COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations

Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, Brian Y. Lim
School of Computing, National University of Singapore, Singapore
{ashraf, chris, mohan, brianlim}@comp.nus.edu.sg

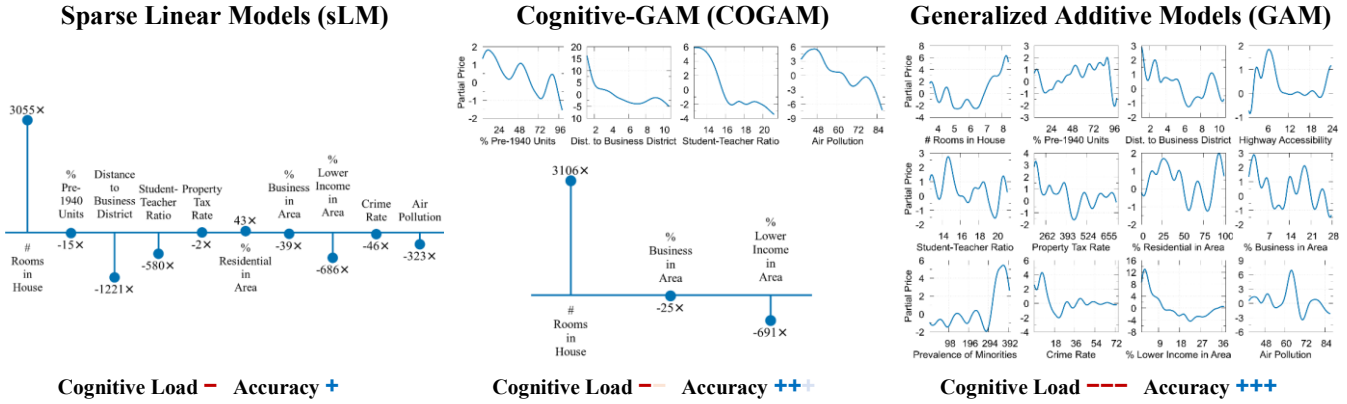

Figure 1. Three explanation visualizations with increasing accuracy at the cost of increasing cognitive load. Left: Sparse Linear Model has lowest cognitive load but lowest accuracy; Right: GAM has highest accuracy but highest cognitive load; Middle: Cognitive-GAM balances cognitive load and accuracy by increasing accuracy with marginal increase in cognitive load.

ABSTRACT

Interpretable machine learning models trade off accuracy for simplicity to make explanations more readable and easier to comprehend. Drawing from cognitive psychology theories in graph comprehension, we formalize readability as visual cognitive chunks to measure and moderate the cognitive load in explanation visualizations. We present Cognitive-GAM (COGAM) to generate explanations with desired cognitive load and accuracy by combining the expressive nonlinear generalized additive models (GAM) with simpler sparse linear models. We calibrated visual cognitive chunks with reading time in a user study, characterized the trade-off between cognitive load and accuracy for four datasets in simulation studies, and evaluated COGAM against baselines with users. We found that COGAM can decrease cognitive load without decreasing accuracy and/or increase accuracy without increasing cognitive load. Our framework and empirical measurement instruments for cognitive load will enable more rigorous assessment of the human interpretability of explainable AI.

CHI 2020, April 25–30, 2020, Honolulu, HI, USA.
 ACM ISBN 978-1-4503-6708-0/20/04...$15.00.
DOI: https://doi.org/10.1145/3313831.3376615

Author Keywords

explanations; explainable artificial intelligence; cognitive load; visual explanations; generalized additive models

CSS Concepts

• **Human-centered computing~Human computer interaction (HCI)**; User studies;

INTRODUCTION

The growing importance of Artificial Intelligence (AI) in society has driven calls for it to be explainable to improve user understanding and trust [3, 17, 22, 67]. Recent research in eXplainable Artificial Intelligence (XAI) has proposed many approaches to help users to interpret inferences from machine learning models [1, 65]. Techniques range from explaining models with linear equations of salient features [40, 47, 53], rules and decision trees [48, 69, 71], partial dependence [31], etc. Reading such explanations can help users to understand model predictions, but this comes at a cognitive cost since users will have to expend effort to cognitively encode and apply these explanations. To support this, XAI researchers typically use sparsity requirements to limit the number of features and complexity of the generated explanations [62, 73]. A popular technique, sparse linear models, achieves sparsity by reducing the number of salient features to be shown, and implies a linear relationship between each feature and the outcome (see Figure 1, left). Sparse linear models (sLM) trade off accuracy for improved readability and explain which features are important by their weights. Machine Learning (ML) researchers typically delegate to application developers to choose the sparsity level, but this is abstract and is a challenge to designers and

developers. Our **first key insight** is to formally quantify the cognitive load of a ML model explanation to provide a practical basis to set its sparsity and complexity to a desired readability level, based on its number of visual chunks.

In contrast to simpler explanations, some researchers have focused on expressive and accurate explanations. Popular techniques such as partial dependence plots [21] and generalized additive models (GAMs) [24] convey nonlinear relationships between each factor independently with the outcome, allowing users to understand in more detail how an outcome will change for a different feature value (see Figure 1, right). Our **second key insight** is that a single nonlinear relationship can be more meaningful and memorable than multiple linear weights, i.e., one nonlinear graph may contain more information than multiple linear factors without requiring higher cognitive load to interpret.

