# Unifying Concept Representation Learning

**Amit Dhurandhar** **Amir-Hossein Karimi** **Sara Magliacane**

**Stefano Teso** **Efthymia Tsamoura** **Zhe Zheng**

## 1 Main Proposal

**Motivation**. Several areas at the forefront of AI research are currently witnessing a convergence of interests around the problem of *learning high-quality concepts from data*. Concepts have become a central topic of study in ***neuro-symbolic integration*** (NeSy). NeSy approaches integrate perception – usually implemented by a neural backbone – and symbolic reasoning by employing concepts to glue together these two steps: the latter relies on the concepts detected by the former to produce suitable outputs (De Raedt et al., 2021; Tsamoura et al., 2021; Garcez et al., 2022; Dash et al., 2022; Giunchiglia et al., 2022). Concepts are also used in ***Explainable AI*** (XAI) by recent post-hoc explainers (Dhurandhar et al., 2018a; Luss et al., 2021; Kazhdan et al., 2020; Bereska & Gavves, 2025) and self-explainable architectures (Koh et al., 2020; Puri et al., 2021; Dhurandhar et al., 2018b; Espinosa Zarlenga et al., 2022) as a building block for constructing high-level justifications of model behavior. Compared to, e.g., saliency maps, these can portray a more abstract and understandable picture of the machine's reasoning process, potentially improving understandability, interactivity, and trustworthiness (Karimi et al., 2022; Kim et al., 2018; Poeta et al., 2023; Teso et al., 2023), to the point that concepts have been called the *lingua franca* of human-AI interaction (Kambhampati et al., 2022). Both areas hinge on learned concepts being "*high-quality*". Concepts with misaligned semantics may compromise the meaning of model explanations (Mahinpei et al., 2021; Margeloiu et al., 2021; Sucholutsky et al., 2023), reliability of NeSy architectures (Marconato et al., 2023b; Wang et al., 2024) and human understanding (Marconato et al., 2023a).

Recent works (Marconato et al., 2023b;a) propose to leverage disentangled representations to mitigate *concept leakage*, i.e., the presence of irrelevant information in the learned concepts. ***Causal Representation Learning*** (CRL) aims to identify latent causal variables and causal relations from high-dimensional observations, e.g., images or text, with theoretical guarantees (Schölkopf et al., 2021). As such, CRL is a generalization of disentangled representation learning, when the latent variables are dependent on each other, e.g., due to causal relations. CRL has been increasingly popular, with a plethora of methods and theoretical results (Khemakhem et al., 2020; Lippe et al., 2022; Lachapelle et al., 2022; Yao et al., 2022; Huang et al., 2022; Lippe et al., 2023; Ahuja et al., 2024; Lachapelle et al., 2024; Yao et al., 2024; Zhang et al., 2024; Xu et al., 2024). The potential of leveraging CRL to learn more robust and leak-proof concept is an emerging area of research with a growing number of approaches (Marconato et al., 2023a; Kong et al., 2024; Rajendran et al., 2024; Liu et al., 2025; Fokkema et al., 2025), but many open questions remain.

In particular, what properties high-quality concepts should satisfy is unclear, and – despite studying the same underlying object – research in these areas is proceeding on mostly independent tracks, with minimal knowledge transfer. Separate branches differ in their working definitions of what concepts are and what desiderata they ought to satisfy, on what data and algorithms they should be learned with, and on how to properly assess their quality. Unfortunately, efforts at adapting ideas and techniques are limited at best, meaning that approaches in one area completely ignore insights from the others. As a result, the central issue of how to properly learn and evaluate concepts is largely unanswered. This state of affairs hampers progress on foundations and applications alike.

