Bank of England

Infusing economically motivated structure into machine learning methods

Staff Working Paper No. 1,144
September 2025

Marcus Buckmann and Galina Potjagailo

Bank of England

Staff Working Paper No. 1,144

Infusing economically motivated structure into machine learning methods

Marcus Buckmann[1] and Galina Potjagailo[2]

Abstract

This paper discusses how economic theory can be integrated into machine learning (ML) models to enhance their interpretability and applicability for policy analysis. While ML methods offer considerable flexibility and strong predictive performance, they are often criticised for their 'black box' nature and lack of economic transparency. A growing body of research addresses this limitation by introducing structure into ML models − most notably through Block-Additive Models (BAMs) and theory-consistent monotonicity constraints. BAMs group predictors into economically meaningful blocks and impose additivity across blocks, while permitting non-linearities and interactions within them. This architecture enables clear attribution of each block's contribution to the model's predictions. Monotonicity constraints further improve interpretability by aligning the model's directional responses with economic theory, allowing for the separation of opposing effects − such as distinguishing between supply and demand-driven components of inflation. Empirical evidence shows that these structured ML approaches retain strong predictive performance while yielding economically meaningful narratives.

Key words: Interpretable machine learning, theory-aligned constraints, macroeconomic analysis.

JEL classification: C10, C14, C53.

---

(1) Bank of England. Email: marcus.buckmann@bankofengland.co.uk
(2) Bank of England. Email: galina.potjagailo@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to Adrian Paul and Anthony Savagar. This paper will appear in the forthcoming book Central Banking, Monetary Policy, and Artificial Intelligence, edited by Marcos Centurion-Vicencio, Louis-Philippe Rochon and Guillaume Vallet.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email: enquiries@bankofengland.co.uk

ISSN 1749-9135 (on-line)

1 Introduction: How does machine learning fit into central banks' toolkits for economic analysis?

The succession of major economic shocks in recent years—including the COVID-19 pandemic, the Russia–Ukraine war and its associated energy disruptions, and the increasing relevance of climate risks—has exposed the limitations of conventional macroeconomic models. Such shocks can have highly non-linear effects, and might generate shifts in economic relationships that traditional parametric models would struggle to capture. In response, central banks and policy institutions are exploring new methodologies that can incorporate high-dimensional, non-standard data and flexibly account for structural instabilities.

Machine learning (ML) methods offer powerful tools for forecasting, as they can approximate complex, non-linear functional forms while processing a wide range of indicators. Common approaches include tree-based ensembles—such as random forests (Breiman, 2001), gradient boosting (Friedman, 2001), and Bayesian Additive Regression Trees (BART, Chipman et al., 2016)—as well as neural networks (Amari, 2006). Central banks are also increasingly adopting ML for broader purposes, including data collection, financial supervision, and the construction of new indicators using large language models (de Araujo et al., 2024; Cipollone, 2024; Benford, 2024).

However, the direct application of machine learning to economic policy analysis remains constrained by a fundamental challenge: the trade-off between predictive accuracy and economic interpretability. Although ML methods can extract signals from data with high precision, they typically lack the structural foundations necessary for policy use. Central banks must not only generate accurate forecasts but also ensure these forecasts are grounded in sound economic reasoning and can be clearly communicated to both policymakers and the public. To this end, forecasts are accompanied by narratives that elucidate the underlying transmission channels, aiming to shape and anchor agents' expectations around the central bank's inflation target and broader economic outlook. Economic interpretability is essential for explaining key drivers—such as distinguishing between demand-driven inflation and supply shocks—and for guiding appropriate policy responses. Standard ML approaches, when employed as black boxes, often fall short in this respect, as they neither impose economic structure nor facilitate the identification of underlying economic mechanisms.

One way to reconcile the flexibility of ML with the structural needs of economic analysis is to embed economic restrictions directly into the ML framework. Imposing constraints will not alter the fundamental nature of machine learning (ML) models as predictive, reduced-form tools. Nonetheless, recent advances in the literature suggest that ML models can be designed as to align with economic theory through simple structural constraints.

One approach to imposing constraints involves machine learning models with additive, blockwise structures, which we refer to as Blockwise Additive Models (BAMs). Rather than learning unconstrained interactions among all predictors, variables are grouped into blocks that represent theoretically meaningful economic drivers—for example, real variables, nominal variables, financial indicators, and expectations. The block structure reflects the modeller's intuition that groups of variables jointly capture an underlying economic driver, or at

the very least, exhibit similar predictive patterns that differ across blocks. This preserves interpretability while allowing flexible within-block non-linearities. We will examine recent advancements in machine learning approaches proposed in the literature for applications to inflation (Goulet Coulombe, 2024; Buckmann et al., 2025) and bond pricing (Bianchi et al., 2021).

Another approach we highlight involves the use of monotonicity constraints. To ensure economic interpretability, it may not be sufficient to simply separate blocks of determinants if these blocks do not uniquely identify the direction of association between a predictor and the variable of interest. To address this ambiguity, the modeller can impose directional constraints on the relationships learnt between predictors and the target variable. By enforcing prior economic knowledge on the associations learnt—such as the expectation that higher demand should place upward pressure on inflation—ML methods can produce results that are both statistically robust and economically plausible. Buckmann et al. (2025) pursue this route within a BAM based on gradient boosting, as we will discuss. ML models can also incorporate identifying information—such as shock series or instruments used as predictors—to further anchor model components to specific economic determinants. Furthermore, components of machine learning models can be constrained in their stochastic properties to distinguish between slow-moving and cyclical elements, analogous to approaches commonly employed in unobserved component models. Goulet Coulombe (2024) pursues this approach in a neural BAM.

While the literature we discuss here is still in its nascent tracks, it represents a first step in extending the applicability of ML methods from purely predictive tools into more interpretable frameworks that central banks can use to analyse non-linear determinants of economic variables such as inflation. There can exist a trade-off between imposing structure and maximising predictive performance, but imposing economic priors may also increase the predictive power, if the imposed constraints reduce the risk of overfitting spurious patterns in short time series. By embedding economic structure ex-ante, rather than relying solely on ex-post interpretability methods such as Shapley values, these models retain the flexibility of ML while improving their applicability to policy analysis.

Traditional methods have remained central in the toolkits of policy institutions, due to their balance of economic and statistical interpretability alongside solid forecasting performance. As illustrated in Figure 1, different approaches span a spectrum between purely predictive models and those grounded in economic theory. Vector autoregressions (VARs), unobserved components models (UCs), and dynamic factor models (DFMs) are reduced-form approaches that largely allow the data to speak for itself. In contrast, semi-structural and structural DSGE models embed explicit assumptions about microeconomic behavior. While reduced-form models often offer a better fit to the data, they are more agnostic in interpretation; augmenting them with identifying assumptions—as in structural VARs and structural DFMs—enhances their economic interpretability. These models are typically linear, although introducing non-linearities and regime shifts can improve flexibility. However, this additional flexibility increases the need for regularization through priors or restrictive assumptions on the degree of time variation, particularly to ensure stable estimates when

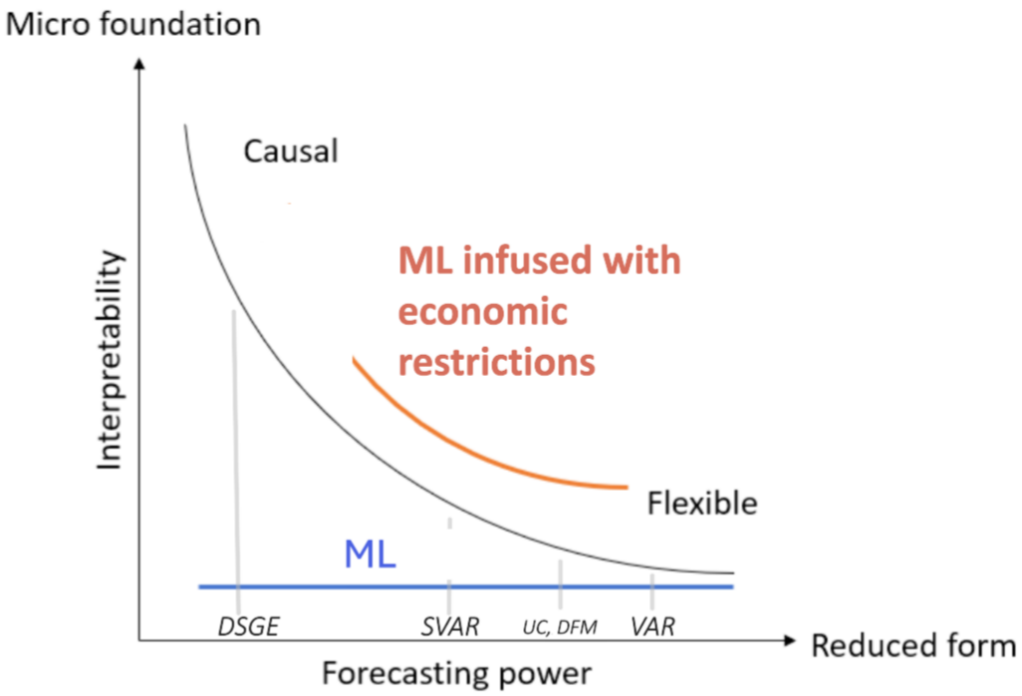

Figure 1: ML from purely predictive tools towards interpretability.
Source: Broadly based on an illustration by Luca Onorante, Workshop on Macroeconomic Analysis and Forecasting for Policy and Practice, University of Strathclyde, November 2024.

working with short samples.

