Causality in Flux: Continual Adaptation of Causal Knowledge via Evidence Matching

Jonas Seng[1,*] JONAS.SENG@TU-DARMSTADT.DE
Florian Peter Busch[1,2,*] FLORIAN_PETER.BUSCH@TU-DARMSTADT.DE
Kristian Kersting[1,2,3] KERSTING@CS.TU-DARMSTADT.DE
[1] *TU Darmstadt, 64289 Darmstadt, Karolinenplatz 5*
[2] *Hessian Center for AI (hessian.AI)*
[3] *German Research Center for AI (DFKI)*

Abstract

Utilising causal knowledge in machine learning (ML) systems yields more robust models with the capability of performing certain extrapolations. However, much of current causality research focuses on deriving causal models in isolation, hence current systems are not capable of updating and improving causal knowledge when new observations arrive. Drawing inspiration from human learning, Continual Learning (CL) aims at updating models given a sequential stream of evidence. Leveraging common patterns and past experiences to gradually improve causal knowledge in ML models is a crucial step towards more robust CL systems. In this work, we propose to learn and update causal models in a lifelong learning setting where causal knowledge explaining newly arriving observations is inferred from similar previously seen observations. We call this framework *evidence matching*. Further, an analysis of real world data supporting our motivation is provided.

Keywords: Causality, Causal Discovery, Continual Learning

1. Motivation

The world provides humans with sequences of new observations continuously. In response, humans repeatedly adapt and extend their world model based on these evidences, enabling them to perform accurate predictions. At the same time, a pivotal component of human intelligence is the incorporation of causal knowledge enhancing the robustness and accuracy of predictions across environments. Frequently, human cognition elucidates novel observations through the adaptation of pre-existing causal explanations derived from analogous past experiences. For example, upon achieving a satisfactory explanation for the motion of a car, humans effortlessly formulate a subtly adjusted rationale for the movement of an airplane. Hence, two key ingredients required for a robust, flexible and accurate learning system are the capability of continuously updating a model and the usage of causal knowledge. Continual Learning (CL) aims to adapt the capability of continuously updating models to Machine Learning (ML) (Mundt et al., 2020; Parisi et al., 2019; McCloskey and Cohen, 1989). Nonetheless, the predominant focus in CL revolves around traditional Machine Learning (ML) models known for their reliance on correlational information exclusively. It is widely accepted that incorporating causal capabilities (Pearl, 2009) into ML

[*] Equal contribution

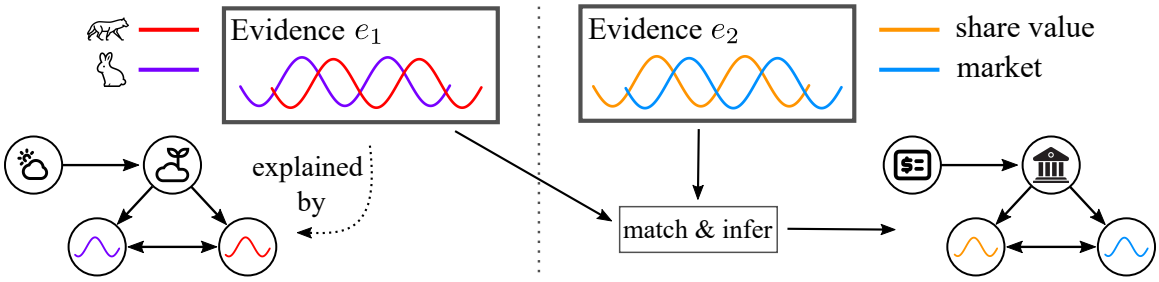

Figure 1: **Infer Causal Structure using former Evidence.** If the causal structure of evidence obtained formerly (e_1 in this case) matches the current evidence e_2, then previous causal knowledge for e_1 can help infer the causal structure of e_2.