Integrating our two key insights, we propose Cognitive-GAM (COGAM) to generate explanations as a combination of a GAM and a sLM that balances cognitive load and accuracy. We extend the concept of visual chunks from graph comprehension [45, 70] to multiple graphs and visual elements. This lets us quantify cognitive load for COGAM visualized as a grid of nonlinear line and linear bar (lollipop) charts (see Figure 1, middle). In contrast to GAMs and sLMs, COGAM provides a range of solutions with different cognitive loads and accuracies. We claim that selected COGAM solutions can i) decrease cognitive load without decreasing accuracy and/or ii) increase accuracy without increasing cognitive load. We test these claims in simulation and user studies. First, we calibrated linear and nonlinear visual chunks to explanation reading time in a user experiment and confirmed our claims in a simulation study with four datasets. Next, we identify COGAMs that have improved cognitive load and accuracy compared to a full-fledged baseline GAM and a sLM for further evaluation in a user study with 398 users. Thereby, we also developed novel interpretability measures for human simulatability [16, 35] that incorporate a user's mental model. Our contributions are:

- A method to quantitatively measure the cognitive load of interpretable machine learning models in terms of visual chunks along with the calibration of linear and nonlinear representations to reading times.
- An algorithmic approach to moderate cognitive load by regularizing GAMs to generate explanations that can improve cognitive load and/or accuracy.
- Survey instruments to evaluate cognitive load and human simulatability in machine learning model explanations, and evaluation of the trade-off between cognitive load and accuracy of explanations.

This work formalizes a human factor concern for interpretable machine learning, namely cognitive load, and by quantitatively defining the trade-off between cognitive load and accuracy, finds a range of alternative explanations.

Our contributions enable application developers to set explanations for desired cognitive load levels rather than the more abstract sparsity or nonlinear constraints, and thus improve the usability of interpretable machine learning. We also introduce the research approach of quantifying a human factor requirement for eXplainable AI (XAI), supporting selection or optimization for the human objective, and evaluating with simulation and human-subjects studies. Researchers can apply these methods to study other human requirements to improve the human-centeredness of XAI.

BACKGROUND AND RELATED WORK

Developing human-centric XAI requires appreciating the capabilities and limitations of explanation interfaces, understanding how psychological factors constrain or bias user interpretation, and how to assess human perception and cognition to best interpret explanations.

Sparse Linear Models and Generalized Additive Models

Sparse Linear Models (sLMs) are very popular explanation models due to their interpretation as linear equations or weighted averages. Recently, sLMs are commonly visualized as a tornado plot (vertical bar chart) [34, 40, 47] to provide explanations of feature attribution [66]. They are sparse because they focus on the most important features (input variables) that influenced the model prediction and remove less important ones [15, 73]. sLMs provides a very simple explanation method that is not mentally demanding to interpret, but suffers from poor accuracy.

In contrast, Generalized Additive Models (GAMs) [25] are gaining popularity due to their expressiveness to accurately represent the model. They have been used in healthcare [9], to explain black-box models [61], in explanation interfaces for empirical studies [4, 26] and included in recent toolkits to strengthen the support for using GAMs [44, 54]. GAMs treat each feature independently and models them nonlinearly with the prediction. Hence, they are visualized as a grid of line graphs, where each graph shows the nonlinear curve of how the model prediction changes with the feature. These are similar to partial dependence plots and can be used to support counterfactual reasoning [31, 41]. Each graph is interpretable, since the user need only read a line graph. However, this can be overwhelming if many features are included, and thus many graphs are shown.

Identifying this contrast between sLM and GAM regarding readability and accuracy, we developed COGAM as a hybrid explanation to bridge these two approaches and benefit from both improved readability and improved accuracy. We next define readability in terms of cognitive psychology.

Explanation insights from Cognitive Psychology

While much research in XAI has focused on improving explanation faithfulness (accuracy with respect to the prediction model), there is an increasing push to consider human factors such as cognitive psychology and social sciences to develop more human-centric XAI [16, 35, 41, 66]. Lombrozo et al. found that people seek explanations which are causal in nature [36, 41], are plausible [37] and can enable counterfactual reasoning. This is supported with

contrastive (e.g., [6]) and counterfactual (e.g., [49, 64]) explanations. Wang et al. studied human reasoning processes to define an XAI framework to mitigate cognitive biases [66]. The framework provides design guidelines to compose explanations to fit users' reasoning goals but does not explicitly instruct how to tune explanations for a specific human factor requirement. Here, we are interested to improve readability by reducing cognitive load.

Defining explanation complexity in terms of semantic textual chunks, Narayanan et al. [43] found that increase in explanation complexity increases the time taken to understand the explanations. Our approach differs in two ways: 1) we propose an algorithmic approach to automatically generate explanations with varying cognitive loads while they manually tweak the textual chunks; 2) we model cognitive load as visual chunks and calibrate it for line and bar charts explanations that are relevant for sLM and GAM. Next, we elaborate on our approach drawn from graph and visualization literacy literature.

Cognitive Load in Line and Bar Graphs Comprehension

Given the ubiquity of graphs, much has been studied regarding graph comprehension to understand how users perceive, encode, and perform operations with graphs [28, 58]. A user's graph comprehension depends on how well the user can encode the visual information and extract conceptual relationships [46]. This process is affected by the user's innate cognitive abilities and graph literacy, and the complexity of the information. This depends on the format of the graph and differs for line and bar graphs. For example, users frequently describe trends in line graph description tasks, while they describe individual outcomes for bar charts [57, 59]. For line graphs, factors such as the symmetry of the graph curve, local monotonicity and the number of trend reversals, influence graph comprehension times [7, 56, 70]. For bar graphs, the perception of height, position and bar separation affect their comprehension [57, 58, 60]. We draw on these measurable aspects of line and bar graphs to quantify cognitive load of visual chunks in our explanations.

