Interpretable Neural System Dynamics: Combining Deep Learning with System Dynamics Modeling to Support Critical Applications

Riccardo D'Elia[1]

[1]*University of Applied Sciences and Arts of Southern Switzerland, Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland*

Abstract

The objective of this proposal is to bridge the gap between Deep Learning (DL) and System Dynamics (SD) by developing an **interpretable neural system dynamics** framework. While DL excels at learning complex models and making accurate predictions, it lacks interpretability and causal reliability. Traditional SD approaches, on the other hand, provide transparency and causal insights but are limited in scalability and require extensive domain knowledge. To overcome these limitations, this project introduces a Neural System Dynamics pipeline, integrating Concept-Based Interpretability, Mechanistic Interpretability, and Causal Machine Learning. This framework combines the predictive power of DL with the interpretability of traditional SD models, resulting in both causal reliability and scalability. The efficacy of the proposed pipeline will be validated through real-world applications of the EU-funded AutoMoTIF project, which is focused on autonomous multimodal transportation systems. The long-term goal is to collect actionable insights that support the integration of explainability and safety in autonomous systems.

Keywords

Explainable Artificial Intelligence (XAI), Neuro-symbolic AI, Causal Machine Learning, System Dynamics.

1. Context and Motivation

The field of System Dynamics (SD) has long focused on modeling complex systems that underpin many application domains. In transportation logistics, for example, dynamical systems are used to model supply chain operations, traffic congestion, fleet management, and urban mobility planning.

Traditional SD models rely on differential equations and expert-defined rules to represent the evolution of a system over time. These models provide interpretable causal pathways and have long been valued for their transparency and accountability. However, they are constrained by simplifying assumptions that often fail to capture the full complexity of real-world systems and suffer from limited scalability as the number of interacting variables increases.

Contemporary Deep Learning (DL) techniques offer a promising alternative to overcome the aforementioned limitations of traditional SD modeling. DL algorithms are indeed capable of learning automatically the non-linear relationships that underpin dynamical systems' behaviors from large-scale data [1], thereby supporting the development of scalable and highly precise predictive models. Yet, these gains do not come costless. Unlike traditional SD models, which involve concepts and inferential rules that are easily understandable by their users, DL models operate as sort of "black boxes", whose semantics and the decision-making logic behind their outputs remain mostly incomprehensible to users. Furthermore, DL algorithms exploit predictions based on correlations, ignoring causal dependencies and mechanisms; this is referred to as a lack of *causal reliability* [2]. Several methods have recently been proposed to overcome the opacity issues of DL models, which fall under the umbrella term of eXplainable AI (XAI). These methods provide valuable insight into how the DL models operate and produce their results. However, existing XAI techniques mostly fail to provide models with well-defined semantics that are

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

riccardo.delia@idsia.ch (R. D'Elia)
0009-0009-6172-4049 (R. D'Elia)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

interpretable to their users. This is because most available XAI methods are *post-hoc*, inspecting the behaviors of naturally opaque models after training rather than trying to embed interpretable features directly within a model's structure. Moreover, these methods are mostly incapable of addressing causal reliability problems as these fall beyond their usual target scope. This severely limits the usefulness of these methods in system dynamics, where interpretability and causal reliability are equally fundamental challenges that go hand in hand. Additionally, the difficulty of certifying deep learning systems due to their limited explainability poses significant safety concerns in critical applications [3]. This gap is increasingly being explored in the emerging field of *Neuro-symbolic AI*, which integrates neural networks and symbolic reasoning to create interpretable and data-driven models. Recent advances in this field [4] align with the goals of this research, which seeks to combine System Dynamics with Deep Learning, as detailed in Fig. 1. To address the challenges of interpretability and causal reliability in DL-based dynamical systems modeling, this project moves away from post-hoc techniques and instead focuses on the construction of an **interpretable by design neural systems dynamic** framework. A plethora of different methods will be implemented to allow the combination of DL with the formalism of SD. In particular, the focus will be on techniques from the fields of *Concept-Based Interpretability* [5], as well as *Mechanistic Interpretability* [6] and *Causal Machine Learning* [7]. The pipeline will be structured as follows, with a more detailed description provided in section 3.