Machine learning methods offer clear advantages in addressing greater complexity and flexibility in economic prediction problems (Döpke et al., 2017; Medeiros et al., 2021; Masini et al., 2023; Lenza et al., 2023; Joseph et al., 2024). However, they are typically viewed primarily as predictive tools. While their forecasting performance can be strong, it is not guaranteed, as illustrated by the blue line in Figure 1. A parallel and growing strand of the literature integrates ML techniques—such as BART or other Bayesian approaches, random forests, and neural networks—with standard econometric models, where they serve as non-parametric tools for estimating non-linear coefficients within frameworks like VARs, local projections, asset pricing models, and DSGE models (Fernández-Villaverde and Guerrón-Quintana, 2021; Hauzenberger et al., 2023; Huber, 2023; Paranhos, 2024; Barbaglia et al., 2025). In contrast, the direct use of machine learning as a tool with inherent economic interpretability remains relatively underexplored. The recent innovations in structured ML approaches discussed in this paper seek to bridge this gap by moving ML toward partially identified representations.

The remainder of this paper is structured as follows. Section 2 examines the concept of interpretability as it is commonly defined in the machine learning literature, focusing on how input variables influence predictions. We discuss the main approaches typically used to achieve this: post-hoc interpretability tools and machine learning methods with

inherently interpretable structures. In Section 3, we explore how economically motivated structures can be incorporated into machine learning models a priori, either through block structures or monotonicity constraints, and review recent studies that have pursued this approach. Section 4 concludes by highlighting avenues for future research and the potential for a broader role of machine learning in central banks' analytical toolkits.

2 Interpretability in the machine learning context

Interpretability in prediction models has multiple dimensions. For economic policy, coherence with expert knowledge and theory, as well as communicability, are especially important. While we return to these aspects later, we now focus on a narrower definition common in the ML literature: "the degree to which an observer can understand the cause of a [model's] decision" (Miller, 2019). Under this definition, interpretability means understanding how input variables influence predictions—regardless of whether the model aligns with economic theory, accurately captures the data-generating process, or predicts well. footnote Interpretability is also closely linked to sparsity (Molnar, 2020; Burkart and Huber, 2021; Rudin et al., 2022). Miller (2019) show that people tend to prefer concise explanations that involve only a few contributing factors. However, in macroeconomics, sparse models often underperform compared to dense ones, and identifying stable sparse representations is challenging due to the high correlation among indicators (Giannone et al., 2021; Cross et al., 2020; Chu and Qureshi, 2023; Coulombe et al., 2024). Models like random forests, gradient boosting, and neural networks are generally considered non-interpretable. Although their individual operations—such as threshold comparisons, weighted summations, and monotonic transformations—are straightforward, the sheer volume and complexity of these operations render the models opaque to users.

The machine learning literature offers two broad approaches to achieving interpretability in this context. The first involves post-hoc interpretability tools, such as Shapley values, which help users make sense of black-box models after estimation; we discuss these in Section 2.1. However, we argue that Shapley values alone do not render black-box models truly interpretable, because complex models will inevitably produce complex explanations. The second avenue involves additive machine learning models, which effectively capture non-linear relationships between predictors and outcomes while retaining interpretability. We discuss these models in detail in Section 2.2. This class of models also serves as a foundation for incorporating economic structure, as further elaborate on in Section 3.

2.1 Post-hoc frameworks for signal interpretation

A wide variety of methods have been developed to interpret the predictions of machine learning models (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sundararajan et al., 2017; Ribeiro et al., 2018; Goldstein et al., 2015). These frameworks are applied after model training, i.e. *post-hoc*, and do not alter the model's output. In economic forecasting, they can reveal economically meaningful associations between variables and can point model

misspecifications or spurious patterns learnt from the data.

Arguably, the most prominent post-hoc interpretability framework is based on Shapley values (Shapley, 1951). Lundberg and Lee (2017) proposed SHAP (SHapley Additive exPlanations). SHAP decomposes a machine learning model's prediction for an observation, $\hat{y}_i$, into the sum of the contributions from each indicator: $\hat{y} = \sum_{j=1}^{M} \phi_j x + \phi_0$, where ϕ_j is the Shapley value associated with indicator j and phi_0 is the baseline value (typically the mean predicted value). Shapley values can be applied to any type of ML model and have seen widespread use in macroeconomic applications, including forecasts of financial crises (Bluwstein et al., 2023; Casabianca et al., 2022) and inflation (Aras and Lisboa, 2022; Lenza et al., 2023; Joseph et al., 2024).

However, post-hoc methods such as Shapley values are not a panacea; they do not render black-box models as interpretable as simple linear models. Complex models inherently produce intricate explanations for their predictions. When explaining an individual prediction, for instance why a credit risk model denied a loan for a certain applicant, the sum of all indicators' Shapley values explain the prediction. However, in macroeconomics, the goal is broader understanding, not isolated outcomes (e.g. the predicted inflation in a specific month). Fully grasping a model with many interactions requires analysing a large set of Shapley values, including interaction effects (Murdoch et al., 2019). Rudin (2019) describes this issue clearly: "*Explanations must be wrong. [...] If the explanation was completely faithful to what the original model computes, the explanation would equal the original model[...]*". The issue of complex explanations is particularly pressing for algorithms like random forests, gradient boosting, and neural networks that do not tend to learn sparse models from data but instead use a large number of variables and their interactions.

A second limitation of Shapley values is their approximate nature. Their estimation relies on permutations and approximations that may lead to inaccurate measures of variable importance (Kumar et al., 2020; Janzing et al., 2020; Sundararajan and Najmi, 2020; Huang and Marques-Silva, 2023, 2024). Moreover, they are frequently misunderstood or misapplied by practitioners, as shown by Kaur et al. (2020). The authors observed that inherently interpretable additive models (which we discuss in the next sub-section) are used more accurately and impose a lower cognitive load.

2.2 Interpretable machine learning and additive models (AMs)

Inherently interpretable non-linear ML methods, such as decision trees, do not require post-hoc interpretability tools. Their predictions can be easily explained by following the tree's decision nodes. However, single decision trees are rarely used in economic forecasting due to their limited predictive power and non-smooth outputs.[1] Additive models offer a more suitable alternative for economic forecasting.

Additive models (AMs)(Hastie and Tibshirani, 1990) capture non-linear relationships between predictors and outcomes while remaining highly interpretable. An AM with k

[1]The same applies to decision lists (Rivest, 1987; Wang and Rudin, 2015), another interpretable model class often used in medical decision-making.

indicators learns separate smooth functions for each variable and sums them to generate a prediction. Formally, the model takes the form:

$$F(\mathbf{x}) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k), \tag{1}$$

where each $f_i(x_i)$ is a smooth function of predictor i learnt from the data.[2] AMs do *not* allow for interactions of predictors and the marginal effect of a predictor is directly visible from its corresponding function $f_i(x_i)$.[3]

AMs offer several advantages for predictive economic modeling. First, their additive structure resembles linear models–the predominant approach in econometrics. Linear models are a special case of AM when each $f_i \forall i$ is restricted to be linear. Second, standard approaches to learning AMs employ smooth functions f_i, which is a desirable property for economic models. Third, the functions f_i can incorporate expert knowledge through constraints, such as enforcing a positive association between the predictor x_i and the outcome. We discuss this in detail in Section 3.2. Finally, as we discuss in Section 3.1, AMs can be extended to learn from blocks of variables that represent distinct economically meaningful determinants.

A key component of AMs is the smoothing function used to estimate the relationship between each predictor and the outcome. Traditionally, non-parametric regression techniques like splines are employed (Hastie and Tibshirani, 1990). Other work has used boosting to train additive models. Boosting is indeed a natural choice since it builds predictions by sequentially *adding* the contributions of its base learners (Lou et al., 2012). When each base learner is restricted to a single predictor, the model retains its additive structure.

