling bandits with delayed feedback. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 29–31, 2024. doi: 10.11576/dataninja-1163.

[23] Alexander Moriz, Dominik Wolfschläger, and Robert H. Schmitt. Study on the influence of texture variation on the validation performance of a synthetically trained object detector. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 69–71, 2024. doi: 10.11576/dataninja-1175.

[24] Raphael Fischer and Mirko Bunse. Improving trust in ai through sustainable and trustworthy reporting. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 56–57, 2024. doi: 10.11576/dataninja-1171.

[25] Moritz Lange, Raphael C. Engelhardt, Wolfgang Konen, and Laurenz Wiskott. Beyond trial and error in reinforcement learning. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 58–61, 2024. doi: 10.11576/dataninja-1172.

[26] Milena Belosevic and Hendrik Buschmeier. Linguistic-based reflection on trust calibration in conversations with llm-based chatbots. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 19–22, 2024. doi: 10.11576/dataninja-1160.

[27] David M. Schmidt and Philipp Cimiano. Question answering from healthcare fora. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 16–18, 2024. doi: 10.11576/dataninja-1159.

[28] Andreas Besginow, Jan David Hüwel, Markus Lange-Hegermann, and Christian Beecks. Finding commonalities in dynamical systems with gaussian processes. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI:*

Opportunities, Innovation, and Achievements for Reliable Approaches, pages 26–28, 2024. doi: 10.11576/dataninja-1162.

[29] Sanaullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Advancements in neural network generations. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 43–46, 2024. doi: 10.11576/dataninja-1167.