As a first step, concept-based interpretability methods will be employed to identify a set of semantically meaningful high-level variables (termed "concepts") that describe understandable characteristics and magnitudes of interest. CML techniques will then be implemented to detect the causal dependencies among the selected high-level variables. Finally, mechanistically interpretable modeling techniques will be leveraged to infer a set of interpretable structural dynamic equations that govern the system's behavior. Such equations will be determined by taking into account the previously identified causal dependencies. This will not only allow for increased interpretability but will also contribute to anchoring the models to the real-world causal structure, making them substantially more reliable and trustworthy for safety-critical applications.

As a final result, this pipeline should be able to return neural models that track the evolution of systems over time, both by operating on **semantically meaningful** and **actionable** variables. It will be implemented and evaluated on a real-world scenario from the EU-funded project *AutoMoTIF*, where SD is involved in modeling the interoperability of multi-modal transportation terminals. A more detailed description of the real-world application is provided in Sec. 1.1.

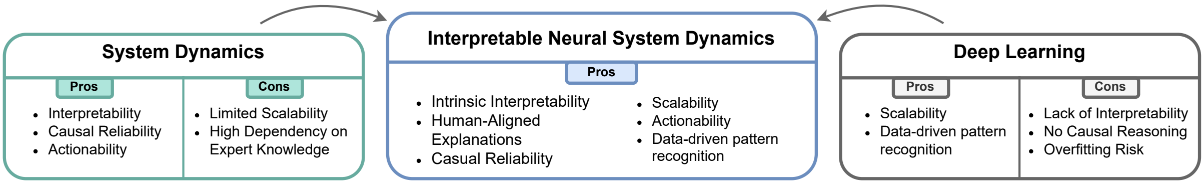

Figure 1: A unified framework combining System Dynamics and Deep Learning for Interpretable Neural System Dynamics.

1.1. Real-World Application: EU Project AutoMoTIF

The doctoral research proposed herein will be carried out under the EU-funded *AutoMoTIF* project[1] (*Automation towards multimodal transportation and integration of freight*). The project's core focus lies in the formulation of strategies, the development of business and governance models, and the generation of regulatory recommendations. These are designed to facilitate the integration and interoperability of automated transport systems. The project's overarching objective is to automate multimodal freight flows and logistics supply chains within the intra-European network, thereby enhancing operational efficiency and addressing existing regulatory and technological gaps.

[1] https://automotif-project.eu

Within this project, System Dynamics plays a crucial role in modeling and optimizing multimodal terminal operations, helping stakeholders to analyze and predict system behavior under different operational conditions. However, the complexity of these environments calls for data-driven AI approaches, which, despite their predictive power, often lack interpretability and causal reliability — critical aspects for risk assessment and certification. This challenge aligns with the broader Trustworthy AI paradigm, which underscores transparency, reliability, and human oversight [8]. Within this framework, the Trustworthy Autonomous Systems (TAS) research field focuses on developing methods to enhance AI accountability, explainability, and resilience in real-world deployments. Reflecting these concerns, the EU AI Act classifies AI-driven transport automation as a high-risk domain, requiring rigorous risk assessment, explainability, and robustness [9]. By integrating this research into *AutoMoTIF*, the proposed pipeline will be tested in a real-world and high-pressure environment where understanding is crucial for ensuring safety, compliance, and trust in autonomous systems.

2. The Many-Faces of the Interpretability Challenge

Whilst the importance of eXplainable AI is becoming increasingly acknowledged, achieving interpretability remains a complex process. The opacity of DL algorithms represents a major challenge for contemporary AI research. In particular, the DL community must navigate various forms of opacity, which vary based on the different aspects of a model's structure and functioning that are focused on, as well as on the specific users and stakeholders involved [10]. The present project focuses specifically on two primary kinds of opacity that are of central relevance for research, notably *semantic* and *mechanistic* opacity [11].