To encourage sparsity—and thus improve interpretability—in AMs, *component-wise boosting* can be applied (Bühlmann and Yu, 2003; Buehlmann, 2006; Schmid and Hothorn, 2008; Hothorn et al., 2010; Groll and Tutz, 2012), where, in each boosting iteration, one candidate base learner is trained on each of the predictors, but only the one that reduces the error the most is added to the ensemble. This often yields sparse models in practice.[4]

The success of neural networks in prediction problems has spurred the development of additive models based on neural networks (Agarwal et al., 2021). These models employ a constrained architecture where separate sub-networks are trained on individual indicators, and combined linearly in the output layer. While boosting learns the base learners sequentially from data, neural networks allow simultaneous optimisation of sub-networks. Agarwal et al. (2021) show that neural AMs perform on par with additive models learnt

[2]In the literature these models are usually referred to as generalised additive models (GAM), which, as generalised linear models (Nelder and Wedderburn, 1972) are formulated with a link function $g(\hat{F})$ where g is a monotone, differentiable function chosen to match the distribution of the dependent variable (e.g., logit for binomial, log for Poisson, identity for Gaussian variables). Without loss of generality and for simplicity, we omit the link function g and omit the term *generalised*.

[3]Bordt and von Luxburg (2023) show the direct correspondence between Shapley values and AMs: variable contributions can be directly extracted from an AM without separate Shapley value estimation.

[4]See Bühlmann et al. (2014) for a comparison of component-wise boosting to Lasso regularisation, the more prominent approach to training sparse models.

with boosted trees.[5]

While additive models enhance interpretability, they sacrifice expressiveness (i.e., the ability to capture complex patterns in the data) by not capturing interactions between variables. As a result, they often underperform compared to black-box models such as random forests or gradient boosting (Lou et al., 2012; Zschech et al., 2022). To overcome this limitation, several studies have developed algorithms that selectively incorporate pairwise interactions, thereby improving predictive performance while retaining much of the interpretability Lou et al. (2013); Yang et al. (2021); Enouen and Liu (2022).

In economic forecasting, Kauppi and Virtanen (2021) use component-wise boosting with regression splines to predict macroeconomic time series.[6] Other studies apply linear component-wise boosting models, primarily for variable selection in macroeconomic prediction (Lehmann and Wohlrabe, 2016; Zeng, 2017).

How non-linear additive models (AMs) compare to black-box methods like random forests in economic prediction remains an open question. While black-box models often outperform in standard machine learning tasks, these results may not extend to economic time series, which typically involve small sample sizes and highly correlated variables.[7] Rather, constrained models, such as additive models might actually perform better, as their structure acts as regularisation, helping to avoid spurious patterns that fail to generalise to the test data. Additional constraints—such as imposing monotonic relationships—could further enhance performance. We explore this in detail in Section 3.2.

3 Machine learning with economically plausible constraints

In this section, we move beyond the basic notion of interpretability as merely identifying signals that drive predictions. Instead, we emphasize the integration of economic intuition by organizing variables into economically meaningful blocks and imposing constraints on functional forms. These structural choices enhance interpretability by ruling out estimates that are economically implausible (see Elliott and Timmermann, 2016).

A key pillar of a broader definition of interpretability is trust. As Rudin et al. (2022) note: "*Despite common rhetoric, interpretable models do not necessarily create or enable trust—they could also enable distrust. They simply allow users to decide whether to trust them*". If a model's outputs contradict expert knowledge, it is unlikely to gain acceptance. The alignment of a model's decision processes and the user's knowledge or intuition is

[5]Compared to boosted AMs, the authors argue that neural networks offer a more flexible learning paradigm, for example allowing more than one dependent variable. Neural additive models might also offer a more parsimonious representation than additive models with many decision trees. Luber et al. (2023), Chang et al. (2021), and Kraus et al. (2024) extend the work on neural AMs by proposing more parameter-efficient neural network architectures and training paradigms that perform competitively.

[6]Their two-step approach first fits a linear model, then applies the boosted spline model to the residuals.

[7]Rudin (2019) argues more broadly for machine learning models with tabular data: "*It is a myth that there is necessarily a trade-off between accuracy and interpretability.*"

widely recognised as essential in the literature on interpretable machine learning (Doshi-Velez and Kim, 2017; Selbst and Barocas, 2018; Rudin, 2019; Simkute et al., 2021) and is equally crucial when communicating model outputs to economic policymakers.

3.1 Blockwise Additive Models (BAMs)

We define *Blockwise Additive Models* (BAMs) as predictive models characterized by an additive structure over groups–or *blocks*–of variables, rather than individual variables as standard additive models (AMs). While we introduce the term *blockwise additive models* to the literature, the underlying concept has been explored in both the economics and machine learning literature under different names or frameworks. Models aligning with our definition of BAMs have been introduced in empirical studies applying machine learning to decompose inflation drivers (Goulet Coulombe, 2024; Buckmann et al., 2025) and forecast bond prices (Bianchi et al., 2021). We discuss these applications in detail in Section 3.1.1.

The defining characteristic of these models is that the outcome is predicted as a sum of functions applied to subsets of features:

$$F(\mathbf{x}) = \sum_{j=1}^{p} f^j(\mathbf{x}^j), \tag{2}$$

where each $\mathbf{x}^j$ represents a feature block, and $f^j(\mathbf{x}^j)$ is the associated *block score* learnt for block j. As with standard additive models, these functions can be learnt using machine learning techniques, such as neural sub-networks or boosted decision trees. This structure allows for interactions *within* blocks, but not *across* them.

BAMs offer greater interpretability than standard black-box machine learning models—provided that each block of variables and its associated score f^j have a meaningful interpretation. This design mirrors a core principle in macroeconomics: economic outcomes are driven by a limited number of structural shocks or unobserved latent forces. A common empirical framework to capture such latent drivers is the factor model, which assumes that the co-movement of many macroeconomic series can be summarized by a few static or dynamic factors (Stock and Watson, 2016).[8] While traditional factor models are typically linear, introducing time variation or non-linearity often requires strong parametric assumptions or Bayesian priors to regularize model complexity (Stock and Watson, 2009). In contrast, BAMs flexibly estimate the relationship between the target variable and economic indicators using machine learning techniques such as decision trees—allowing for non-linearities and interactions within blocks without imposing rigid structural assumptions.

The block structure in BAMs reflects the modeller's intuition that groups of variables jointly represent an unobserved or underlying economic driver of

[8]In classical factor models, residual fluctuations are assumed to be idiosyncratic and uncorrelated. Generalized dynamic factor models allow for more flexible structures, including lagged factor loadings, enabling them to capture asynchronous and frequency-specific co-movements across time series (Forni et al., 2000).

the target variable. The pinning down of "interpretable" blocks parallels structural dynamic factor models, in which factors are extracted from pre-defined groups of variables to facilitate economic interpretation (Kose et al., 2003; Potjagailo and Wolters, 2023), rather than being derived solely through statistical decomposition. Furthermore, BAMs do not impose assumptions like orthogonality between blocks or other structural constraints that are common in traditional factor models.

Similar to Partial Least Squares (PLS) approaches (Mateos-Aparicio, 2011), but in contrast to standard factor models, BAMs learn block-level scores by directly optimizing predictive performance. In doing so, they unify two steps that are typically conducted sequentially in econometric workflows: (1) extracting components using an unsupervised method (such as factor models or filters), and (2) employing these components in a supervised model.

The assignment of variables to blocks is determined by the modeller, who applies economic judgment to define the composition and interpretation of each block. For example, a modeller aiming to forecast GDP growth might structure the BAM into blocks representing different sectors of the economy, or distinguish between real and nominal indicators, global and domestic influences, or exogenous factors and endogenous policy variables. This approach enables BAMs to learn potentially complex, non-linear associations among indicators *within* each block, while preserving an additive (and thus conditionally linear) structure *between* blocks.

Similar to factor models, the block structure in BAMs reduces the need to enforce sparsity at the level of individual variables to achieve interpretability. **Interpretation is primarily centred on the block level, rather than on individual predictors.** Nevertheless, the modeller may still wish to communicate the contributions of individual variables within a block. This can be done using post-hoc interpretability techniques, such as Shapley values. Importantly, these individual signals remain interpretable, as they are naturally contextualized within the economically meaningful determinant represented by the block to which they belong.

BAMs also share conceptual similarities with autoencoders (Berahmand et al., 2024), a neural network approach designed to learn lower-dimensional, non-linear representations of high-dimensional input data, without imposing structural constraints such as orthogonality. However, key distinctions remain. First, unlike BAMs, autoencoders learn representations from the entire set of input variables, without incorporating any prior grouping or block structure. Second, autoencoders are trained in an unsupervised manner—i.e., the latent representation is not optimised to predict a target variable.

3.1.1 BAMs in economics and finance applications

Recent empirical work demonstrates the growing appeal of BAMs for interpreting and forecasting macroeconomic variables such as inflation and bond returns. While differing in model architecture and application, all three studies underscore how BAMs can combine predictive performance with economic interpretability by structuring models around theoretically meaningful blocks of indicators.

Bianchi et al. (2021) forecast US Treasury bond returns using a neural network with a custom architecture that "*takes into account the economic structure of the input data.*" The model features multiple sub-networks, with one dedicated to bond forward rates, alongside others corresponding to groups of macroeconomic variables such as output and income, the labour market, and housing. These sub-networks operate independently within the hidden layers and are only combined at the output layer. This design reflects the structure of a BAM, as it preserves interpretability by maintaining separability between economically meaningful blocks. The authors show that this BAM performs as well as, or better than, an unconstrained neural network that allows full interaction across all variables. This leads them to conclude that the strong performance stem primarily from non-linearities *within* blocks, rather than from interactions *between* blocks. While the primary focus of the paper is predictive performance, the authors also leverage the interpretability of the BAM structure by analysing the learnt block scores.