models is helpful in constructing more robust ML systems and progressing towards systems with certain extrapolation capabilities. To this end, the ML community has begun exploring how causal knowledge can be incorporated into ML models, enabling them to perform inference for causal questions (Xia et al., 2021; Zečević et al., 2021). However, as of now, this exploration has been predominantly limited to classical supervised environments which lack continuous streams of evidence. We believe that one key component still missing in the design of CL systems incorporating causal knowledge in their predictions is the ability to guide the adaptation process of causal models by leveraging patterns manifested in observed evidences – termed *evidence matching*. The following example gives an intuition:

Example. Examine the renowned Lotka-Volterra rule (Lotka, 1920), asserting that prey and predator population sizes exhibit a cyclical pattern with a slight phase shift. Initially, prey populations peak, followed by a decline as predator populations rise and eventually reach their peak. Subsequently, with a decrease in the predator population due to a scarcity of prey, the prey population gradually recovers.

A similar dependency can be observed on financial markets: During economic growth, stock prices are rising as well as the volume of stocks being held by shareholders. At some point, the volume peaks and due to certain (often unmeasured) factors, shareholders start selling their shares, letting stock prices fall. When economy recovers, shareholders start to buy shares again, initiating a new cycle.

Fig. 1 visualizes our idea by highlighting how similar causal structures can render similar patterns in observations. By exploiting previous evidence e_1 for which the causal structure was already identified, one can infer the causal relations explaining current evidence e_2 efficiently by making use of the similarity of the evidences.

We suspect that exploiting such patterns can make adaptation faster, more efficient and more robust. Consequently, in this work we make the following contributions:

1. We present a meta-learning inspired framework exploiting patterns in the evidences observed to guide the adaptation of causal models.

2. We introduce evidence matching along with theoretical assumptions and describe how it relates to the independent mechanisms principle.

3. Based on the Tübingen cause-effect pairs dataset (Mooij et al., 2016), we show empirical evidence that patterns in observed data exist and can be used to reason about the underlying causal structure.[1]

2. Related Work

Causal Discovery. Structural Causal Models (SCMs) define a principled framework to model complex joint distributions, predict the effects of interventions and perform counterfactual reasoning (Pearl, 2009). The causal graph induced by an SCM dictates the causal directions between variables. Recovering this graph from data is a fundamental task in various disciplines which has been tackled by various approaches, termed causal discovery (Sprites et al., 2000; Chickering, 2003; Shimizu et al., 2006; Zanga and Stella, 2023). Ke et al. (2022) recently proposed a supervised learning task to identify relations between synthetic datasets and the structure of Causal Bayesian Networks (CBNs) leveraging observational and interventional data. Lopez-Paz et al. (2015) predict causal directions from data leveraging a model trained on a set of causal discovery tasks. However, they only consider the standard setting (no CL) and thus lack capabilities to deal with continual updates and changes. The Tübingen causal-pairs dataset (Mooij et al., 2016) comprises real-world datasets featuring pairs of variables, each accompanied by a ground truth causal direction.

Continual Learning. Continual learning (CL) in general and lifelong learning (Chen and Liu, 2018) in particular are concerned with learning problems where new information is made available over time. Typically, CL methods rely on regularization (Kirkpatrick et al., 2017), (pseudo-)rehearsal (Shin et al., 2017), or architecture modifications (Rusu et al., 2022) to address this challenge. Mundt et al. (2020) highlight the connections between open-world learning, active learning and continual learning, emphasizing the limitations of the closed world assumption. This aligns with our motivation of leveraging evidence across different domains. Javed and White (2019) propose to meta-learn an expressive representation for all tasks in a sequence of tasks and train specialized neural network heads to solve a certain task. Meta-learning the representation helps exploiting similarities among tasks, hence allowing for quick adaptation of classification heads.

3. Guiding Causal Model Adaptation via Evidence Matching

Evidence matching aims to compare representations of former evidence for which causal knowledge is available to infer causal knowledge of new evidence by exploiting similarities among observations. To that end, we consider the set of all possible evidences $\mathcal{E}$ and assume that causal knowledge is represented as a graph structure.

3.1. Evidence Matching and Independent Mechanisms

We start by revisiting Structural Causal Models (SCMs).