COGAM VISUALIZATION DESIGN

Visualizations of GAM as line charts and simple linear model (sLM) as bar charts is typically used separately, but we combine them in a hybrid visual explanation to balance cognitive load and accuracy (see Figure 1, middle). These explanations visualize how *features* (input variables in the model) relate to the model's prediction outcome. We designed COGAM to visualize explanations as a simplified GAM that is also accurate. We simplify in two ways: 1) by smoothing each nonlinear line graph in a GAM to be less "wiggly" (or curvy), to become straight lines, or even zeroing to remove them; and 2) by further combining all linear line graphs to a single bar chart. This aims to reduce cognitive load by reducing the number of visual chunks due to fewer changes in slopes in each line graph, fewer visual elements with bars instead of line graphs, and fewer features with the removal of zeroed terms. Just like a GAM, COGAM is a global explainer (e.g., [8, 26, 31]), so users need only remember one explanation that applies to any *instance* (use case), unlike instance-based explainers (e.g., [40, 47]) that give different explanations for each instance. Next, we describe how COGAM differs from GAM and sLM.

Compared to GAM visualizations of showing line graphs in a grid (see Figure 1, right), COGAM is simpler and shows fewer line graphs where each line graph is also smoother. COGAM only retains the most informative (accurate) nonlinear relations between a feature and outcome to communicate trends, ranges and extremes, and help to facilitate comparisons between different points to support counterfactual reasoning [26, 33].

In a GAM, linear relationships can be shown as line charts, but for a sLM, linear relationships are typically visualized with bar charts, such as vertical tornado plots [32, 34, 47], where the width or area of a bar indicates the influence or *attribution* ($= w_i x_i$) of a feature x_i. A sLM takes up less visual space as a single bar chart than multiple line graphs. With COGAM we seek to use the space-savings of bar charts, but express the slope w_i of linear line graphs, thus we employ lollipop charts [13]. This communicates the slope as a *multiplication factor* w_i instead of an attribution and retains its suitability as a global explanation. Each lollipop in the chart has the same circle area but different position to represent the weight, so it is less likely to be misinterpreted as attribution. A more influential feature (larger w_i) will have a lollipop further from the axis, but note that its attribution $w_i x_i$ depends on the feature value x_i. Compared to the bar charts of sLM (see Figure 1, left), COGAM can substitute one or more bars to single nonlinear line graphs to communicate more accurate and richer relationships between key features and the outcome, i.e., COGAM can reduce the number of features in an explanation compared to sLM.

After grouping features to be presented as line graphs (top) or lollipops (bottom), we further group features semantically to aid interpretation. In summary, COGAM visualizes nonlinear features using a line graph, linear features using a lollipop chart, and omits unimportant features. Next, we describe our technical approach to simplify GAM to smooth, linearize, or remove features.

TECHNICAL APPROACH

In this section, we describe how we quantify cognitive load, and integrate it algorithmically to develop COGAMs. We have two aims to manage cognitive load in machine learning model (ML) explanations: i) quantitatively define cognitive load such that it can be estimated and measured, and ii) model ML explanations as a hybrid combination of linear and nonlinear expressions that can achieve different cognitive load and accuracy levels.

Cognitive Load as Visual Chunks

Cognitive load is a measure of the amount of cognitive resources required for performing a mental task [28]. We seek to quantify cognitive load as a countable property on

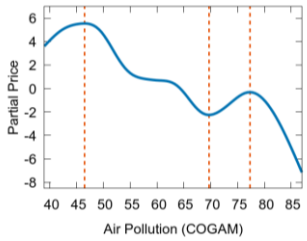

Figure 2. Four visual chunks in a nonlinear line graph separated by changes in trend direction at critical stationary points.

which we can develop a computational model, beyond just a qualitative goal that can be satisfied by iterative interaction design. Work in interpretable ML typically considers a data-centric perspective that making an explanation sparser, with fewer data features, improves its simplicity [16]. Thus, many explanations are generated in terms of the number scores or weights, and data structures (e.g., trees, salient features, words, pixels). In contrast, we consider the human-centric perspective that considers the philosophy and psychology of reasoning and decision making, domain knowledge and expertise, human factors, visualization, UI design, usability, familiarity, etc. [1, 41, 65]. Specifically, we leverage on cognitive psychology [41, 65] to define simplicity in terms of cognitive load due to what a user sees and remembers. Since explanations are often visualized as charts, we focus on human understanding of line and bar charts and draw on research on graph comprehension and visualization literacy [33, 58, 59]. We describe the cognitive load for line charts, each bar in a bar chart, and for the combined visualization.

Trends and Trend Reversals in Line Charts

Based on the Gestalt principle of *continuity*, that people tend to join items into lines and anticipate trends [2, 70], Carswell et al. [7] modeled graph complexity in terms of trends and reversals. They found that graph comprehension time increases with the number of trend changes in a graph, with reversal in trends further increasing the time compared to slow changes in trend [7, 70]. This is because trend reversals create breaks in the visual perception of continuity and results in visual chunking of the line graph at points of trend reversals. Figure 2 illustrates different chunks in a line graph of a nonlinear relationship due to trend reversals around critical stationary points (minima or maxima). Hence, we count the number of sections separated by critical stationary points, i.e., $n_{csp}^{(i)} + 1$. Note that if there are no critical stationary points, there will be one visual chunk; this could be a straight line or a slightly curved line. The number of visual chunks in the *l*ine graph of the *i*th feature is

$$VC_{lg}^{(i)} = \begin{cases} VC_{nl}^{(i)} = n_{csp}^{(i)} + 1 & \text{, for nonlinear graph} \\ VC_{ll}^{(i)} = \alpha & \text{, for linear graph} \end{cases} \quad (1)$$

Since users may perceive chunks in linear line graphs differently from nonlinear chunks (as we confirm in our calibration study later), we introduce the α weighting parameter. Alternatively, we can represent each simple linear relationship as a bar in a lollipop chart. We discuss implications for cognitive load next.