Goulet Coulombe (2024) introduces the *Neural Phillips Curve*, a neural network-based BAM designed to forecast and decompose the determinants of U.S. inflation. The model groups input variables according to the structural components of a New Keynesian Phillips Curve framework. Specifically, the input blocks correspond to long-run inflation expectations, short-run inflation expectations, real economic activity, and commodity and energy prices, with an extended specification that also includes financial indicators. This blockwise structure enables the model to reflect economic relationships while leveraging the flexibility of neural networks to learn non-linear associations within each group.

In contrast to the shallow network architecture employed by Bianchi et al. (2021), the Neural Phillips Curve model features deep sub-networks, each consisting of five hidden layers. This depth enables the model to capture complex non-linear relationships and interactions among variables *within* each block, while preserving additivity *across* blocks. Each sub-network culminates in a single output neuron, and the final prediction is obtained by summing the outputs of these neurons. This is illustrated in Figure 2, which replicates Figure 1 from Goulet Coulombe (2024).

The author emphasizes the economic interpretability of his model, noting that it enables model predictions to be directly decomposed into the sum of contributions from the Phillips Curve components, which he refers to as "latent states." The model draws on a large set of time series, from which it extracts the optimal "summary statistic" for each component by maximizing predictive performance. The long-run expectations component is proxied by a time trend, while the short-term expectations component is informed by survey expectations and lagged price series. The real activity block draws on labour market, industrial production, and national accounts data, whereas the energy price block incorporates oil, gas, and metal prices. Each indicator is included with four lags and three moving averages.

In an extension to the base model, Goulet Coulombe allows for a time-varying slope of the Phillips Curve, decomposing the contribution of real activity variables into a slowly evolving slope coefficient and a more cyclical output gap. This approach draws on identification assumptions common in unobserved component models and time-varying parameter factor models, where structural parameters are assumed to change gradually over time, cap-

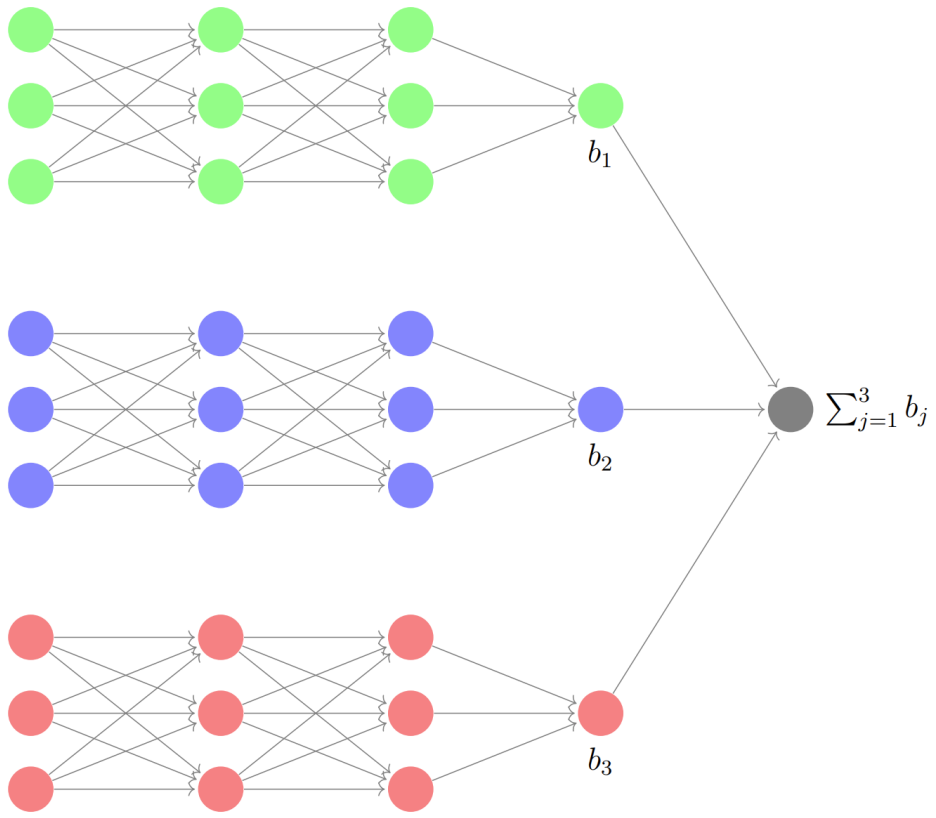

Figure 2: Illustration of a neural BAM. Source: Replication of Figure 1 in Goulet Coulombe (2024).

turing long-run shifts, while cyclical components fluctuate at higher frequencies (see, e.g., Chan et al., 2018). The same methodology is applied to estimate time-varying coefficients for the short-run expectations and commodity price components. The author demonstrates that the model performs competitively relative to a wide set of benchmarks—including a linear Phillips Curve and a random forest—and notably outperforms all benchmarks after 2020 by better capturing the sharp rise in inflation.

Finally, Buckmann et al. (2025) examine the non-linear drivers of inflation in the United Kingdom using the *Blockwise Boosted Inflation Model*, a BAM based on a blockwise boosted tree approach. The model learns associations between inflation and five blocks of variables: an expectations-informed trend and global and domestic demand and supply components. The latter draw on real activity variables, identified shock series, and indicators of cost pressures and supply-chain disruptions. To distinguish between demand and supply blocks, the authors impose sign restrictions on the learnt associations within each block–another approach to incorporating economic theory into machine learning which we discuss in greater detail in the next section. For each block, up to 200 trees are trained sequentially, with each block conditioned on the others. The authors use cross-validation to assess the predictive contributions of the individual components. An out-of-sample forecasting exercise

further demonstrates that the model performs as well as less interpretable models such as a random forest and standard boosted tree model.

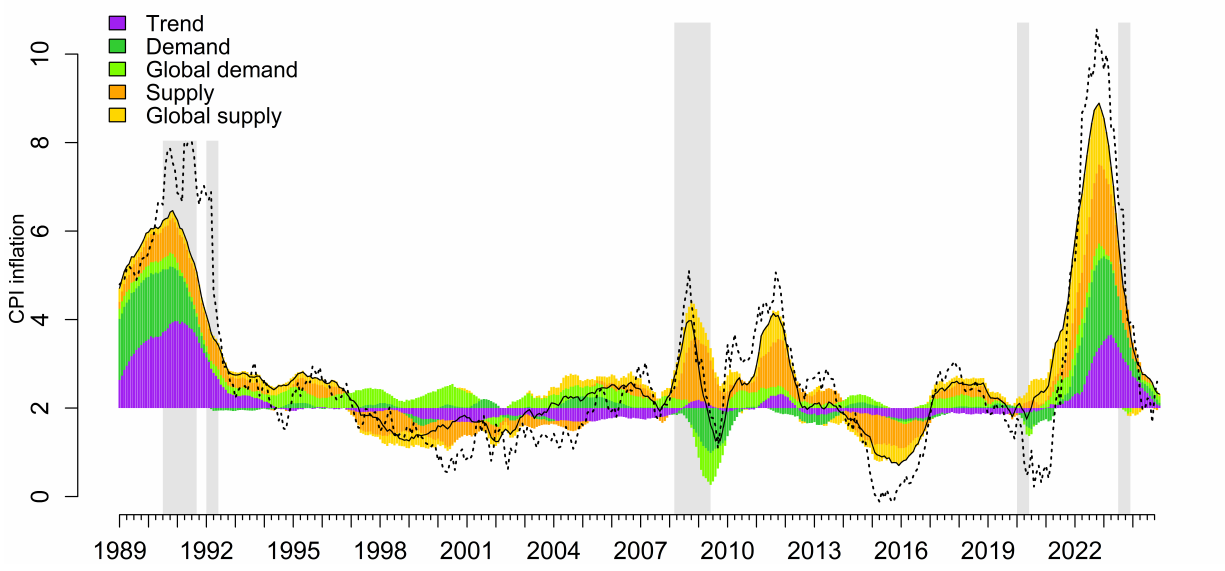

Figure 3: Decomposition of UK inflation into inflation trend, demand and supply contribution from a blockwise boosted tree model. Source: Buckmann et al. (2025).

Buckmann et al. (2025) also show that the model delivers an intuitive time-series decomposition of inflation into blockwise contributions, illustrated in Figure 3. Monthly inflation predictions (black line) and the component contributions are smoothed over 12 months. Notable patterns include a negative demand contribution during the global financial crisis and pronounced supply-side drags during periods of falling global energy prices and increased supply chain integration. In the most recent inflation episode, the model attributes the surge to a combination of strong supply-side effects, a positive demand contribution, and an upward shift in the inflation trend.