Semantic opacity refers to the challenge of deciphering what a model's learned representations mean in terms understandable to humans. This issue is particularly relevant in Neural Networks (NNs), where internal representations are often abstract and distributed without explicit meanings. Humans generally process information through high-level concepts, whereas NNs operate within a multi-dimensional feature space that obscures the semantics of the features involved and how these relate to categories comprehensible by the layman user.

Mechanistic opacity refers to the difficulty of detailing precisely the mechanisms through which the various components of a model interact with one another and thus contribute to generating the overall model's behavior [12]. This issue is particularly significant in large-scale NNs with millions or even billions of parameters, where computations are spread across numerous layers and involve several non-linear transformations.

In the XAI literature, these two different opacity forms have been mostly addressed separately by referring to different paradigms and implementing distinct strategies and techniques. This fragmented approach obstructs the creation of a cohesive, mathematically rigorous framework capable of addressing multiple interpretability challenges that stem from considering these two aspects of opacity together. Explaining the mechanisms underpinning the inference process of a DL model (e.g., via equation modeling) contributes minimally to the overall interpretability of the model's behavior if the features remain low-level and semantically meaningless. Conversely, mapping low-level features to high-level concepts has limited value if the model's decision-making mechanisms remain opaque.

Causal Reliability. Related to the aforementioned forms of opacity is another fundamental issue, which is orthogonal to the problem of (mechanistic and semantic) opacity but intrinsically connected to it. This is the problem we refer to as *causal reliability* [2]. This problem concerns the (in)ability of a model to track the real-world causal mechanisms operating beyond observable data generation and take them into account when drawing predictions. DL models are built to identify correlations among features and generate predictions solely based on them while ignoring the causal mechanisms. This poses DL algorithms in contrast with traditional mechanistic models, widely involved especially in the field of system dynamics. The latter, indeed, embeds an explicit representation of the causal mechanisms beyond data. The lack of reliance on real-world causal mechanisms represents a major limitation for DL algorithms, notably as it undermines their robustness and generalizability, especially in *out-of-distribution* contexts [13]. Furthermore, this issue limits the actionability of DL models, limiting the possibility of users intervening properly in their inferential processes and analyzing related interventional and counterfactual scenarios [14].

The Need for an Integrated Approach. Opacity and causal reliability have been mostly treated as separate issues in contemporary AI research. Indeed, while opacity represents the target problem of XAI, causal reliability is at the heart of another growing research field, that of *Causal Machine Learning* (CML) [13, 7]. The two fields have developed separately with little connection among each other [15]. However, the two problems have a close relationship, especially when we focus on modeling dynamical systems. In response to these limitations, this doctoral project aims to propose a cohesive framework that jointly addresses the two aforementioned forms of opacity and, at the same time, produces causally reliable models. The project can be seen as an attempt to combine the three research domains of semantic ("concept-based") explainability, mechanistic interpretability, and causal reliability, with specific reference to the field of dynamical systems modeling and its applications.

3. Research Strategy and Rationale

This project proposes a novel *Interpretable Neural System Dynamics (INSD) pipeline* that combines Concept-Based Interpretability, Causal Learning, and Mechanistic Interpretability to construct causally-reliable neural system dynamics models that operate on human-interpretable variables while preserving the flexibility and scalability of DL approaches (Figure 2). The pipeline consists of three distinct learning steps:

1. *Concept Learning:* In the first step, concept-based interpretability (CBI) methods are used to extract high-level semantically interpretable variables ("concepts") from raw data.
2. *Causal Learning:* In the second step, causal machine learning (CML) and causal discovery (CD) techniques are leveraged to identify the causal dependencies among these high-level concepts, thereby representing them in the form of a *causal directed graph*.
3. *Equation Learning:* In the third step, mechanistic interpretability methods are involved to derive explicit and interpretable dynamic equations that allow to predict the behavior of the target-system over time.