Countable units in Lollipop (Bar) Chart

Since the lollipop chart is a different representation, we consider that each bar has a different cognitive load than a visual chunk in a line graph. Therefore, we denote the visual chunk for the linear *l*ollipop *b*ar of the *i*th feature as

$$VC_{lb}^{(i)} = \beta \quad (2)$$

where β is a weighted factor for the relative cognitive load of lollipop compared to a line graph. Note that $\beta > 0$.

Combined Cognitive Load in COGAM Explanation

The total cognitive load depends on the number of features shown and the representation of each feature. We define the total cognitive load CL for a COGAM visualization as proportional to the sum of visual chunks due to nonlinear and linear lines and linear lollipop representations, i.e.,

$$\begin{aligned} CL &\propto \sum_{i \in NL} VC_{nl}^{(i)} + \sum_{i \in LL} VC_{ll}^{(i)} + \sum_{i \in LB} VC_{lb}^{(i)} \\ &\propto |VC_{nl}| + \alpha |LL| + \beta |LB| \end{aligned} \quad (3)$$

where VC_{nl} is the set of nonlinear visual chunks, NL, LL and LB are the sets of nonlinear, linear line graph and lollipop chunks, respectively, and $|\cdot|$ represents their counts. Eq (3) describes that fewer features and nonlinear chunks will decrease cognitive load. Note that features omitted from the visualization contribute no cognitive load. We assume that users do not spend attention to find the features.

In summary, we model the cognitive load of a visual explanation by counting visual chunks based on the number of trend changes in line graphs Eq (1) and number of lollipops in a lollipop chart Eq (2). The overall cognitive load of the combined explanation is a weighted sum of visual chunks of trend changes and lollipops Eq (3). We calibrate the cognitive load CL to a memorization-based reading task time and estimate α and β in a calibration user study that we describe later. Next, we describe the technical approach to moderate model explanations with cognitive load.

Cognitive Generalized Additive Model (COGAM)

COGAM simplifies GAM by selectively smoothing "wiggly" curves, linearizing or removing features to reduce cognitive load while limiting the loss in explanation accuracy. In this section, we provide technical details to 1) describe the functional form of GAM with a *regularization* term to control for smoothness 2) discuss how *sparsity* (removing features) and *partial linearity* (making some features linear) affect accuracy and cognitive load and 3) describe our algorithm to find COGAM solutions using a greedy approach to select which features become sparse or linear, in addition to regularization to control for smoothness.

Generalized Additive Model (GAM)

Generalized Additive Model (GAM) [24] models the relationship between an outcome Y and features as a sum of shape functions of each feature. A GAM is represented as

$$g(E(Y)) = f_0 + \sum_{i=1}^{d} f_i(x^{(i)}) \quad (4)$$

where f_0 is the intercept term, f_i is the shape function modeling the nonlinear relationship between the ith feature $x^{(i)}$ and the outcome, and d is the number of features. Each shape function, also called *term fit*, can be modeled as a spline, line, or categorical factor. Each shape function is often visualized with a line graph and the overall outcome is a simple addition of the individual partial outcome for each feature. Hence, GAMs are often considered to be intelligible [9, 26, 39] for a wide variety of applications. Each feature $x^{(i)}$ is treated independently, so f_i does not depend on any other feature. Estimating a GAM involves minimizing the objective function over all instances j for a given dataset.

$$\text{minimize} \sum_j^n \left(y_i - \sum_i^d f_i\left(x_j^{(i)}\right)\right)^2 + \lambda \sum_i^d \int f_i''(t)^2 \, dt \quad (5)$$

where λ represents the weight for the regularization term to control the smoothness of shape functions by penalizing their "wiggliness" as measured by the total change in magnitude of the slope. Lower λ leads to better accuracy, but wigglier shape functions to the data; while higher λ leads to smoother shape functions, but with lower accuracy. There are many approaches to estimating a GAM and we point the interested reader to Binder and Tutz [5].

GAM with Sparsity and Partial Linearity

The basic GAM formulation only controls the wiggliness of shape functions. However, it is often desirable to set some shape functions to linear (*partial linearity*) when those features exhibit linear relationships or set some shape functions to zero (*sparsity*) when they are relatively poor predictors. As a result, recent GAM techniques such as GAMSEL [10] and SPLAM [38] have formulated new optimization objectives to induce sparsity and partial linearity to find one GAM solution that has the best accuracy. Since our focus is not just on model accuracy, but also on cognitive load, we fit GAMs for multiple combinations of sparse, linear and nonlinear term fits to get solutions with varying cognitive load and accuracies. Since fitting all possible combinations may be computationally intractable depending on the number of features, we use a greedy feature selection approach which is described next.

Greedy Search for COGAM Solutions

We define a COGAM solution as a GAM model trained such that it contains a set of features L with linear term fits, a set of features NL with nonlinear term fits and a set of features Z that are zeroed (excluded from the model), such that $|Z| + |L| + |NL| = d$. For each combination of term fits (e.g., nonlinear spline, linear line, zero), we first select active features $\notin Z$ in decreasing order of *mutual information score* between a feature and the outcome (i.e., how well a feature predicts the outcome). Features with lowest scores are zeroed first. For the features that are active, we select which features are modelled as linear lines based on decreasing order of the R^2 score between each feature and the outcome (i.e., higher R^2 suggests a more linear relationship). The remaining features are modelled as nonlinear splines. We estimate the GAM at each combination of term fits using Penalized Iteratively Reweighted Least Squares [68] with an L2 penalty on the mean square error (MSE) to regularize linear features and a penalty on the second derivative of the shape function f'' to regularize nonlinear features. While iterating over the combination of term fits, as the number of active terms in a model decreases and/or as more features are made linear, the spline fit for nonlinear features can become wigglier. This is because they tend to overfit to make up for the loss of accuracy because of the features that were excluded or made linear. To mitigate this overfitting, we progressively increase the lower bound of the λ search grid based on the best λ selected at the previous iteration step. The pseudocode for the algorithm is provided in the supplementary material.