**Purpose** This workshop brings together researchers from NeSy, XAI and CRL and from both industry and academia, who are interested in learning robust, semantically meaningful concepts. By facilitating informal discussion between experts and newcomers alike, it aims to *tie together these currently independent strands of research and promote cross-fertilization.*

The workshop accepts submissions on the following ***topics***:

- Foundations of concept representations and learning in XAI, CRL and NeSy.
- Supervised and unsupervised techniques for learning concepts from observational and interventional data, raw inputs, and pre-trained embeddings.
- Techniques for learning concepts in non-standard settings, e.g., causal abstraction.
- Design and evaluation of concept-based XAI techniques and self-explainable concept-based models.
- Interactive human-machine concept acquisition and alignment.
- Applications of concept-based AI systems, including but not limited to, reasoning, causality, formal verification, interactive learning, and explainability.
- Metrics and evaluation techniques for assessing the quality of learned concepts, with a focus on down-stream applications.

The workshop will feature a **short paper track** focusing on late-breaking results and reassessment of existing results to facilitate participation for younger researchers. All accepted submissions will have to abide to the LLM guidelines proposed by ICLR, except short papers, for which LLM usage will be restricted to polishing.

**Related events** This is the first edition of this workshop. Past events have not targeted learned concepts explicitly, especially not in a multi- and inter-disciplinary fashion. Venues dedicated to explainable AI (e.g., the World Conference on XAI and workshops and the International Workshop on Explainable and Interpretable Machine Learning), for instance, do not emphasize the role of concepts in XAI. Similarly, venues on Neuro-Symbolic AI (primarily, the NeSy conference and the International Joint Conference on Learning and Reasoning) focus on relational data, reasoning techniques and scalability issues rather than on the properties of learned concepts. Causal representation learning has had several dedicated workshops (2022 @ UAI, 2023 @ Max Planck Institute in Tübingen, 2023 @ NeurIPS, and 2024 @ NeurIPS), however it is focused on causality aspects only. The most closely related workshops are nCSI @ NeurIPS 22 (19 accepted papers), UniReps at NeurIPS 23 (72 papers) and NeurIPS 24 (68 papers). The former focused on NeSy and causality but ignored XAI, while the latter focused more broadly on representation learning and, in part, mechanistic interpretability, rather than symbols and their semantics.

**Attendance** Given the above figures, the increasing attention that concepts are receiving, and our own experience with previous workshops, we believe our workshop's will draw significant interest from the research community and estimate around 60-80 submissions, 30-40 accepted papers, and about 100 attendees.

**Logistics** Our schedule is designed to offer a good balance of invited and contributed talks and opportunities for interaction. We will have 6 invited speakers, listed below, each with 30 minutes (talk + Q&A) and two sessions of 30 minutes for 4 contributed talks selected from the highest ranked papers (15 minutes for each talk + Q&A). We will have two 1-hour poster sessions (with around 15-20 posters each), one before lunch and the other in the afternoon. At the end of the workshop, we will have a panel discussion with the invited speakers. To make the event more interesting, we will "challenge" the panelists to comment on problematic issues and hot topics revolving around concepts. The schedule provides ample room for discussion among participants in poster sessions (intentionally scheduled next to coffee and lunch breaks) and in the panel discussion, so as to foster interaction among participants. Our tentative schedule is (times are UTC-3, deadlines are AoE):

| | | | | | |
|---|---|---|---|---|---|
| 09:00 - 09:10 | Introduction | 13:30 - 14:00 | Invited talk 4 | Jul 11 | Website online, send out CFP. |
| 09:10 - 09:40 | Invited talk 1 | 14:00 - 14:20 | Coffee break | | |
| 09:40 - 10:10 | Invited talk 2 | 14:20 - 14:50 | Invited talk 5 | Aug 22 | Paper submission deadline. |
| 10:10 - 10:40 | Contributed talks | 14:50 - 15:20 | Invited talk 6 | Sept 15 | Review deadline. |
| 10:40 - 11:00 | Coffee break | 15:20 - 15:50 | Contributed talks | Sept 22 | Accept/Reject notification. |
| 11:00 - 12:00 | Poster session | 15:50 - 16:50 | Poster session | Nov 22 | Upload deadline for spotlight materials. |
| 12:00 - 13:00 | Lunch break | 16:50 - 17:50 | Panel discussion | | |
| 13:00 - 13:30 | Invited talk 3 | 17:50 - 18:00 | Closing remarks | Dec 6 - 7 | Workshop |