3.1.2 BAMs in the machine learning literature

While additive models (AMs) are well established in the machine learning literature, BAMs have received comparatively less attention—likely due to the field's traditional emphasis on data-driven pattern discovery over structured model design. Nonetheless, several contributions point to blockwise structures. Hothorn et al. (2010) introduces the `mboost` R package for *model-based* boosting, enabling BAM estimation, while Mayer et al. (2021) show how popular boosting libraries like xgboost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) can implement BAMs by only allowing interactions of variables within blocks, demonstrating high interpretability in housing price applications.

Other contributions are more tangential to BAMs. Obster and Heumann (2024), drawing on ideas from group Lasso (Simon et al., 2013), propose a linear component-wise boost-

ing method for variable selection both within and between blocks, which can be viewed as a linear BAM. Chen and Ye (2024) extend a neural additive model to allow interactions of variables within blocks, effectively training a BAM. Their method outperforms standard additive models and matches unconstrained neural networks on three financial datasets, while offering greater interpretability. In contrast to our definition of BAMs, block structure is inferred from the data rather than imposed by the modeller. The authors caution that "detecting statistical interactions without considering domain knowledge can result in oversimplified and unreasonable models."

3.2 Monotonicity constraints

To ensure economic interpretability, it may not be sufficient to merely separate blocks of determinants if these blocks do not uniquely pin down the direction of association between a predictor and the variable of interest. In practice, indicators grouped within the same block can exhibit different or even opposing—associations with the target variable. As a result, the model might learn effects that contradict expert knowledge or overlook relevant signals due to offsetting influences within a block.

To address such ambiguities, the modeller can impose directional constraints on the relationships between predictors and the target variable. For example, in a model predicting the risk of lung cancer, it is reasonable to expect a positive monotonic relationship with smoking: holding other risk factors constant, the more cigarettes an individual smokes per day, the higher their risk of developing lung cancer. Similarly, in an inflation prediction model, demand-driven inflation should be positively associated with indicators of economic activity, such as GDP growth, whereas supply-driven inflation is expected to show a negative association. Imposing directional constraints helps ensure that the model reflects these theoretically grounded relationships.

In the machine learning literature, such constraints on the direction of association between an input and output variable are known as *monotonicity constraints* and are often regarded as an important element of model interpretability (Martens et al., 2011; Freitas, 2014; Burkart and Huber, 2021; De Bock et al., 2024). **A monotonicity constraint restricts the direction of association between a feature and the predicted outcome.** For instance, a positive monotonicity constraint requires that an increase in a feature must not lead to a decreasing in the predicted outcome if all other features are kept constant.

Formally, for a predictive model f a monotonicity constraint requires that:

$$f(x_1, x_2, \cdots, x, \cdots) \geq f(x_1, x_2, \cdots, x', \cdots)$$

when $x \geq x'$ is a positive (increasing) constraint; or

$$f(x_1, x_2, \cdots, x, \cdots) \leq f(x_1, x_2, \cdots, x', \cdots)$$

when $x \leq x'$ is a negative (decreasing) constraint.

Monotonicity constraints limit the flexibility of machine learning models, which can, in principle, negatively affect predictive performance. To manage this trade-off,

various studies have proposed algorithms that balance the degree to which monotonicity constraints are imposed with the goal of maintaining high accuracy (Ben-David, 1995; Cano et al., 2019). **At the same time, monotonicity constraints can enhance predictive performance by acting as informed priors. They can also be viewed as a form of regularization, preventing the model from learning overly complex functional forms that might lead to overfitting.** As such, monotonicity constraints can be particularly advantageous when training data is limited or the prediction task is inherently difficult.

To address such ambiguities, the modeller can impose directional constraints on the relationships between predictors and the target variable. For example, in a model predicting the risk of lung cancer, it is reasonable to expect a positive monotonic relationship with smoking: holding other risk factors constant, the more cigarettes an individual smokes per day, the higher their risk of developing lung cancer. Similarly, in an inflation prediction model, demand-driven inflation should be positively associated with indicators of economic activity, such as GDP growth, whereas supply-driven inflation is expected to show a negative association. Imposing directional constraints helps ensure that the model reflects these theoretically grounded relationships.

3.2.1 Monotonicity constraints vs. sign restrictions for shock identification

In standard econometric models—especially vector autoregressions (VARs) and dynamic factor models (DFMs)—sign restrictions are widely used to incorporate economic intuition, particularly for identifying supply and demand drivers of variables like output, oil prices, or inflation. While both sign restrictions and monotonicity constraints embed a priori economic knowledge, they operate differently and it is important not to conflate the two.

Sign (or zero) restrictions in VARs (Arias et al., 2018) constrain the dynamic responses of variables to structural shocks by filtering the set of admissible impulse responses—those that satisfy the specified sign or zero restrictions over a limited number of horizons—from all possible residual decompositions. Applied as a post-estimation procedure, they identify orthogonal structural shocks based on contemporaneous correlations among variables and can shape the joint response of multiple variables to a common shock.

By contrast, monotonicity constraints in machine learning models are imposed directly during model estimation, restricting the functional relationship between individual predictors and the target variable across the entire dataset. These constraints shape predictions throughout the feature space, not merely in response to a shock. Although they do not yield an identified series of structural shocks, when applied to multiple indicators simultaneously, monotonicity constraints can enable a nuanced decomposition of the prediction into contributions from each determinant.

3.2.2 Implementing monotonicity constraints in ML models

Several approaches have been developed to effectively implement monotonicity constraints, particularly within neural networks and decision tree-based methods (for reviews, see Cano

et al., 2019; Nanfack et al., 2022). In neural networks, monotonicity can be induced by either (i) creating a specific, constrained architecture that guarantees monotonicity or (ii) modifying the loss function. A constraint architecture may limit the expressiveness of the models and thus degrade their performance, or the implementation of the constraints comes with high computational complexity (e.g. Sill, 1997; Daniels and Velikova, 2010; You et al., 2017). However, recent advances suggest that these drawbacks can be effectively mitigated (Runje and Shankaranarayana, 2023; Kitouni et al., 2023). Modified loss functions are applicable to any neural network architecture, but cannot guarantee that the learnt neural network is consistently monotonic (Sill and Abu-Mostafa, 1996; Gupta et al., 2019). Consequently, other studies have introduced additional computational procedures to certify that monotonicity is satisfied (Liu et al., 2020; Sivaraman et al., 2020), albeit at increased computational costs.

In the work on monotonic decision trees, several methods adjust the loss function to penalise monotonicity violations but cannot guarantee monotonicity (Ben-David, 1995; González et al., 2015). Even methods that *locally* blocks splits violating monotonicity constraints do not ensure that the resulting tree satisfies *global* monotonicity, i.e. across all nodes (Bartley et al., 2019). The popular boosting library `xgboost` implements global monotonicity constraints by tracking the minimum and maximum permissible predicted values during tree growth.[9] Bonakdarpour et al. (2018) propose an alternative method that enforces monotonicity by directly adjusting tree predictions, without altering the tree structure (see also van de Kamp et al., 2009). Bartley et al. (2019) introduce an approach tailored to random forests that reportedly outperforms other monotonicity-enforcing methods in this context.

3.2.3 Applications of monotonicity constraints in forecasting

Monotonicity constraints are particularly influential in predictive applications within high-stakes domains such as medicine (Pazzani et al., 2001; Royston, 2000) and credit scoring (Chen and Li, 2014; Chen and Ye, 2022). Notably, Chen and Ye (2022) propose a monotonic neural additive model, combining two key pillars of interpretability–additivity and monotonicity constraints. Chen and Zhang (2023) apply this model on datasets from criminology, education, health care, and finance.

Monotonicity constraints have also gained traction in economic and financial forecasting, primarily due to their potential to enhance predictive performance. Campbell and Thompson (2008) and Li and Tsiakas (2017) demonstrate that imposing monotonicity constraints on predictors in linear models can significantly improve the accuracy of stock return forecasts. Similarly, Wen et al. (2022) finds that applying such constraints enhances predictive performance in oil price forecasting. Other studies extend these findings to non-linear models: Fisher et al. (2020) show that monotonicity constraints improve forecasts of firms'

[9]Bartley et al. (2019) show that this approach can introduce significant biases in deeper trees, as commonly found in random forests. In boosting applications, however, this method performs well since the trees are typically shallow and the sequential learning process helps correct for any induced biases.

expected returns using a Bayesian non-linear additive model, while Richman and Wüthrich (2024) report similar gains in the context of insurance pricing with neural networks.