Note that since the combinations of nonlinear, linear, and zero term fits are discrete and the degree of nonlinearity of the the features are dependent on the dataset, there may be no COGAM solutions that have high accuracy and low cognitive load. We carefully study their distribution in a simulation study described in the next section.

SIMULATION STUDY ON COGNITIVE LOAD ACCURACY TRADE-OFF WITH FOUR DATASETS

Since COGAM finds many explanation models with varying cognitive load - accuracy trade-offs, we conducted a simulation study to characterize the COGAM solutions across four diverse datasets (Boston Housing [23], Ames Housing [11], Concrete Strength [72], and Wine Quality [12]). We selected *regression* problems (predicting continuous numeric values) instead of classification, since and they are easier for users to reason with when viewing line graphs and to give most flexibility to our analysis. For each explanation model, we calculate accuracy as the R^2 score on a *held-out* (excluded from training) test dataset that indicates how well the explainer's prediction matches the true value. We calculate cognitive load as described in Eq(3). For simplicity, we excluded categorical features, because they would be one-hot encoded or discretized (splitting an n-level categorical variable into n binary 0 or 1 variables), which is equivalent to assigning each value to a lollipop bar.

Figure 3 compares the visual chunk (top) and calibrated cognitive load (bottom) with accuracy of different explanation models and illustrates the spread of COGAM solutions that trade off cognitive load and accuracy variously. More accurate models are towards the right and simpler models are towards the bottom. Explanations with the best characteristics — low cognitive load and high accuracy — are towards the bottom-right of the graphs. In general, as the accuracy increases so does the number of visual chunks (diagonal trend with positive correlation), because COGAM can take advantage of the inherent sparsity and partial linearity in the dataset to find solutions with lower number of visual chunks. As expected, we found that GAMs have higher accuracy than (sparse) linear models (sLM & LM) but also higher cognitive load. We expect COGAM

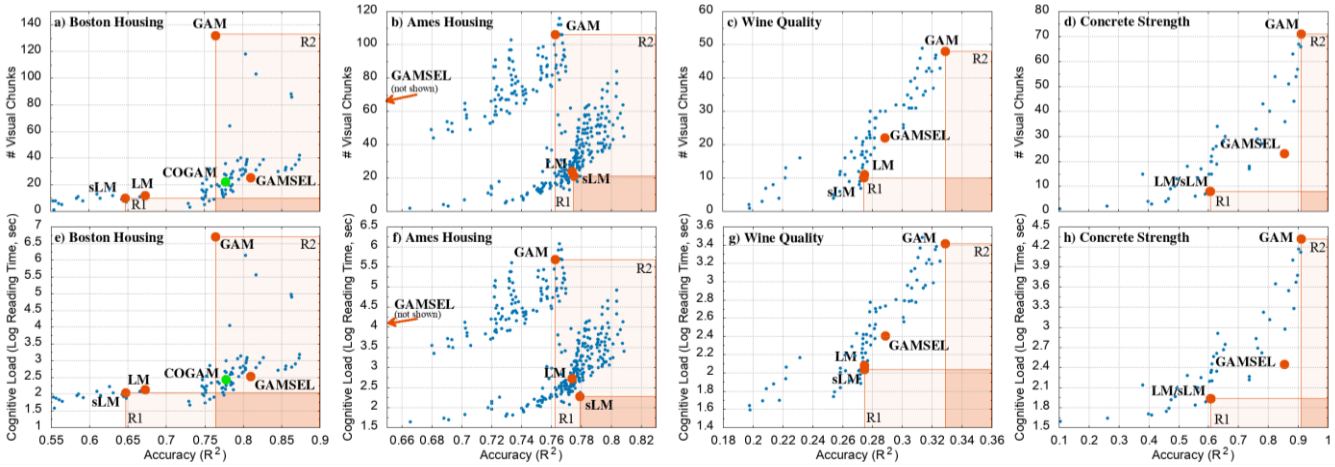

Figure 3. Simulation results on four datasets showing the trade-off between cognitive load (top: as number of visual chunks; bottom: as calibrated reading time) and accuracy (regression R^2) for different explanation models. Each COGAM is denoted by a blue dot. Baselines LM, sLM, GAM and GAMSEL are denoted by red dots. We selected the COGAM denoted by the green dot (e) for the user study.

solutions to mainly fall in between these points of reference. For some datasets, COGAMs compromise accuracy for cognitive load (diagonal trend interpolated between sLM and GAM, e.g., in Figure 3c); for others, remarkably, COGAMs can have *both* high accuracy and high cognitive load, i.e., accuracy as high as GAM and cognitive load as low as sLM (reversed "L" shape of plots in Figure 3a, d). Curiously, for the Ames Housing dataset (Figure 3b), sLM accuracy is slightly higher than GAM; this suggests that many features are naturally linear in relation to the outcome. Furthermore, some COGAM explanations have lower cognitive load at similar accuracy than sLM; this suggests some features in sLM were accommodated by fewer nonlinear functions. We verified these by looking at the COGAM visualizations.

After examining the spread of COGAMs, we can select one solution that has the best trade-off for cognitive load and accuracy. We highlight two regions of interest in which selecting a COGAM can provide better cognitive load and accuracy than either GAM or sLM:

- R1) Any COGAM here will have the same or higher accuracy, but no higher cognitive load, than sLM.
- R2) Any COGAM here will have the same or lower cognitive load, but no lower accuracy than the GAM.

Ideally, a good choice of COGAM will be in the region overlapped by R1 and R2, but such an explanation model may not exist, selecting a viable explanation nearby is reasonable. We also plot GAMSEL among COGAMs for reference, since it also optimizes for sparsity and smoothness with accuracy. However, since GAMSEL does not explicitly model cognitive load, it does not necessarily produce an explanation that best balances cognitive load and accuracy.