After the three learning steps, the final model integrates the learned concepts, causal relationships, and governing equations to emulate the underlying dynamical system. Unlike traditional black-box neural networks, this model offers full interpretability, enabling users to trace predictions back to meaningful variables and causal influences. Furthermore, it provides actionable insights, allowing decision-makers to simulate interventions, predict long-term effects, and better understand the system's behavior under different conditions. This ensures both transparency and practical applicability, bridging the gap between deep learning's flexibility and human-comprehensible system dynamics. The following paragraphs provide a detailed breakdown of each methodological component.

3.1. Understanding System Dynamics Through Concept-Based Interpretability

When applied to system dynamics, DL algorithms typically generate latent representations of system states that are usually not understandable and are arduous to interpret. For instance, in the context of epidemiological modeling, deep learning algorithms have the potential to discern underlying patterns of disease transmission; however, they have difficulty formulating these patterns using conventional epidemiological factors, such as *contact rate* or *incubation period*. This semantic opacity restricts their reliability for critical decision-making processes. Concept-based Explainable AI introduces a potentially effective solution to these limitations by aligning AI reasoning with human-understandable abstractions rather than *opaque* latent representations [5]. Despite these innovations, this approach remains opaque regarding other aspects, such as the intrinsic mechanisms underlying concept representational learning.

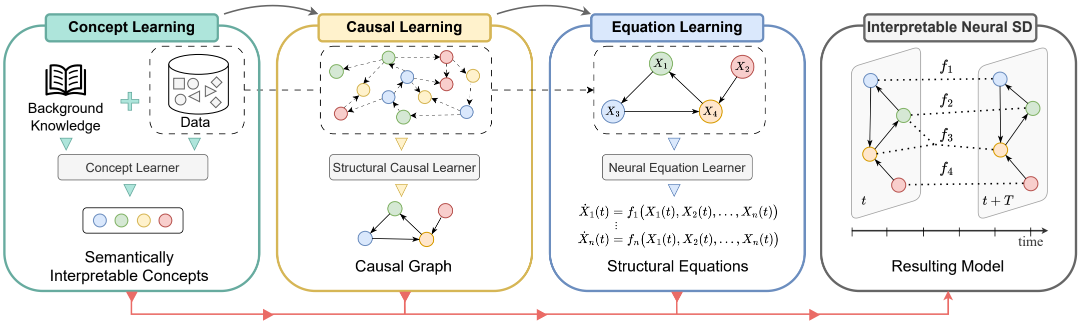

Figure 2: Overview of the INSD pipeline, from concept learning to causal and equation learning, ensuring interpretability and causal reliability in the resulting model.

Concept-based models are currently underdeveloped concerning the temporal dimension, i.e. the evolution of interpretable concepts over time. This project would enable human interventions over evolving representations, constituting a significant advance. In addition, this class of methods faces generalization and compositionality challenges because, similar to standard deep learning architectures, they are essentially associative models [14]. This suggests that their decision-making process is not aligned with the underlying causal mechanisms of the world. They must distinguish regularities in data that reflect true causal relationships from those that are spurious. It is, therefore, crucial to comprehend this distinction to develop a robust and reliable understanding of phenomena, as well as to support intervention planning and ensure the application of fairness constraints [13].

Running Example: Automated Terminal Operation in *AutoMoTIF*. Within an intermodal terminal setting, while a DL model might forecast freight congestion patterns, it may not clarify the reasons behind delays. By employing concept-based interpretability, logistics-related concepts such as terminal workload, handling efficiency, and waiting times for various transport modes are incorporated. This ensures predictions reflect real-world operational factors accurately. Consequently, this approach makes AI-driven simulations more transparent and actionable for terminal operators.