In macroeconomic forecasting, theory-driven regularisation can be particularly valuable due to typically small samples sizes and high intercorrelation between variables, which can lead to overfitting and unstable parameter estimates if the model is not constrained. However, to date, only a few studies have applied sign restrictions in machine learning models for macroeconomic applications. Chalaux and Turner (2023) develop a machine learning algorithm for selection of linear statistical models to predict economic downturns across 20 OECD countries, selecting predictors based on both empirical relevance and theoretically motivated monotonicity constraints. They differentiate between short-term sign restrictions applied to quarter-on-quarter and year-on-year changes and "long-term" sign restrictions imposed on annual changes over 3–5 years. For instance, the authors explain that "a strong positive growth rate in real house prices over the previous five years as well as a sharp fall over the previous quarter might both signal the increased likelihood of a downturn." Lin et al. (2024) incorporate monotonicity constraints when forecasting exchange rates using boosted trees.

3.2.4 Monotonicity constraints to separate determinants of inflation

In a recent study, Buckmann et al. (2025) apply monotonicity constraints within a BAM using boosted trees to disentangle supply- and demand-driven determinants of inflation (see also Section 3.1.2). Indicators of real activity predict inflation through both a demand and a supply block, meaning their effect can be attributed to either source. To separate these channels in line with economic theory, the authors impose opposite monotonicity constraints depending on the block. For example, the unemployment rate gap is constrained to have a negative association with inflation within the demand block and a positive association within the supply block. Specifically, the decision trees in the demand block are restricted to splits where a lower unemployment gap corresponds to higher inflation, while splits where the two move in the same direction are excluded.

Unlike VARs, which identify shocks through simulated dynamic responses with imposed sign restrictions (see Section 3.2.1), these constraints directly shape the functional relationship between predictors and the target variable across the entire sample. By training numerous trees on a wide range of real activity indicators, the model exploits variation over time and across indicators to separate supply- and demand-like contributions at each point in time. Although the model remains predictive rather than fully structurally identified, the monotonicity constraints offer a theory-based method for disentangling supply- and demand-driven influences on inflation. As shown in Figure 4, these constraints help uncover meaningful cyclical demand effects that are missed in unconstrained models. Consequently, a decomposition like that in Figure 3, estimated without monotonicity constraints, fails to capture plausible demand cycles and understates the role of supply-side factors.

To visualise the non-linearities captured by the boosting model, the authors estimate Shapley values (see Section 2.1) and plot them in a scatter plot against the realisations of the corresponding indicator variable, as also shown here in Figure 5. This reveals pronounced

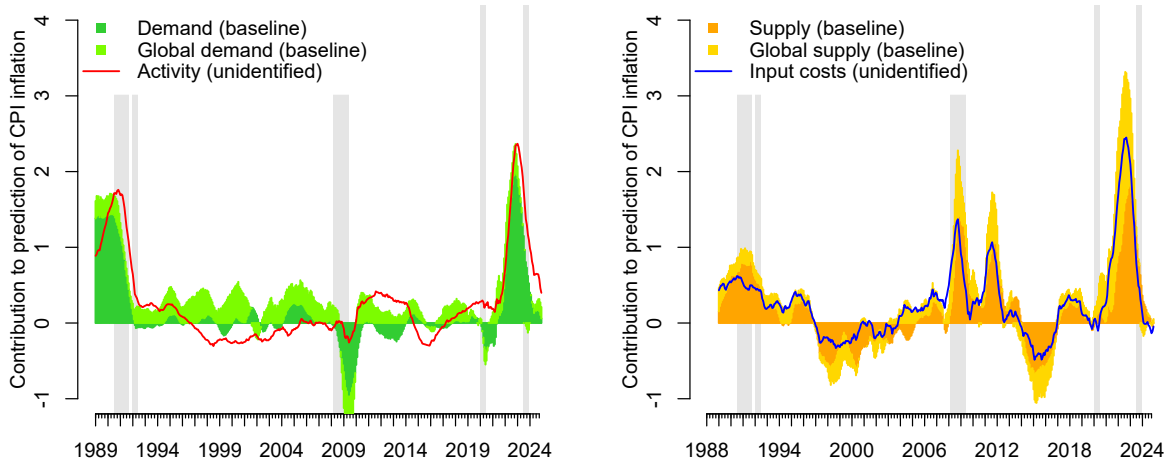

Figure 4: Contribution to inflation from demand and supply blocks with monotonicity constraints (baseline) versus blocks of indicators without restrictions (unidentified). Source: Buckmann et al. (2025).

non-linear associations between inflation and both the unemployment rate gap and the vacancy-to-unemployment (v/u) ratio gap —a measure of labour market tightness—in the demand block. The authors observe a stronger predictive effect on inflation at low levels of unemployment and, similarly, at high levels of the v/u ratio. These non-linearities are consistent with evidence from time series analyses and theoretical models that suggest a non-linear Phillips curve relationship between inflation and labour market tightness, that has been particularly relevant in the recent inflation surge. In the supply block, the effect of labour market variables—imposed to have the opposite sign by the monotonicity constraints—is less pronounced and exhibits little evidence of non-linearity. Instead, other variables in the supply block, such as two indices of global supply chain disruptions, display marked non-linearities.

Despite the relatively restrictive design—imposing both a block structure and monotonic constraints—the authors show that their model comparably to an unconstrained boosting model and a random forest in out-of-sample forecasts. This suggests that restrictions do not necessarily come at the expense of predictive performance.

Finally, in upcoming work, De Polis et al. (2025) measure the evolution of inflation risks over time via a flexible time-varying parameter model. They impose theory-driven monotonicity constraints on predictor loadings towards the moments of the predictive distribution of inflation in the short-run and the long-run. As such, the authors impose the unemployment gap and unit labour costs to have negative associations with the mean and the skew of the inflation distribution in the short run, whereas an international commodity price index is restricted towards a positive association. In the long run, money growth, trend unit labour costs and the long-run real interest rate are restricted to have positive

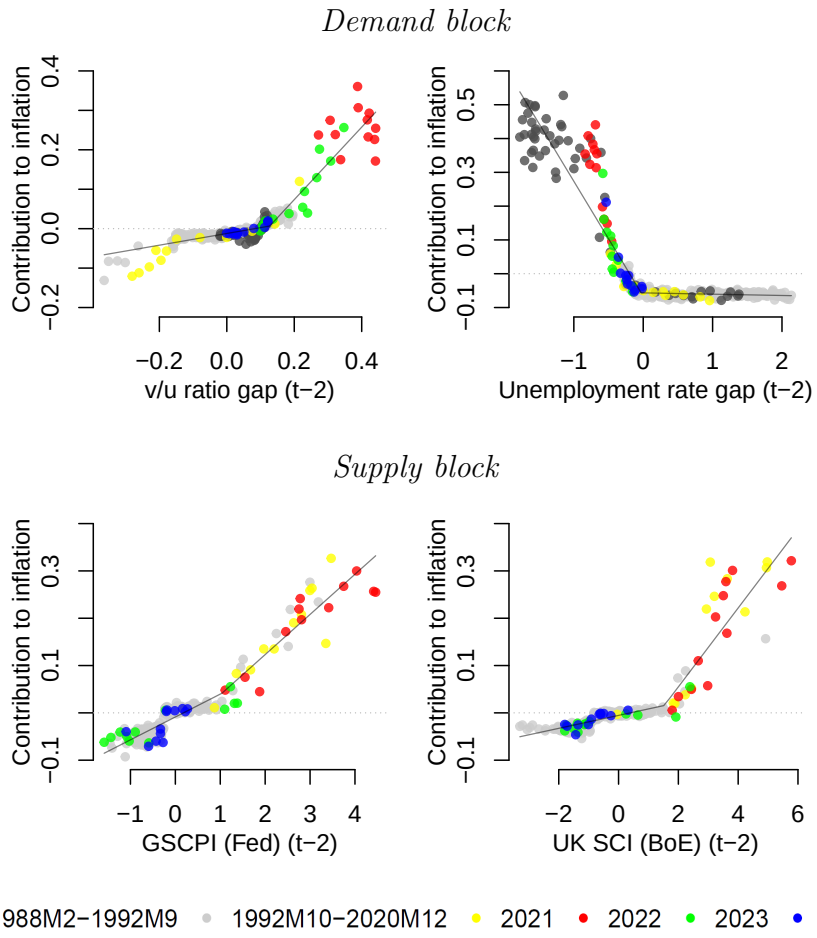

Figure 5: Non-linear demand and supply associations between selected indicators and UK inflation, learnt by blockwise boosted tree model.
Note: Functional forms reflect the contribution to the inflation prediction (y-axis) from a blockwise boosted tree model, against the realisation of the indicator at a given lag (x-axis). Colours indicate specific realisations for selected months. The lines show best-fit segments. Top panel shows selected indicators within the demand block, bottom panel shows selected indicators within the supply block. Source: Buckmann et al. (2025).

predictor loadings towards the mean and skew of the inflation distribution. This yields a decomposition of long-run inflation risks into meaningful determinants over time. For instance, they find that Phillips curve-type effects show up in inflation skewness.