In summary, we found that GAMs have higher accuracy but higher cognitive load than sLM, and that there are many COGAMs with intermediate accuracy and cognitive load. Depending on dataset, COGAM can find explanations that reduce cognitive load by compromising acuracy, and find explanations that have cognitive load comparable to sLM and accuracy comparable to GAM. These theoretical results are based on real data, and we further validate the findings and benefits with calibration and user evaluation studies.

EVALUATING COGAM READABILITY AND USEFULNESS

We evaluate COGAM in two steps. First we calibrate the cognitive load of a COGAM to the time taken by a user to read its visual explanation by testing multiple COGAM solutions with different number and types of visual chunks. Second, we use the calibrated cognitive load to select one COGAM (see green dot Boston Housing Figure 3e) and evaluate it against baselines sLM and GAM in terms of cognitive load (to interpret the explanations) and *human simulatability* (measuring human understanding of machine learning model explanations). The calibration and comparison studies share many common details such as the dataset (Boston Housing), between-subjects design, survey instruments, and statistical analysis methods to retain internal validity. We performed two iterations of pilot testing with HCI researchers and on MTurk which resulted in clarification of instructions, adding controls, simplified UI and reduction of questions per task. Next, we describe the final experiment method, procedure, design, and results.

Experiment Method: User Tasks and Survey Instruments

A user is required to study an explanation visualization about a machine learning (ML) model, specifically, housing price prediction on the Boston Housing dataset and demonstrate their understanding through human simulatability tasks. We implemented survey questions and usage log instruments to measure and control factors affecting the users' ability to read, understand, and apply the visual explanations. Next, we describe experiment tests and tasks.

Visual Literacy Assessment Tests (VLAT). As discussed in related work on graph comprehension theories, the reading time for a graph depends on the number of visual chunks. However, it is also confounded with the user's graph and visualization literacy; poorer literacy would lead to longer reading time. Hence, we extend VLAT that tests literacy on

interpretation tasks of various visualizations [33] to multi-chart interfaces with line and lollipop charts by including questions about multiple chart comparisons and multivariate counterfactual reasoning, which are relevant human simulatability tasks for XAI models [16, 41].

Memory Tests. Since a COGAM explanation visualization contains multiple line and bar charts, the time to read or memorize the charts may depend on the user's memory capacity to remember a span of items and visual spatial ability to remember shapes in individual line charts. Hence, we use the Digit Span Test and Visual Pattern Test that measure the person's verbal and visual-spatial memory capacity, respectively [14, 50].

Reading and Memorization Task. This is the **key task** to measure the user's cognitive load to read and study the explanation to a nontrivial level of understanding. We prime the user to study (and memorize) the visualization and warn that it would not be shown when answering subsequent questions. This approach is similar to prior graph comprehension studies that had participants memorize single charts to draw them from memory later [7]. Our task is notably more challenging, since users must study multiple charts. The amount of time spent on this page before clicking next is recorded as the reading time. The user is then asked to rate the mental effort to read the visualization on a 7-point Likert scale on the next page [51].

Recall Task. This measures how well users can retrieve, from memory, details of feature-outcome relationships visualized in the explanation without any prompts. It involves the user, from memory, drawing each nonlinear relationship on an empty line chart canvas, and specifying the multiplier influence of each linear feature.

Recognition Task. This measures a user's ability to recognize a feature when they are shown a chart drawing indicating the feature-outcome relationship (without label). It involves matching the correct feature names to line graph or lollipop charts indicating the feature-outcome relationship.

Description Task. Users see their recalled drawings or multiplier influence values and describe the relationship between the feature and outcome. This provides contextual information to help us to understand which characteristics of each chart is salient or memorable.

Counterfactual Task. Users get to view the explanation visualization again and answer multiple questions of the type, "If Factor X changes from a to b, then by how much does the Outcome increase (or decrease)?" These questions measure the user's ability to apply the explanations to new situations. They require deeper understanding than the common approach of reading explanations to simulate the outcome of a one input value [16, 43].

Post Survey Questionnaire. Using Likert scale and free text questions, this asks about overall *ease of use* and *helpfulness* of the explanation as well as the charts for specific features.

Procedure

Each participant in any given condition was exposed to only one explanation visualization. In the calibration study, one explanation visualization has a fixed number and types of visual chunks, whereas for the comparison study an explanation visualization is generated from the selected COGAM or GAM, sLM. All participants went through the following procedure.

1. Introduction and accepting consent form.
2. Tutorial to give a refresher on reading line and bar (lollipop) charts and comprehension test (adapted VLAT [33]). Participants who scored ≥80% qualify to continue.
3. Memory tests starting with digit span test from 3 digits up to 9, followed by visual patterns test from 3×3 grid, 3×4, and so on, up to 5×5. Users continue to the main survey regardless of levels achieved.
4. Reading and memorization task.
5. Recall task, as first task to reduce forgetting.
6. Recognition task, after recall to avoid priming.
7. Counterfactual task, along with explanation visualization being shown again.
8. Post survey questionnaire.

Since asking about all features in any explanation can lead to user fatigue, we limit questions on tasks 4 through 7 to four features. The four features were selected from the explanation such that two were fixed across all treatment conditions, and the remaining were randomly selected where one was nonlinear and the other linear where applicable. This avoids biasing results based on the features being tested.

Measures for Cognitive Load

In addition to *reading timing* and the *self-reported perceived load*, we assess cognitive load with memory performance, i.e., how well the user remembers the explanation. Better memory scores indicate lower cognitive load, since the user can retain more information about the explanation, given the same reading time. We score memory based on measuring the user's mental model (recall reconstruction) and recognition performance.