3.2. The Role of Causality in DL-based System Dynamics Models

Causal reasoning plays a crucial role in System Dynamics, as these models explicitly represent causal connections between system variables. Contrarily, DL-based approaches rely on statistical correlations rather than true causal frameworks, which limits their ability to provide strong, interpretable predictions. A promising direction for overcoming these limitations is provided by recently developed CML techniques and, in particular, the framework of Neural Causal Models (NCMs). These techniques aim to uncover the underlying causal structure of a system by learning a graph that captures causal dependencies between concepts. Based on this learned graph, we can then infer equations that describe the system's evolution in a causally reliable manner, ensuring that the resulting models generalize more robustly and provide deeper insights into the underlying mechanisms governing the data.

Running Example: Automated Terminal Operation in *AutoMoTIF*. A DL model might predict regular train delays at an intermodal terminal and identify a correlation between high truck traffic and these delayed departures. However, without causal reasoning, the underlying cause can remain unclear. A causal model could determine whether truck congestion is directly causing train delays or if another external factor, such as inefficient crane operation, is primarily responsible. By simulating counterfactual scenarios, such as "What if we increased crane availability?", causal DL enables logistics operators to make proactive and data-driven decisions.

3.3. Understanding System Dynamics Through Mechanistic Interpretability

Mechanistic Interpretability seeks to unravel DL models by reverse-engineering them to reveal their internal structures and decision-making processes. This research field is highly pertinent to System Dynamics, where comprehending a model's inner workings holds equal importance as its predictive accuracy. Traditional System Dynamics models use clearly defined equations and feedback loops, which make them inherently easy to interpret. The formal use of a causal loop diagram to describe a feedback system naturally leads to a connection with graph-based AI architectures like Graph Neural Networks (GNNs). GNNs offer an intuitive and organized way to represent dynamical systems, bringing benefits regarding intrinsic interpretability by exploiting relational inductive bias. Specifically, mechanistic interpretability techniques can help to trace information propagation to understand internal interactions. This is essential to develop structured, human-interpretable representations of dynamic processes.

> **Running Example: Automated Terminal Operation in *AutoMoTIF*.** Consider a simulation of an intermodal terminal where a DL-based model recommends rerouting trucks via a secondary access road. Without mechanistic interpretability, the reasoning behind this decision remains unclear — whether it stems from predicted congestion, infrastructure constraints, or other operational factors. Taking advantage of the intrinsically interpretable model, we can identify modular components within the model, track the flow of information within the model, and uncover the key interactions influencing routing choices. This structured understanding enhances both the interpretability and trustworthiness of AI-driven logistics systems.

4. Research Questions and Objectives

This research is guided by the following key questions:

RQ1: How can a system dynamics framework ensure *transparency* and *accountability* while maintaining the predictive power of deep learning?
Working Hypothesis: Integrating traditional equation-based modeling with deep learning can achieve this balance.

RQ2: How can traditional modeling techniques and deep learning be effectively combined to leverage their strengths?
Working Hypothesis: An optimal integration may involve combining three key areas of XAI research: *concept-based interpretability*, *mechanistic interpretability*, and *causal machine learning*.

RQ3: How can methods from the aforementioned distinct fields be integrated to develop an interpretable neural system dynamics framework?
Working Hypothesis: Concept-based interpretability can be used to learn high-level concepts and map raw data to meaningful variables, causal learning to infer dependencies among these variables, and mechanistic interpretability to define and parameterize system equations.

To refine this framework, the investigation focuses on: *(i)* how concept-based techniques can identify high-level variables from raw data; *(ii)* how causal learning can infer dependencies using data and background knowledge; and *(iii)* how to determine the structure and parameters of governing equations in an interpretable manner. Additionally, the framework's adaptability for modeling intermodal transportation logistics is examined, with an emphasis on ensuring safety and accountability. The integration of traditional equation-based modeling with deep learning, leveraging methods from these XAI fields, is hypothesized to achieve an interpretable and reliable system dynamics framework. This approach ensures *partial verifiability, actionability, and control* in real-world applications.