Definition 1 (Structural Causal Model) *A Structural Causal Model (SCM) is a tuple* $\xi = (\mathbf{U}, \mathbf{V}, \mathcal{F}, p)$ *where* $\mathbf{U}$ *is a set of exogenous (i.e. unmodelled) noise variables,* $\mathbf{V}$ *is the*

1. Our code is available at https://github.com/J0nasSeng/causal-structure-meta-learning.git.

set of endogenous (i.e. observed) variables, and p is a joint probability distribution over $\mathbf{U}$ s.t. all $U_i, U_j \in \mathbf{U}$ are independent and $\mathcal{F}$ is a set of functions describing relations between variables in $\mathbf{V}$

$$V_j = f_j(\mathbf{Pa}(V_j), U_j) \tag{1}$$

where $V_j \in \mathbf{V}$ is an endogenous variable, $\mathbf{Pa}(V_j)$ describes the set of parents of V_j, and U_j is exogenous noise.

Each SCM induces a a Directed Acyclic Graph (DAG) G. Assume dataset $\mathcal{D} = \{(\mathbf{X}_1, G_1), \ldots, (\mathbf{X}_n, G_n)\}$ where $\mathbf{X}_i = (\mathbf{x}_1^i, \ldots \mathbf{x}_k^i) \in \mathcal{E}$ is a set of k i.i.d. observations with dimension d, where each observation is sampled from distribution p_i induced by the causal graph G_i. Considering $\mathcal{D}$ as a set of tasks, learning a function $g_\theta : \mathcal{E} \to \mathcal{G}$ can be viewed as a meta-learning problem where each $(\mathbf{X}_i, G_i)$ corresponds to a task. Each task comes with observations $\mathbf{X}_i$ and a model G_i explaining $\mathbf{X}_i$. Hence, learning g_θ can be seen as learning to learn a model G_i given observations $\mathbf{X}_i$. Inspired by this meta-learning perspective, we propose to minimize a loss $\mathcal{L}$ of the following form:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(\mathbf{X}_i, G_i) \sim \mathcal{D}}[\mathcal{L}(g_\theta(\mathbf{X}_i), G_i)] \tag{2}$$

It is well known that recovering the full causal structure from purely observational data is impossible in general (Pearl, 2009), hence Eq. 2 cannot be solved optimally using standard ML methods. However, it is possible to recover the causal structure up to Markov Equivalence, i.e. to identify the correct independence structure among variables. We introduce the following assumptions to allow for evidence matching:

Assumption 1 (Availability of Evidence) *For each incoming $(\mathbf{X}, G)$ there exist $(\mathbf{X}', G') \in \mathcal{D}$ s.t. $G \equiv_\mathcal{M} G'$ where $\equiv_\mathcal{M}$ means Markov equivalent.*

Assumption 2 (Clusters of Markov Equivalent Graphs) *For $\mathcal{D}' \subset \mathcal{D}$ s.t. $G_i \equiv_\mathcal{M} G_j \, \forall G_i, G_j \in \mathcal{D}', i \neq j$ and $\mathcal{D}'' = \mathcal{D} \setminus \mathcal{D}'$, assume*

$$\sigma < \frac{\max_{G_i, G_j \in \mathcal{D}'} D_{KL}(p(V(G_i)) || p(V(G_j)))}{\min_{G_i \in \mathcal{D}', G_j \in \mathcal{D}''} D_{KL}(p(V(G_i)) || p(V(G_j)))}$$

where D_{KL} is the KL-divergence, σ is a constant and $p(V(G_i))$ denotes the distribution over random variables $V(G_i)$ the graph G_i is defined over.

Assumption 2 ensures that Markov equivalent graphs can be clustered w.r.t. their induced distributions and thus w.r.t. samples we observe which allows matching of observations. The ultimate goal is to learn a function g_{θ^*} inferring a graph close to the true causal DAG G that induced the distribution $p(V(G))$ from which the observed samples $\mathbf{X}_i$ come, by matching $\mathbf{X}_i$ to former observations. This leads to the following conjecture.

Conjecture 1 *Recovering graphs beyond the MEC via evidence matching is impossible without a relaxation of the independent mechanisms principle.*

Algorithm 1: Evidence Matching Framework. Our framework relies on an encoding function computing representations of observations and a function to match observations and retrieve the closest graph.