Recall Reconstruction Score

We seek to capture how accurately users recall the relationships in the explanations that they saw. Ideally, a suitable metric would be compatible for both line chart drawings and slider displacements for the lollipop bars. However, potential methods like relative difference depend on the user's and the explanation's absolute values, and this will be difficult to scale for these different representations.

Instead, we reconstruct the explanation from the user's drawing and lollipop positioning as a "mental" model in the format of GAM, COGAM, or sLM, but with different parameters. To do this, we reconstruct shape functions based on the users recalled linear and nonlinear relationships. To reconstruct the shape function for a line graph, we perform a univariate spline interpolation. To reconstruct the effect of a lollipop graph, we add a linear term weighted by the user recalled multiplier influence. This model can be considered as "trained" from the user's belief and memory recall, and

can be used similarly as the explainer model to infer outcomes from instance data. Hence, we use a held-out test set over which we compute the user's partial predictions. Note that, since the models are additive, we can treat the prediction as if the rest of the features were missing.

Finally, to determine how well the user's mental model matches the shown explanation model, we compute the R^2 score between the user's mental model partial predictions and the partial predictions of the explainer model. The recall reconstruction is this R^2 score between the user and explainer. Note that this is similar to *human simulatability* in interpretable machine learning [16, 35] that seeks to measure the similarity between the user's and explainer's inference, but those works do not compute a user's mental model.

Recognition Score

The recognition score measures how well the user can correctly recognize and associate the name of a feature and its shape as seen in a chart or its multiplier influence seen in a lollipop bar. This is a binary score of 1 for every correct match, and the sum is the total Recognition score.

Measures for Mental Model Faithfulness (Accuracy)

We seek to access how accurately the user's mental model can be applied to new situations. In interpretable ML, explanation model faithfulness is defined as the similarity in prediction between the explainer and predictor or ground truth. Here, we define *mental model faithfulness* as the similarity in prediction between the user's mental model and predictor or ground truth. Therefore, mental model faithfulness increases if the user better understands how the predictive model or ground truth phenomenon works. We compute mental model faithfulness in terms of counterfactual performance and recall faithfulness.

Counterfactual Performance

We measure counterfactual performance as the relative difference between the user's answer and the model. In addition to measuring how accurate the answers are, we also time the counterfactual question for each feature tested to assess the difficulty of reading the explanation.

Recall Faithfulness

To compute recall faithfulness, we leverage on most of the method to compute recall reconstruction. However, instead of comparing the user's mental model with the explainer model, we compare the user's mental model with the ground truth labels. This measures how similar to the ground truth the user could predict. Therefore, the recall faithfulness is this R^2 score between the user and ground truth.

Pre-Study: Calibrating Cognitive Load

Since COGAM employs two different visualization elements (line charts and lollipop bars), the measured cognitive load may differ between the types of visual chunks Eq (3). Hence, we conducted a study to calibrate their respective cognitive loads. Although we define cognitive load in terms of visual chunks, it is typically measured subjectively from survey questions (e.g., Paas scale [51]), or objectively using sensors (e.g., electrodermal activity, EEG) or with task performance time. We chose to calibrate against time taken to read or memorize a visual explanation with different visual chunks, since it is an objective measure and we can measure this with an instrumented online study without expensive sensor equipment. We study the cognitive load for multiple variants of COGAM to provide many readings across a range of visual chunk combinations for a more accurate estimation.

Treatment Conditions. Each participant viewed one of 27 variants of COGAM based on three independent variables: the number of non-zeroed features (3 levels: 4, 8, 12) and the percentage of features that are nonlinear (5 levels: 0, 25, 50, 75, 100%), and representation of linear features (2 levels: straight line in a graph or lollipop bar). The latter two independent variables enable us to calibrate α, β in Eq (3).

Participants. We recruited participants from Amazon Mechanical Turk. To ensure reliable results, we targeted workers with high qualifications (>5000 completed HITs with >97% approval rate). Participants were compensated US$2.50 and completed the survey in about 30 to 40 minutes. While 1318 participants attempted the survey, only 305 (151 Female, average age 39) participants passed the screening. We discuss implications of our strong screening later.

Calibrating Visual Chunks with Cognitive Load Time

Our calibration study results verify that the reading time for explanation increases with the number of line graphs and the complexity of the nonlinear relationships. We fit a multivariate linear model ($R^2 = 0.34, R^2_{adj} = 0.30$) with a log transform of the cognitive load time $log\ (t_{CL})$ as dependent variable; number of nonlinear/linear line graphs $|NL|$, $|LL|$, total number of nonlinear visual chunks $|VC_{nl}|$, and number of lollipop bars $|LB|$ as independent variables; and participant age and Recall accuracy as control variables:

$$\log\ (t_{CL}) = w_1|NL| + w_2|VC_{nl}| + w_3|LL| + w_4|LB| + \varepsilon \quad (6)$$

where $w_1 = 0.385$, $w_2 = -0.073$, $w_3 = 0.071$, $w_4 = 0.049$, ε represents terms from control variables and intercept. All IVs were very significant (p<.01) and CVs were significant (p<.05). In our calibration study, we had a mostly fixed number of nonlinear visual chunks for the number of nonlinear line graphs, so we substitute with $|NL| = \kappa|VC_{nl}|$ where $\kappa = 0.293$ is estimated from a linear regression (p<.0001). Formatting Eq (6) as in Eq (3), we get

$$\log\ (t_{CL}) = 0.039|NL| + 0.071\ |LL| + 0.048\ |LB| + \varepsilon \quad (7)$$

Interestingly, the cognitive load due to nonlinear chunks is lower than for linear chunks in line graphs; this suggests that users mentally process multiple nonlinear chunks in batches. Normalizing Eq (7), we calibrate parameters in Eq (3) as $\alpha = 1.83, \beta = 0.686$. Since $\beta < \alpha$, this means lollipop chunks are faster to read than linear line graphs, validating our choice to visualize linear relationships with lollipop bars.