Specific Objectives and Milestones. The project's **main objective** is to design, implement, and evaluate an integrated pipeline combining system dynamics modeling with deep learning, applying it to *AutoMoTIF* and assessing generalisability. To achieve this objective, the following intermediate milestones are planned:

M1: State-of-the-Art Analysis. Review concept-based, mechanistic, and causal learning interpretability via a systematic literature review [**D1**: survey paper].

M2: Pipeline Blueprint. Design a blueprint for integrating dynamical system modeling with deep learning through literature review and theoretical modeling [**D2**: short conference paper].

M3: Use-Case Implementation. Develop and validate the pipeline on a toy example [**D3**: conference paper (e.g., NeurIPS, ICML)].

M4: Real-World Scalability Study. Scale the pipeline to *AutoMoTIF* [**D4**: applied research paper].

M5: Generalization Study. Assess applicability to other real-world scenarios [**D5**: generalization study report or journal paper].

M6: Dissertation Completion. Compile research findings into the PhD thesis [**D6**: dissertation].

The project will run for 48 months. It started in January 2025, and termination is planned for December 2028. The structure of the project's timeline is reported in Fig. 3.

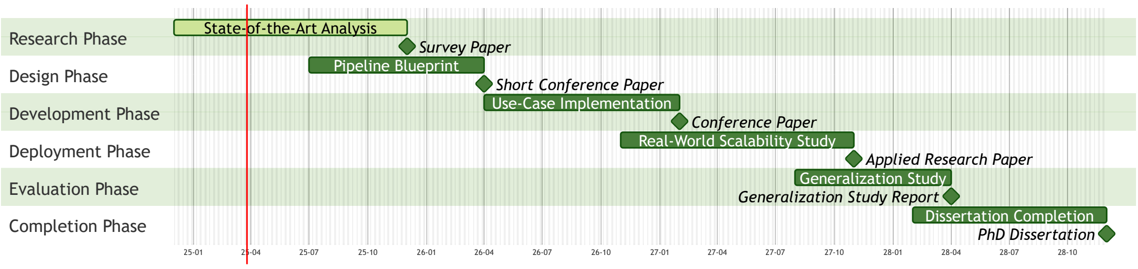

Figure 3: Gantt diagram of the doctoral project.

Expected Contribution and Impact. Through the development of a unified interpretability framework for DL-based System Dynamics models, this research aspires to bridge the current divide between theoretical advancements in eXplainable AI and their application in high-stakes, real-world environments. By bringing together causal, mechanistic, and concept-based perspectives within a cohesive methodology, the project is expected to deliver not only novel algorithms and formal models but also practical tools that empower users to understand, trust, and effectively intervene in AI-driven processes. The anticipated contributions extend beyond the specific context of multimodal logistics, offering generalizable insights for any domain that relies on the interplay of Deep Learning and System Dynamics. These insights may be applied to a wide range of fields, including, but not limited to, transportation, healthcare, environmental monitoring, and finance. Ultimately, this work aims to establish both a conceptual and operational foundation for interpretable, trustworthy AI in settings where transparency and accountability are not optional but essential for safety, compliance, and societal acceptance.

Acknowledgments

This work was partly supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract no. 24.00184 (AutoMoTIF project). The project has been selected within the EU Horizon Europe programme under grant agreement no. 101147693. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the funding agencies, which cannot be held responsible for them.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] G. Quaranta, W. Lacarbonara, S. F. Masri, A review on computational intelligence for identification of nonlinear dynamical systems, Nonlinear Dynamics 99 (2020) 1709–1761. URL: http://link.springer.com/10.1007/s11071-019-05430-7. doi:10.1007/s11071-019-05430-7.