The workshop will follow the dates suggested by ICLR: **submission deadline** on 30 January 2026; **reviewer assignments** on 2 February 2026; **reviews due** 22 February 2026; **paper notification** on 1 March 2025. This schedule gives reviewers almost three weeks to process their assigned papers (up to three) and some time to organizers to address potential delays in the process. All dates are AoE.

**Outreach and Materials** We will advertise the workshop on social networks – Twitter/X, BlueSky, LinkedIn, etc. – and on relevant mailing lists (e.g., the Machine Learning mailing list, the Queer in AI group, the Black in AI group) as soon as we receive the acceptance notification. All accepted papers will be hosted on OpenReview and all recordings will be uploaded to an online video repository (likely YouTube, for visibility) and linked to the workshop's website.

**Invited Speakers.** We invited the following 6 speakers:

- **Qiaochu Chen** (**confirmed**) is an Assistant Professor at the University of Alberta. Her research bridges programming languages, formal methods, and natural language processing to develop neurosymbolic programming languages and synthesis algorithms. She has created domain-specific languages that blend symbolic pattern matching with neural semantic understanding for data analytics. To make these languages practical, she develops synthesis techniques that learn programs from multi-modal specifications using neural-guided search and quantitative synthesis methods. Her research has been published in top venues including PLDI, OOPSLA, NeurIPS, and ACL. Chen is a recipient of the EECS Rising Stars award and received her PhD from the University of Texas at Austin.
- **Biwei Hwang** (**confirmed**) is an Assistant Professor at Halicioğlu Data Science Institute, UC San Diego. Her research interests include Causal Discovery and Inference, Causality-Empowered ML/AI and Foundation Models, and Computational Science. On the causality side, Biwei's research has delivered more reliable and practical causal discovery algorithms by considering many of the challenges to conventional statistical inference, including distribution shifts, selection bias, and latent confounders. On the ML side, her work has shown that the causal view provides a clear picture for understanding advanced learning problems and allows going beyond the data in a principled, interpretable manner.
- **Mateja Jamnik** (**confirmed**) is a Full Professor of AI at University of Cambridge and an Associate Fellow at the Leverhulme Centre for the Future of Intelligence. Her research focuses on human-like computing, combining reasoning and ML to improve explainability in AI, with applications in personalised medicine and tutoring. Key interests include symbolic-neural integration, multi-modal learning, knowledge representation, and cognitive AI. She founded women@CL and advised the UK House of Lords on AI.
- **Subbarao Kambhampati** (**confirmed**) is a professor of computer science at Arizona State University. Kambhampati studies fundamental problems in planning and decision making, motivated in particular by the challenges of human-aware AI systems. He is a fellow of AAAI, AAAS, and ACM and a recent recipient of the AAAI Patrick H. Winston Outstanding Educator award. He served as the president of the Association for the Advancement of Artificial Intelligence, a trustee of the International Joint Conference on Artificial Intelligence, the chair of AAAS Section T (Information, Communication and Computation), and a founding board member of Partnership on AI. His research as well as his views on the progress and societal impacts of AI have been featured in multiple national and international media outlets.