Data: Representation storage $\mathbf{R}$, observations $\mathbf{X}_i$
$\mathbf{r} \leftarrow$ encode_observations($\mathbf{X}_i$) ;
if $\mathbf{R}$ *is not empty* **then**
  $G_{local} \leftarrow$ match_and_infer($\mathbf{r}$, $\mathbf{R}$) ;
  $G_{local} \leftarrow$ refine_local_graph(G_{local}, $\mathbf{X}_i$);
end
else
  $G_{local} \leftarrow$ identify_causal_graph($\mathbf{X}_i$);
end
store $(\mathbf{r}, G_{local})$ in $\mathbf{R}$;
return G_{local} ;

The motivation of this paper is based on the observation that similar mechanisms can co-occur more frequently with similar graph structures (see Fig. 1). To identify a causal structures beyond the MEC via evidence matching, we believe that – given a shared causal structure – the probability of a set of functions explaining a formerly seen evidence must influence the probability of the set of functions explaining observations seen later. Hence, mechanisms might not be independent of each other. We argue that this aligns with the observation that humans tend to quickly infer causal relationships from evidences without explicitly testing the causality of these relationships. This strategy allows for quick identification of causal structures in various scenarios. Adopting such a strategy to adjust causal models in CL systems can, therefore, be highly beneficial.

In Alg. 1 we propose a possible framework to update causal knowledge in a CL setting using matching of sequentially obtained evidence. Note that we assume that causal knowledge is represented by a graph structure, however our framework is not limited to this representation. Other, similar algorithms for evidence matching are also conceivable.

4. Exploiting Evidence Matching

Evidence matching holds promise for making identification of causal structures easier and more efficient. An intriguing approach to leverage former evidence could involve learning representations of entire datasets and making predictions based on these representations. While such an approach eliminates the need for hand-crafted features, ideally the learned representations provide high quality estimates of sufficient statistics of obtained observations. There are multiple ways to use such a representation for causal discovery. In its most simple form, the model can directly output a causal graph. But the model output could also be less specific, instead simply constraining the search space over causal graphs, hence warm starting subsequent causal discovery algorithms. If available, additional sources of information, such as details about the datasets' environments or information about variable names could further provide beneficial information. Here, the application of large language models to this additional information (e.g. Kıcıman et al. (2023)) holds promise.

However, note that we acknowledge that there is no guarantee in general that similar data implies similar causal structures, as generally two different causal graphs can generate the same observational data. Hence, providing uncertainty estimates for predictions should be part of the design of such systems. Also, in cases where certainty in identifying the correct causal directions is essential, evidence matching can prove valuable for selecting interventions, ultimately helping to identify causal structures. This is because certain causal directions are more probable than others, resulting in a minimal set of interventions to be performed to identify causal structures. Furthermore, the availability of a larger base of datasets could not only assist in testing assumptions commonly used in causal discovery algorithms but also facilitate the exploration of new criteria and assumptions, potentially leading to the development of novel algorithms. It could also serve as a basis of determining the conditions under which assumptions are applicable.

For evidence matching, the CL setting is of particular importance as the entire concept is built around continually increasing the accumulated knowledge about causal patterns for future predictions. Here, models trained on different environments or using different data might utilize different patterns for evidence matching. For instance, if a model is initially trained on datasets capturing physical processes and subsequently training is continued with datasets related to economy, the model could fail to perform well on physical datasets if no measures are taken to prevent this. But even within a specific setting, unmeasured influences (variables) can change over time, thus resulting in new kinds of datasets with previously unseen properties. In such a setting of "Causality in Flux", i.e. where patterns to detect causal directions may change over time, avoiding forgetting is desirable. In particular, learning in a continual manner is important as, in this setting, a single model is trained on *datasets of datasets*. Remember, that each dataset is used to output a single causal graph. In a sense, a full dataset is just a single variable for evidence matching. When training using this kind and amount of data, storing all data trained on previously requires a lot of capacity. Even if all previous data is available, training on such a huge amount of data can be time consuming and computationally expensive. Thus, updating a model in a CL setting follows naturally. If a model for evidence matching is only applied in a specific environment such as economy, another difficult task is to decide when new data contradicting current model behavior is representative for the problem in the future and should be learned. If, otherwise, this new data is determined to be outliers, it may be better to exclude them from training or leave open the possibility to revert to previous model behavior.