[2] A. Termine, G. Primiero, Causality Problems in Machine Learning Systems, in: The Routledge Handbook of Causality and Causal Methods, 1 ed., Routledge, New York, 2024, pp. 325–341. URL: https://www.taylorfrancis.com/books/9781003528937/chapters/10.4324/9781003528937-37. doi:10.4324/9781003528937-37.

[3] F. Flammini, C. Alcaraz, E. Bellini, S. Marrone, J. Lopez, A. Bondavalli, Towards Trustworthy Autonomous Systems: Taxonomies and Future Perspectives, IEEE Transactions on Emerging Topics in Computing 12 (2024) 601–614. URL: https://ieeexplore.ieee.org/document/9979717/. doi:10.1109/TETC.2022.3227113.

[4] B. P. Bhuyan, A. Ramdane-Cherif, R. Tomar, T. P. Singh, Neuro-symbolic artificial intelligence: a survey, Neural Computing and Applications 36 (2024) 12809–12844. URL: https://link.springer.com/10.1007/s00521-024-09960-z. doi:10.1007/s00521-024-09960-z.

[5] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, E. Baralis, Concept-based Explainable Artificial Intelligence: A Survey, 2023. URL: http://arxiv.org/abs/2312.12936. doi:10.48550/arXiv.2312.12936, arXiv:2312.12936 [cs].

[6] L. Bereska, E. Gavves, Mechanistic Interpretability for AI Safety – A Review, 2024. URL: http://arxiv.org/abs/2404.14082. doi:10.48550/arXiv.2404.14082, arXiv:2404.14082 [cs].

[7] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, R. Silva, Causal Machine Learning: A Survey and Open Problems, 2022. URL: https://arxiv.org/abs/2206.15475. doi:10.48550/ARXIV.2206.15475, version Number: 2.

[8] A. Bellogín, O. Grau, S. Larsson, G. Schimpf, B. Sengupta, G. Solmaz, The EU AI Act and the Wager on Trustworthy AI, Communications of the ACM 67 (2024) 58–65. URL: https://dl.acm.org/doi/10.1145/3665322. doi:10.1145/3665322.

[9] M. Borrelli, S. Musch, S. Khan, Applying the EU AI Act to the Automotive Industry: Ensuring Explainability and Transparency, SSRN Electronic Journal (2024). URL: https://www.ssrn.com/abstract=4819226. doi:10.2139/ssrn.4819226.

[10] K. Sokol, P. Flach, Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence, 2022. URL: http://arxiv.org/abs/2112.14466. doi:10.48550/arXiv.2112.14466, arXiv:2112.14466 [cs].

[11] A. Facchini, A. Termine, Towards a Taxonomy for the Opacity of AI Systems, in: V. C. Müller (Ed.), Philosophy and Theory of Artificial Intelligence 2021, volume 63, Springer International Publishing, Cham, 2022, pp. 73–89. URL: https://link.springer.com/10.1007/978-3-031-09153-7_7. doi:10.1007/978-3-031-09153-7_7, series Title: Studies in Applied Philosophy, Epistemology and Rational Ethics.

[12] L. Kästner, B. Crook, Explaining AI through mechanistic interpretability, European Journal for Philosophy of Science 14 (2024) 52. URL: https://link.springer.com/10.1007/s13194-024-00614-4. doi:10.1007/s13194-024-00614-4.

[13] B. Scholkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Toward Causal Representation Learning, Proceedings of the IEEE 109 (2021) 612–634. URL: https://ieeexplore.ieee.org/document/9363924/. doi:10.1109/JPROC.2021.3058954.

[14] J. Pearl, Causality: Models, Reasoning, and Inference, 2 ed., Cambridge University Press, 2009. URL: https://www.cambridge.org/core/product/identifier/9780511803161/type/book. doi:10.1017/CBO9780511803161.

[15] G. Carloni, A. Berti, S. Colantonio, The role of causality in explainable artificial intelligence, 2023. URL: http://arxiv.org/abs/2309.09901. doi:10.48550/arXiv.2309.09901, arXiv:2309.09901 [cs].