- **Been Kim** (**confirmed**) is Senior Staff Research Scientist at Google DeepMind. Her work is focused on improving interaction between humans and machines. She has been a General Chair for ICLR 2024, given keynotes at ICLR 2022, ECML 2020 and at the G20 meeting in Argentina in 2018. Her work has been featured in Quanta Magazine and she has given multiple tutorials and invited talks at leading AI research venues such as NeurIPS, ICML and ICLR. She is also on multiple steering committees, executive boards and has organized multiple workshops in top research venues.
- **Bernhard Schölkopf** (**tentative**) is a Director at the Max Planck Institute for Intelligent Systems, Tübingen, a Professor at ETH Zurich, an Honorary Professor at TU Berlin, a co-founder of ELLIS, a Fellow of the ACM and CIFAR, and a member of the German Academy of Sciences. His research interests span various topics in ML and causality, including CRL. He was general and program chair for the First Conference on Causal Learning and Reasoning (CLeaR 2022), a member of the advisory board and a co-founder of the CLeaR society. He has organized various conferences and workshops, including the CRL Tübingen workshop in 2023 and the NeurIPS 2021 Causality workshop.

### 1.1 Promoting Diversity

We have made a concerted effort to ensure diversity among organizers, speakers, and participants of the workshop. Our organizing team comprises individuals from diverse countries, institutions, career stages and backgrounds. The keynote speakers were chosen to represent the three different communities involved, XAI, NeSY and CRL, across both academia and industry, and from multiple countries and institutions. Our goal is to provide a broad range of insights and experiences, reflecting diverse professional and cultural backgrounds. The diversity among our organizers and speakers fosters a welcoming and inclusive environment, signaling that different viewpoints and communication styles are valued and encouraging participation from a broad spectrum of attendees from different fields and with diverse backgrounds. Our workshop is inherently multidisciplinary and this will be reflected by our call for contributions. This is meant to ensure representation across all the three communities we intend to bring together, drawing in participants with diverse interests and perspectives.

## 2 Organizers

**Organization Team** Our team is highly multidisciplinary, and individual members complement each other's expertise, making it ideal for evaluating contributions on all topics covered by the workshop (XAI, NeSy, and Causality), as shown by the following summary:

| Name | Affiliation(s) | Position | Topics |
|---|---|---|---|
| Amit Dhurandhar ( ✉, 📚 ) | IBM Research | Principal Research Scientist | XAI |
| Amir-Hossein Karimi ( ✉, 📚 ) | U Waterloo | Assistant Professor | Causality, XAI |
| Sara Magliacane ( ✉, 📚 ) | U Amsterdam | Assistant Professor | Causality |
| Stefano Teso ( ✉, 📚 ) | U Trento | Associate Professor | NeSy, XAI |
| Efthymia Tsamoura ( ✉, 📚 ) | Huawei Labs | Technical Expert | NeSy |
| Zhe Zeng ( ✉, 📚 ) | U Virginia | Assistant Professor | NeSy |

Collectively, our team has extensive organizational experience: we have co-organized numerous events, including international conferences (CLeaR, UAI), workshops (NeurIPS × 4, ICML × 2, CVPR × 1, AAAI × 1, UAI × 5, KDD × 2, AAMAS × 1, KR × 1, Industrial - Samsung × 1), tutorials (AAAI × 1, IJCAI × 1, ECML × 1), seminars (Dagstuhl × 1), and panels (ISWC × 1). This is however our first direct collaboration and it will offer junior members valuable mentorship and hands-on involvement. None of the organizers are submitting any other workshop proposal for NeurIPS.

**Amit Dhurandhar**, is a Principal Research Scientist at IBM TJ Watson NY USA. His recent work was featured in Forbes and PC magazine with corresponding technical contribution in leading research venues such as Science, Nature, NeurIPS, ICML, ICLR. His research also has received the AAAI deployed application award, AAAI HCOMP best paper honorable mention as well as being selected as Best of ICDM twice. He has provided multiple invited talks including a industry keynote at ACM CODS-COMAD 2021. He also Co-led the creation of the AI Explainability 360 open source toolkit. He has been an Area Chair and SPC member for top AI conferences as well as has served on National Science Foundation (NSF) panels for the small business innovative research (SBIR) program. He also serves on the invention disclosure committee (IDT) in IBM Research, is an honored listee on Marquis Who's Who 2024 and has received the Distinguished Alumni Award for Career Achievement from the University of Florida 2025.