4 Conclusion

This paper highlights the potential of integrating economic theory into machine learning models through structured architectures and interpretability-enhancing constraints. We have discussed how block-additive models (BAMs) and monotonicity constraints embed economic intuition into machine learning frameworks. By aligning model design with estab-

lished theoretical relationships—whether by grouping predictors into economically meaningful blocks or imposing monotonicity restrictions informed by expert knowledge—machine learning models can be steered toward greater economic interpretability without necessarily sacrificing predictive performance. Although these methods fall short of full structural identification, they offer a pragmatic middle ground between purely data-driven machine learning approaches and traditional econometric models. Post-hoc interpretation tools, such as Shapley values, remain valuable for illustrating the signals a model exploits and can provide clearer insights when applied to models explicitly designed for interpretability.

This opens new avenues for broadening the relevance of machine learning tools for policy institutions such as central banks, where particular value is placed on model forecasts that provide coherent and communicable narratives. Beyond the more established applications of machine learning as purely predictive tools or as tools for estimating non-linear coefficients in standard economic models, these ML methods can also serve to track non-linear determinants of economic aggregates in a way that aligns with economic theory.

Looking ahead, there is substantial scope for further research in this direction. Future work could assess the robustness of these methods in different macroeconomic contexts and examine potential biases introduced by the constraints discussed in this paper. In addition, advancing machine learning techniques toward multivariate models that enable the identification of economic drivers—whether through restrictions on the correlation structure between variables or through multivariate monotonicity constraints—would be a particularly promising area of development.

References

Agarwal, R., L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems 34*, 4699–4711.

Amari, S. (2006). A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers* (3), 299–307.

Aras, S. and P. J. Lisboa (2022). Explainable inflation forecasts by machine learning models. *Expert Systems with Applications 207*, 117982.

Arias, J. E., J. F. Rubio-Ramírez, and D. F. Waggoner (2018). Inference based on structural vector autoregressions identified with sign and zero restrictions: Theory and applications. *Econometrica 86*(2), 685–720.

Barbaglia, L., L. Frattarolo, N. Hauzenberger, D. Hirschbühl, F. Huber, L. Onorante, M. Pfarrhofer, and L. Tiozzo Pezzoli (2025). Interpretable Bayesian machine learning for assessing the effects of climate news shocks on firm-level returns. Available at SSRN 5133162.

Bartley, C., W. Liu, and M. Reynolds (2019). Enhanced random forest algorithms for partially monotone ordinal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, pp. 3224–3231.

Ben-David, A. (1995). Monotonicity maintenance in information-theoretic machine learning algorithms. *Machine Learning 19*, 29–43.

Benford, J. (2024, September). TRUSTED AI: Ethical, safe, and ef-

fective application of artificial intelligence at the Bank of England. https://www.bankofengland.co.uk/speech/2024/september/james-benford-speech-at-the-central-bank-ai-inaugural-conference. Speech at the Central Bank AI Conference.

Berahmand, K., F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu (2024). Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review 57*(2), 28.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *The Review of Financial Studies 34*(2), 1046–1089.

Bluwstein, K., M. Buckmann, A. Joseph, S. Kapadia, and Ö. Şimşek (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics 145*, 103773.

Bonakdarpour, M., S. Chatterjee, R. F. Barber, and J. Lafferty (2018). Prediction rule reshaping. In *International Conference on Machine Learning*, pp. 630–638. PMLR.

Bordt, S. and U. von Luxburg (2023). From Shapley values to generalized additive models and back. In *International Conference on Artificial Intelligence and Statistics*, pp. 709–745. PMLR.

Breiman, L. (2001). Random forests. *Machine learning 45*, 5–32.

Buckmann, M., G. Potjagailo, and P. Schnattinger (2025). Blockwise boosted inflation: non-linear determinants of inflation using machine learning. *Bank of England Staff Working Paper Series* (1143).

Buehlmann, P. (2006). Boosting for high-dimensional linear models.

Bühlmann, P., J. Gertheiss, S. Hieke, T. Kneib, S. Ma, M. Schumacher, G. Tutz, C.-Y. Wang, Z. Wang, and A. Ziegler (2014). Discussion of "the evolution of boosting algorithms" and "extending statistical boosting". *Methods of Information in Medicine 53*(06), 436–445.

Bühlmann, P. and B. Yu (2003). Boosting with the L 2 loss: regression and classification. *Journal of the American Statistical Association 98*(462), 324–339.

Burkart, N. and M. F. Huber (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research 70*, 245–317.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies 21*(4), 1509–1531.

Cano, J.-R., P. A. Gutiérrez, B. Krawczyk, M. Woźniak, and S. García (2019). Monotonic classification: An overview on algorithms, performance measures and data sets. *Neurocomputing 341*, 168–182.

Casabianca, E. J., M. Catalano, L. Forni, E. Giarda, and S. Passeri (2022). A machine learning approach to rank the determinants of banking crises over time and across countries. *Journal of International Money and Finance 129*, 102739.

Chalaux, T. and D. Turner (2023). Doombot: a machine learning algorithm for predicting downturns in OECD countries. *Documents de travail du Département des Affaires économiques de l'OCDE*.

Chan, J. C., T. E. Clark, and G. Koop (2018). A new model of inflation, trend inflation, and long-run inflation expectations. *Journal of Money, Credit and Banking 50*(1), 5–53.

Chang, C.-H., R. Caruana, and A. Goldenberg (2021). Node-gam: Neural generalized additive model for interpretable deep learning. *arXiv preprint arXiv:2106.01613*.

Chen, C.-C. and S.-T. Li (2014). Credit rating with a monotonicity-constrained support

vector machine model. *Expert Systems with Applications 41*(16), 7235–7247.
Chen, D. and W. Ye (2022). Monotonic neural additive models: Pursuing regulated machine learning models for credit scoring. In *Proceedings of the third ACM international conference on AI in finance*, pp. 70–78.
Chen, D. and W. Ye (2024). Generalized Groves of Neural Additive Models: Pursuing Transparent Machine Learning Models in Finance. In *2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, pp. 1–8. IEEE.
Chen, D. and L. Zhang (2023). Monotonicity for AI ethics and society: An empirical study of the monotonic neural additive model in criminology, education, health care, and finance. *arXiv preprint arXiv:2301.07060*.
Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.
Chipman, H. A., E. I. George, R. E. McCulloch, and T. S. Shively (2016). High-dimensional nonparametric monotone function estimation using BART. *arXiv preprint arXiv:1612.01619*.
Chu, B. and S. Qureshi (2023). Comparing out-of-sample performance of machine learning methods to forecast US GDP growth. *Computational Economics 62*(4), 1567–1609.
Cipollone, P. (2024, July). Artificial intelligence: a central bank's view. `https://www.ecb.europa.eu/press/key/date/2024/html/ecb.sp240704_1~e348c05894.en.html`. Speech at the National Conference of Statistics on official statistics at the time of artificial intelligence.
Coulombe, P. G., M. Goebel, and K. Klieber (2024). Dual Interpretation of Machine Learning Forecasts. *arXiv Preprint arXiv:2412.13076*.
Cross, J. L., C. Hou, and A. Poon (2020). Macroeconomic forecasting with large Bayesian VARs: Global-local priors and the illusion of sparsity. *International Journal of Forecasting 36*(3), 899–915.
Daniels, H. and M. Velikova (2010). Monotone and partially monotone neural networks. *IEEE Transactions on Neural Networks 21*(6), 906–917.
de Araujo, D. K. G., S. Doerr, L. Gambacorta, and B. Tissot (2024). Artificial intelligence in central banking. BIS Bulletin 2024.
De Bock, K. W., K. Coussement, A. De Caigny, R. Słowiński, B. Baesens, R. N. Boute, T.-M. Choi, D. Delen, M. Kraus, S. Lessmann, et al. (2024). Explainable AI for operational research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research 317*(2), 249–272.
De Polis, A., L. Melosi, and I. Petrella (2025). The macroeconomic drivers of the balance of inflation risks. Mimeo. Presented at Bank of England Macro Seminar.
Döpke, J., U. Fritsche, and C. Pierdzioch (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting 33*(4), 745–759.
Doshi-Velez, F. and B. Kim (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
Elliott, G. and A. Timmermann (2016). Forecasting in economics and finance. *Annual Review of Economics 8*(1), 81–110.
Enouen, J. and Y. Liu (2022). Sparse interaction additive networks via feature interaction detection and sparse selection. *Advances in Neural Information Processing Systems 35*, 13908–13920.
Fernández-Villaverde, J. and P. A. Guerrón-Quintana (2021). Estimating DSGE models:

Recent advances and future challenges. *Annual Review of Economics 13*(1), 229–252.
Fisher, J. D., D. W. Puelz, and C. M. Carvalho (2020). Monotonic effects of characteristics on returns. Ann. Appl. Stat. 14(4): 1622-1650 (December 2020).
Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics 82*(4), 540–554.
Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter 15*(1), 1–10.
Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.
Giannone, D., M. Lenza, and G. E. Primiceri (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica 89*(5), 2409–2437.
Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics 24*(1), 44–65.
González, S., F. Herrera, and S. García (2015). Monotonic random forest with an ensemble pruning mechanism based on the degree of monotonicity. *New Generation Computing 33*, 367–388.
Goulet Coulombe, P. (2024). A neural Phillips curve and a deep output gap. *Journal of Business and Economic Statistics*.
Groll, A. and G. Tutz (2012). Regularization for generalized additive mixed models by likelihood-based boosting. *Methods of Information in Medicine 51*(02), 168–177.
Gupta, A., N. Shukla, L. Marla, A. Kolbeinsson, and K. Yellepeddi (2019). How to incorporate monotonicity in deep networks while preserving flexibility? *arXiv preprint arXiv:1909.10662*.
Hastie, T. J. and R. J. Tibshirani (1990). *Generalized additive models*. Chapman & Hall/CRC.
Hauzenberger, N., F. Huber, G. Koop, and J. Mitchell (2023). Bayesian modeling of time-varying parameters using regression trees.
Hothorn, T., P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner (2010). Model-based Boosting 2.0. *Journal of Machine Learning Research 11*(71), 2109–2113.
Huang, X. and J. Marques-Silva (2023). The inadequacy of Shapley values for explainability. *arXiv preprint arXiv:2302.08160*.
Huang, X. and J. Marques-Silva (2024). A refutation of Shapley values for explainability.
Huber, F. (2023). Bayesian nonlinear regression using sums of simple functions. *arXiv preprint arXiv:2312.01881*.
Janzing, D., L. Minorics, and P. Blöbaum (2020). Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR.
Joseph, A., G. Potjagailo, C. Chakraborty, and G. Kapetanios (2024). Forecasting UK inflation bottom up. *International Journal of Forecasting*.
Kauppi, H. and T. Virtanen (2021). Boosting nonlinear predictability of macroeconomic time series. *International Journal of Forecasting 37*(1), 151–170.
Kaur, H., H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan (2020). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14.
Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017). Light-

gbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30*.

Kitouni, O., N. Nolte, and M. Williams (2023). Expressive monotonic neural networks. *arXiv preprint arXiv:2307.07512*.

Kose, M. A., C. Otrok, and C. H. Whiteman (2003). International business cycles: World, region, and country-specific factors. *American Rconomic Review 93*(4), 1216–1239.

Kraus, M., D. Tschernutter, S. Weinzierl, and P. Zschech (2024). Interpretable generalized additive neural networks. *European Journal of Operational Research 317*(2), 303–316.

Kumar, I. E., S. Venkatasubramanian, C. Scheidegger, and S. Friedler (2020). Problems with Shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, pp. 5491–5500. PMLR.

Lehmann, R. and K. Wohlrabe (2016). Looking into the black box of boosting: the case of Germany. *Applied Economics Letters 23*(17), 1229–1233.

Lenza, M., I. Moutachaker, and J. Paredes (2023). Density forecasts of inflation: A quantile regression forest approach. ECB Working Paper No 2830.

Li, J. and I. Tsiakas (2017). Equity premium prediction: The role of economic and statistical constraints. *Journal of Financial Markets 36*, 56–75.

Lin, C.-H., T. Liu, and K. Vincent (2024). Enhancing exchange rate forecasting accuracy: Integrating economic theories with boosted trees and monotonic constraints. Available at SSRN 4799906.

Liu, X., X. Han, N. Zhang, and Q. Liu (2020). Certified monotonic neural networks. *Advances in Neural Information Processing Systems 33*, 15427–15438.

Lou, Y., R. Caruana, and J. Gehrke (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, pp. 150–158.

Lou, Y., R. Caruana, J. Gehrke, and G. Hooker (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 623–631.

Luber, M., A. Thielmann, and B. Säfken (2023). Structural neural additive models: Enhanced interpretable machine learning. *arXiv preprint arXiv:2302.09275*.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*.

Martens, D., J. Vanthienen, W. Verbeke, and B. Baesens (2011). Performance of classification models from a user perspective. *Decision Support Systems 51*(4), 782–793.

Masini, R. P., M. C. Medeiros, and E. F. Mendes (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys 37*(1), 76–111.

Mateos-Aparicio, G. (2011). Partial least squares (PLS) methods: Origins, evolution, and application to social sciences. *Communications in Statistics-Theory and Methods 40*(13), 2305–2317.

Mayer, M., S. C. Bourassa, M. Hoesli, and D. Scognamiglio (2021). Structured additive regression and tree boosting. *Swiss Finance Institute Research Paper* (21-83).

Medeiros, M. C., G. F. Vasconcelos, Á. Veiga, and E. Zilberman (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics 39*(1), 98–119.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence 267*, 1–38.

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences 116*(44), 22071–22080.

Nanfack, G., P. Temple, and B. Frénay (2022). Constraint enforcement on decision trees: A survey. *ACM Computing Surveys (CSUR) 54*(10s), 1–36.

Nelder, J. A. and R. W. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society 135*(3), 370–384.

Obster, F. and C. Heumann (2024). Sparse-group boosting: Unbiased group and variable selection. *The American Statistician*, 1–14.

Paranhos, L. (2024). How do firms' financial conditions influence the transmission of monetary policy? A non-parametric local projection approach. *Journal of Econometrics*, 105886.

Pazzani, M., S. Mani, and W. Shankle (2001). Acceptance by medical experts of rules generated by machine learning. *Artificial Intelligence 1*(700.568), 110–065.

Potjagailo, G. and M. H. Wolters (2023). Global financial cycles since 1880. *Journal of International Money and Finance 131*, 102801.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). Model-agnostic interpretability of machine learning. *arXiv Preprint arXiv:1606.05386*.

Ribeiro, M. T., S. Singh, and C. Guestrin (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32.

Richman, R. and M. V. Wüthrich (2024). Smoothness and monotonicity constraints for neural networks using ICEnet. *Annals of Actuarial Science*, 1–28.

Rivest, R. L. (1987). Learning decision lists. *Machine Learning 2*, 229–246.

Royston, P. (2000). A useful monotonic non-linear model with applications in medicine and epidemiology. *Statistics in medicine 19*(15), 2053–2066.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1 (5): 206–215.

Rudin, C., C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys 16*, 1–85.

Runje, D. and S. M. Shankaranarayana (2023). Constrained monotonic neural networks. In *International Conference on Machine Learning*, pp. 29338–29353. PMLR.

Schmid, M. and T. Hothorn (2008). Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis 53*(2), 298–311.

Selbst, A. D. and S. Barocas (2018). The intuitive appeal of explainable machines. *Fordham L. Rev. 87*, 1085.

Shapley, L. S. (1951). Notes on the n-person game—ii: The value of an n-person game.

Sill, J. (1997). Monotonic networks. *Advances in Neural Information Processing Systems 10*.

Sill, J. and Y. Abu-Mostafa (1996). Monotonicity hints. *Advances in Neural Information Processing Systems 9*.

Simkute, A., E. Luger, B. Jones, M. Evans, and R. Jones (2021). Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable. *Journal of Responsible Technology 7*, 100017.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal*

of computational and graphical statistics 22(2), 231–245.

Sivaraman, A., G. Farnadi, T. Millstein, and G. Van den Broeck (2020). Counterexample-guided learning of monotonic neural networks. *Advances in Neural Information Processing Systems 33*, 11936–11948.

Stock, J. H. and M. Watson (2009). Forecasting in dynamic factor models subject to structural instability. *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry 173*, 205.

Stock, J. H. and M. W. Watson (2016). Dynamic factor models: A brief retrospective. In *Dynamic Factor Models (Advances in Econometrics, Vol. 35)*. Emerald Group Publishing Limited.

Sundararajan, M. and A. Najmi (2020). The many Shapley values for model explanation. In *International Conference on Machine Learning*, pp. 9269–9278. PMLR.

Sundararajan, M., A. Taly, and Q. Yan (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR.

van de Kamp, R., A. Feelders, and N. Barile (2009). Isotonic classification trees. In *Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31-September 2, 2009. Proceedings 8*, pp. 405–416. Springer.

Wang, F. and C. Rudin (2015). Falling rule lists. In *Artificial Intelligence and Statistics*, pp. 1013–1022. PMLR.

Wen, D., M. He, L. Liu, and Y. Zhang (2022). Forecasting crude oil prices: do technical indicators need economic constraints? *Quantitative Finance 22*(8), 1545–1559.

Yang, Z., A. Zhang, and A. Sudjianto (2021). GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition 120*, 108192.

You, S., D. Ding, K. Canini, J. Pfeifer, and M. Gupta (2017). Deep lattice networks and partial monotonic functions. *Advances in Neural Information Processing Systems 30*.

Zeng, J. (2017). Forecasting aggregates with disaggregate variables: does boosting help to select the most relevant predictors? *Journal of Forecasting 36*(1), 74–90.

Zschech, P., S. Weinzierl, N. Hambauer, S. Zilker, and M. Kraus (2022). GAM (e) changer or not? An evaluation of interpretable machine learning models based on additive model constraints. *arXiv preprint arXiv:2204.09123*.