On the related issue of performing targeted forgetting of information learned, recent advancements have been made (Wang et al., 2021), (Wang et al., 2023). One primary motivation for this approach is to balance model capacity with predictive performance. The goal is to keep as much free model capacity as possible for upcoming tasks, i.e. to remove information with no or little relevance from the model. In the context of causal models in CL scenarios, abstractions are a promising alternative to forgetting. Finding abstractions corresponds to summarizing multiple variables of a causal graph into one variable and adapting the mechanisms between the summarized variable and its neighbors. This way, a trade-off between model complexity and predictive performance can be achieved, mirroring the principles proposed in recent works on CL. Evidence matching could help identify potential abstractions, thus assisting in finding a balance between model complexity and predictive performance.

Experiment Results					Average
0.83	0.78	0.67	0.78	0.87	0.79 ±0.07

Table 2: **Evidence Matching using Random Forests.** Given the meta-features, we trained a RF 5 times on different train-test splits. It can be seen that the RF performs better than random (baseline is 0.5 accuracy due to our data augmentation) consistently, indicating that there is a dependency between the meta-features and the true causal model.

5. Evidence Matching in Cause-Effect-Pairs

Successful evidence matching relies on recognizing patterns in the data that are dependent on the causal structure. To gauge the practical feasibility, we analyse the Tübingen Cause-Effect-Pairs dataset from Mooij et al. (2016).

5.1. Experimental Setup

The Cause-Effect-Pairs dataset consists of 108 cause-effect pairs consisting of two variables where a causal connection between these variables is known (i.e. either $X \to Y$ or $Y \to X$). Each dataset contains between 109 and 16k i.i.d. samples from the underlying data generating process. We excluded all datasets which contained erroneous values (such as *null*-values), leaving us with 85 datasets.

Name	Formular	Rationale
Skewness	$\frac{\mathbb{E}[X-\mu_X]^3}{\sigma_X^3}$	Normality
Kurtosis	$\frac{\mathbb{E}[X-\mu_X]^4}{\sigma_X^4}$	Normality
Mutual Information	$MI(X_1, X_2)$	Entropy
Correlation	ρ_{X_1,X_2}	Dependency

Table 1: **Meta Features.** We derive four features from the the Cause-Effect-Pair datasets aiming to describe properties of the underlying joint distribution and its marginals s.t. the correct causal graph can be predicted based on these derived features.

Our analysis aimed to identify similarities among mechanisms in the data generating processes yielding the datasets in the Cause-Effect-Pairs collection. The mechanisms induce a joint distribution over all observed variables (in this case X and Y), hence our analysis focused on identifying similarities among these induced joint probabilities across datasets. To this end, we generated a set of meta-features that capture certain distribution-related properties of the data, see Tab. 1 for details. We include augmented datasets simulating the case had the columns been swapped[2], doubling the overall size of our data. The meta-features were used to train a classifier predicting the causal model (i.e. $X \to Y$ or $Y \to X$). We applied t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008) on the resulting meta-features to project the dataset on two dimensions. Tab. 1 shows all features derived from the datasets.

2. If the the original dataset D consists of instances such that $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ with class 0, we also add $D' = \{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$ with class 1.

5.2. Results

We now discuss the results of our analysis.

Classification. We trained a Random Forest (RF) on 5 randomly chosen subsets (70%) and tested it on the respective remaining data (30%). Our results demonstrate that the RF can distinguish between the causal models more effectively than random chance, indicating that the meta-features capture properties of the joint and marginal distributions, enabling the inference of the correct causal model. We show our results in Tab. 2.

Dimensionality Reduction. In Fig. 2, we present the results of the meta-features by t-SNE. We observe that $X \to Y$ and $Y \to X$ models form some clusters in the projected space. While not perfect, this shows that even in a 2-dimensional space, the data does not follow a completely random distribution. This indicates that the independent mechanism assumption does not hold in the Cause-Effect-Pairs datasets, providing support for our hypothesis that exploiting patterns among mechanisms is advantageous for inferring the underlying causal graph when presented with novel evidence.