**Amir-Hossein Karimi**, is an Assistant Professor at the University of Waterloo and a Vector Institute Faculty Affiliate. He leads the Collaborative Human-AI Reasoning Machines (CHARM) Lab, focusing on AI safety, ethics, causal inference, and explainable AI. His work, which has over 3,000 citations, has been presented as oral and spotlight presentations at top venues like NeurIPS, ICML, AAAI, AISTATS, and ACM-FAccT. He has co-organized multiple ICML workshops, including the Workshop on Algorithmic Recourse (2021) and the Workshop on Counterfactuals in Minds & Machines (2023), and has been a review committee member for ICML workshops (2024-5).

**Sara Magliacane** is an Assistant Professor at the Informatics Institute at the University of Amsterdam and an ELLIS Scholar. Her research is on causal representation learning (CRL), causal discovery, applications of causality to ML, especially domain adaptation and RL. She organized the CRL workshops at NeurIPS 2024, 2023 and UAI 2022, the Causality in Time Series workshop at UAI 2023, A causal view on Dynamical Systems at NeurIPS 2022, and the Causality workshops at UAI 2024, UAI 2021 and NeurIPS 2020. She has co-organized bigger events, e.g. the CLeaR and UAI conferences, as a sponsor chair (CLeaR 2022), publication chair (UAI 2022), communication chair (CLeaR 2023), online chair (UAI 2023) and as program chair (UAI 2025).

**Stefano Teso** is an Associate Professor at the University of Trento and ELLIS Member. Stefano's research focuses on concept-based Explainable AI and explainable-by-design models, Neuro-Symbolic AI, and human-in-the-loop Machine Learning. Stefano co-organized tutorials on constraint learning at AAAI'18, IJCAI'18, ECML-PKDD'19, and at the Reasoning Web Summer School '19; a tutorial on explanations in interactive ML at AAAI'22; a workshop on interactive ML at the same venue; and a Dagstuhl Seminar on Explainable AI in 2025. He is a regular PC/SPC/AC at international conferences (NeurIPS, ICML, ICLR, AAAI, IJCAI, UAI, AISTATS, ECML-PKDD) and journals (MLJ, AIJ, TMLR).

**Efthymia (Efi) Tsamoura** is a Technical Expert at Huawei Labs. From 2019 to 2025, Efi was a Senior Researcher at Samsung AI, Cambridge, UK. In 2016, she was awarded a prestigious early career fellowship from the Alan Turing Institute, UK, for her work on logic and databases and, before that, she was a Postdoctoral Researcher in the Department of Computer Science of the University of Oxford. Her main research interests lie in the areas of logic, knowledge representation and reasoning, and neurosymbolic learning, while her recent outcomes involve scaling symbolic reasoning to billions of triples, as well as addressing open problems in neurosymbolic learning. In 2024, Efi was invited by the Royal Society, UK, at the Frontiers of Science on AI meeting to discuss the risks of AI and ways to address them. Efi co-organized multiple workshops on neurosymbolic learning including the 1st and the 2nd Samsung AI Neurosymbolic Workshop, the 3rd and the 4th Knowledge-Infused Learning Workshop (co-located with KDD 2023 and KDD 2024), the Neurosymbolic AI for Agent and Multi-Agent Systems Workshop (co-located with AAMAS 2023), the Knowledge Representation for Hybrid and Compositional AI Workshop (co-located with KR 2021), and the Neurosymbolic AI Panel in the 22nd International Semantic Web Conference in 2023.

**Zhe Zeng**, Tenure-track Assistant Professor at the University of Virginia, USA. Her research focuses on neurosymbolic AI and probabilistic machine learning to achieve trustworthy AI and aid scientific discoveries. She organized the Ensuring Trustworthiness in Multi-Modal Open-World Intelligence workshop at CVPR 2025 and the eighth workshop on Tractable Probabilistic Modeling at UAI 2025. She is a regular PC/SPC at international conferences (NeurIPS, ICML, AAAI, IJCAI, UAI, AISTATS).

**Program Committee** The organizers will be acting as area chairs and emergency reviewers. We will carefully assess conflicts of interest, also in terms of the organizing team. Our diversity in terms of affiliation should allow us easily to avoid area chairing for papers of the same organization. Given the multidisciplinary nature of the workshop, we plan to have 60+ reviewers with different backgrounds that will each review 2–3 submissions. To ensure a thorough and timely review process: 1) Each submission will be assigned to 3 reviewers. 2) No reviewer will be assigned more than 3 papers to review. 3) We have already confirmed a group of Program Committee (PC) members, marked in **bold** below along with their expertise (NeSy: Ⓝ, XAI: Ⓧ, Causality: Ⓒ): Ricardo Dominguez-Olmedo (Max Planck, ⒸⓍ); Julius von Kügelgen (ETH Zürich, Ⓒ); Krikamol Muandet (Helmholtz Center, ⒸⓍ); Matej Zečević (TU Darmstadt, ⓃⓍⒸ); Adrian Javaloy (U Edinburgh, Ⓒ); **Emanuele Marconato** (U Trento, ⓃⓍⒸ); **Samuele Bortolotti** (U Trento, Ⓝ); **Andrea Passerini** (U Trento, ⓃⓍ); Devendra Singh Dhami (TU Eindhoven, ⓃⒸ); **Wolfgang Stammer** (TU Darmstadt, ⓃⓍ); **Antonio Vergari** (U Edinburgh, Ⓝ); Luigi Gresele (U Copenhagen, Ⓒ); **Francesco Giannini** (Scuola Normale Superiore Pisa, ⓃⓍ); Guy Van den Broeck (UCLA, Ⓝ); **Luciano Serafini** (Fondazione Bruno Kessler, Ⓝ); **Pedro Zuidberg Dos Martires** (Örebro U, Ⓝ); Giuseppe Marra (KU Leuven, ⓃⓍ); Victor Gutierrez Basulto (Cardiff U, Ⓝ); Vaishak Belle (U Edinburgh, ⓃⓍⒸ); Gianluca Cima (Sapienza U, Ⓝ); Floris Geerts (U Antwerp, Ⓝ); Christoph Haase (U Oxford, Ⓝ); Ziyang Li (U Pennsylvania, Ⓝ); **Riccardo Massidda** (U Pisa, Ⓒ); **Danru Xu** (U Amsterdam, Ⓒ); **Matyas Schubert** (U Amsterdam, Ⓒ); **Roel Hulsman** (U Amsterdam, Ⓒ); **Hidde Fokkema** (U Amsterdam, Ⓧ); **Fan Feng** (UCSD, Ⓒ); **Nadja Rutsch** (VU Amsterdam, Ⓒ); **Phillip Lippe** (Google, Ⓒ); **Tim van Erven** (U Amsterdam, Ⓧ); **Jacopo Dapueto** (U Genova, Ⓒ); Sebastien Lachapelle (Samsung, Ⓒ); **Christine Bang** (U Copenhagen, Ⓒ); Gemma Moran (Rutgers, Ⓒ); Jiani Huang (UPENN, Ⓝ); YooJung Choi (ASU, Ⓝ); Kareem Ahmed (UCI, Ⓝ); **Meihua Dang** (Stanford, Ⓝ); **Honghua Zhang** (XAI, Ⓝ); **Anji Liu** (NUS, Ⓝ); **Poorva Garg** (UCLA, Ⓝ); **Benjie Wang** (UCLA, ⒸⓍⓃ); **Daniel Israel** (UCLA, ⒸⓃ); **Chi Zhang** (Toyota Research Institute, Ⓒ); **Zilei Shao** (UCLA, Ⓝ) **Yidou Weng** (UCLA, Ⓝ) **Eleonora Poeta** (Politecnico Torino, Ⓧ); Eleonora Giunchiglia (Imperial College London, Ⓝ); Sonia Laguna (ETH Zurich, Ⓧ); Yoshihide Sawada (AISIN Corp, Ⓧ); **Gesina Schwalbe** (U Lübeck, Ⓧ).