Discussion. Considering the theoretical impossibility of discovering causal graphs solely from observational data, evidence matching can not always predict the correct causal graphs. As the meta-features here were chosen manually and are too simplistic to capture all relevant information about the data distributions, these results are not intended to compete within the causal discovery literature. Instead, they serve as an illustration and motivation for the idea that features extracted from causal data can be subsequently used to infer causal relations. While further experimentation and investigation are required to comprehensively assess the applicability of evidence matching, we believe it to be a promising concept with high potential, particularly in combination with other ideas or algorithms.

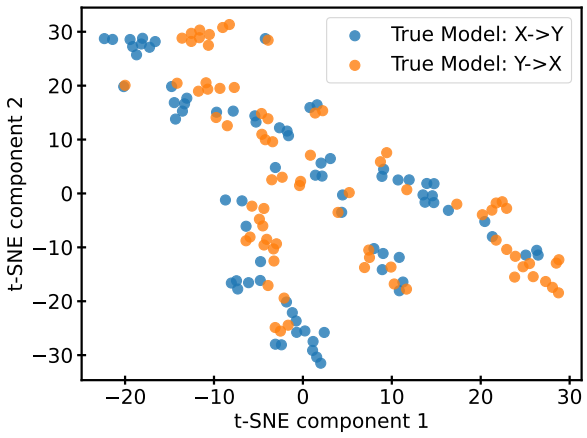

Figure 2: **Dimensionality Reduction of Meta Features (t-SNE).** We applied t-SNE on the meta-features and observe that there are certain clusters depending on the class ($X \to Y$ or $Y \to X$).

6. Conclusion & Outlook

In this work, we stated assumptions to allow for inference of causal relations by matching currently observed evidence with previous evidence. We also conjectured that the independent mechanism assumption must be weakened in order to facilitate such inference. Additionally, we provided empirical evidence with our analysis of the Tübingen Cause-Effect-Pairs datasets.

Further work should consider to build real world datasets where the underlying causal structure is known as such datasets are crucial to develop and evaluate methods following our proposed framework. This will enable a more detailed investigation into the capabilities and potential of evidence matching and the applicability of the independent mechanisms principle.

Acknowledgements

The work was supported by the Hessian Ministry of Higher Education Research, Science and the Arts (HMWK) via the DEPTH group CAUSE of the Hessian Center for AI (hessian.ai) and it has benefited from the HMWK cluster project "The Third Wave of AI". Further, this work benefited from the National High-Performance Computing project for Computational Engineering Sciences (NHR4CES).

References

Zhiyuan Chen and Bing Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. 3:507–554, 2003.

Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances in neural information processing systems*, 32, 2019.

Nan Rosemary Ke, Silvia Chiappa, Jane Wang, Anirudh Goyal, Jorg Bornschein, Melanie Rey, Theophane Weber, Matthew Botvinic, Michael Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure, 2022.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461. PMLR, 2015.

Alfred J. Lotka. Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences*, 6, 1920.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. 24:109–165, 1989.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Martin Mundt, Yong Won Hong, Iuliia Pliushch, and Visvanathan Ramesh. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *CoRR*, abs/2009.01797, 2020.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks, 2022.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7: 2003–2030, 2006.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay, 2017.

Peter Sprites, Glymour Clark, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 1 edition, 2000.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. Afec: Active forgetting of negative transfer in continual learning. *Advances in Neural Information Processing Systems*, 34:22379–22391, 2021.

Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning. *arXiv preprint arXiv:2307.09218*, 2023.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems*, volume 34, pages 10823–10836, 2021.

Alessio Zanga and Fabio Stella. A survey on causal discovery: Theory and practice, 2023.

Matej Zečević, Devendra Dhami, Athresh Karanam, Sriraam Natarajan, and Kristian Kersting. Interventional sum-product networks: Causal inference with tractable probabilistic models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15019–15031. Curran Associates, Inc., 2021.