## References

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *AISTATS*, 2024.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for ai safety-a review. *Transactions on Machine Learning Research*, 2025.

Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 2022.

Luc De Raedt, Sebastijan Dumančić, Robin Manhaeve, and Giuseppe Marra. From statistical relational to neural-symbolic artificial intelligence. In *IJCAI*, 2021.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*, 2018a.

Amit Dhurandhar, Karthikeyan Shanmugam, Ronny Luss, and Peder Olsen. Improving simple models with confidence profiles. *NeurIPS*, 2018b.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *NeurIPS*, 2022.

Hidde Fokkema, Tim van Erven, and Sara Magliacane. Sample-efficient learning of concepts with theoretical guarantees: from data to concepts without interventions. *arXiv preprint arXiv:2502.06536*, 2025.

Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Luis C Lamb, Leo de Penning, BV Illuminoo, Hoifung Poon, and COPPE Gerson Zaverucha. Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 2022.

Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. In *IJCAI*, 2022.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *NeurIPS*, 35:5549–5561, 2022.

Subbarao Kambhampati, Sarath Sreedharan, Mudit Verma, Yantian Zha, and Lin Guan. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. In *AAAI*, 2022.

Amir-Hossein Karimi, Krikamol Muandet, Simon Kornblith, Bernhard Schölkopf, and Been Kim. On the relationship between explanation and prediction: A causal view. *arXiv preprint arXiv:2212.06925*, 2022.

Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now You See Me (CME): Concept-based Model Extraction. *arXiv preprint arXiv:2010.13233*, 2020.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020.

Lingjing Kong, Guangyi Chen, Biwei Huang, Eric Xing, Yuejie Chi, and Kun Zhang. Learning discrete concepts in latent hierarchical models. *NeurIPS*, 2024.

Sebastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi LE PRIOL, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *First Conference on Causal Learning and Reasoning*, 2022.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal Identifiability from Temporal Intervened Sequences. In *ICML Spotlight*, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Biscuit: Causal representation learning from binary interactions. In *UAI*, 2023.

Yuhang Liu, Dong Gong, Erdun Gao, Zhen Zhang, Biwei Huang, Mingming Gong, Anton van den Hengel, and Javen Qinfeng Shi. I predict therefore i am: Is next token prediction enough to learn human-interpretable concepts from data? *arXiv preprint arXiv:2503.08980*, 2025.

Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. Leveraging latent features for local explanations. In *KDD*, 2021.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. *Entropy*, 2023a.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. *NeurIPS*, 2023b.

Andrei Margeloiu et al. Do concept bottleneck models learn as intended? *arXiv:2105.04289*, 2021.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*, 2023.

Isha Puri, Amit Dhurandhar, Tejaswini Pedapati, Karthikeyan Shanmugam, Dennis Wei, and Kush R Varshney. Cofrnets: Interpretable neural architecture inspired by continued fractions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *NeurIPS*, 2021.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. From causal to concept-based representation learning. *NeurIPS*, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Stefano Teso, Öznur Alkan, Wolfang Stammer, and Elizabeth Daly. Leveraging explanations in interactive machine learning: An overview. *Frontiers in Artificial Intelligence*, 2023.

Efthymia Tsamoura, Timothy Hospedales, and Loizos Michael. Neural-symbolic integration: A compositional perspective. In *AAAI*, 2021.

Kaifu Wang, Efthymia Tsamoura, and Dan Roth. On characterizing and mitigating imbalances in multi-instance partial label learning. *arXiv preprint arXiv:2407.10000*, 2024.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *ICML*, 2024.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *ICML*, 2024.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *NeurIPS*, 35, 2022.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *ICML*, 2024.