Using our calibration result, we can estimate the cognitive load for different COGAMs for different datasets (see Figure

3, bottom) to help us select a COGAM for a desired cognitive load level. Next, we evaluate how a carefully chosen COGAM can help users more accurately understand the model performance without much increase in cognitive load, compared to sLM and GAM baselines.

Main Study: Comparing COGAM vs baselines

We evaluate how theoretical improvements calculated from COGAM leads to measurable decrease in cognitive load and increase in mental model understanding compared to baseline models sLM and GAM. As mentioned earlier, all models are trained on the Boston Housing dataset to predict housing prices based on 12 features (see Figure 1). This is an intuitive lay problem to simplify the task for participants. We chose a COGAM (see green dot in Figure 3e) that has comparable accuracy to GAM and comparable cognitive load to sLM to evaluate the best circumstances for COGAM. This COGAM implements linear terms as lollipops. Based on our simulation results, we hypothesize that users will experience cognitive load in the following order: sLM $\leq$ COGAM $<$ GAM (H1); and understand the model prediction in the following order: sLM $<$ COGAM $\leq$ GAM (H2). *Treatment Conditions.* We presented each participant with one of three explanations (3 levels: sLM, COGAM, GAM).

Participants. We recruited participants from MTurk with the same requirements as the calibration study. 1858 participants attempted the survey but only 398 participants (201 female, average age 38) passed the VLAT screening test. The low pass rate ensures participants were sufficiently skilled for the interpretation tasks. Though there is a risk for cognitive saturation, our pilot testing iteration and ultimate high ratings (>4) for Ease of Understanding (Figure 3b) suggest this was unlikely. Participants were reimbursed US$2.50. There were 130 sLM, 131 COGAM, and 120 GAM participants.

Analysis and Findings

We fit a linear mixed effects models on each of the aforementioned dependent variables with Model type as fixed effect. We found significant (p<.001) fixed effects for Log(Reading Time), Ease of Understanding, Recall and Counterfactual scores. We performed post-hoc Tukey HSD testing with Bonferroni correction (at α=.005) to identify specific differences. Figure 4 shows significant results for key dependent variables. Next, we discuss specific findings and supplement with qualitative participant comments.

COGAM is faster to read and more memorable than GAM and is similar to sLM. A Tukey HSD test on Log(Reading Time) found that (sLM=COGAM) <GAM, p=.0043. However, there was no significant effect on Self-reported Cognitive Load (M=5.77), perhaps because participants found all explanations somewhat demanding to study (ceiling effect) due to seldom reading charts, and not be able to compare explanations relatively within-subjects. A Tukey HSD test on Recall Reconstruction (higher is better) found that (sLM = COGAM) > GAM, p<.0001. Some participants struggled to remember GAMs due to their high number of line graphs with nonlinear terms, e.g., *"Some factors were hard to remember because they didn't seem much different from random noise. There was no trend in student-teacher ratio from what I remembered."* [P39], and because recalling is much harder than reading, e.g., *"... the charts were easy to understand when I was looking at them...but when I wasn't it was so hard for me. I felt really disadvantaged when I was having to remember the charts."* [P66]. There was no significant effect on Recognition scores which were equally high for all explanations (M=3.16/4), indicating that recognizing explanations is much easier than recalling them.

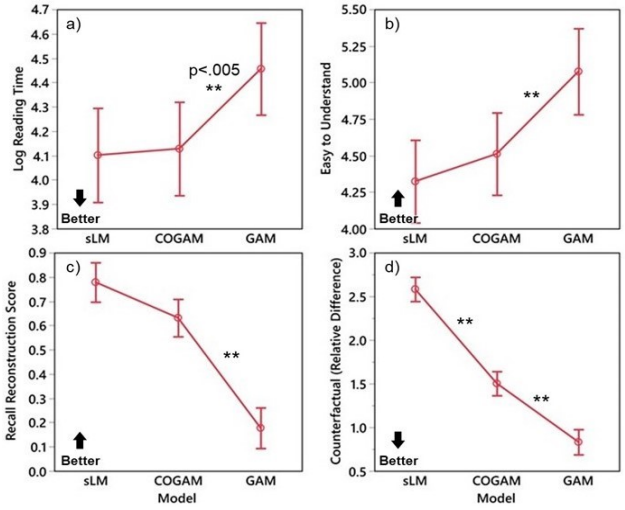

Figure 4. Comparison results showing how COGAM imposes intermediate cognitive load (a, b, c) and supports intermediate mental model faithfulness (d) with respect to sLM and GAM.

Visualizing sLM as Lollipop bar charts does not provide as accurately useful explanations and is surprisingly hard to understand compared to GAMs. Since COGAM is a mix of the two, its ease of understanding and usefulness for counterfactual reasoning are in between. A Tukey HSD test on Counterfactual difference (lower is better) found that sLM>COGAM>GAM, p<.0001, in agreement with H2. Furthermore, a Tukey HSD test on Ease of Understanding found (sLM=COGAM)>GAM, p<.0001. This suggests that there could be more than the lack of expressiveness of sLM that impedes the users' mental model faithfulness. As expected, users could easily identify which features were important with sLM, *"Seeing the lines* [bars] *made it more intuitive to understand how the factors influenced the price and seeing the length of the lines gave a greater impression of the magnitude of that influence"* [P259]. However, some users found it difficult to mentally compute counterfactuals from lollipop charts for e.g., P121 who used COGAM felt that *"the multipliers relied on me doing a large portion of the math and did not have enough information to really figure it out"*. This was exacerbated with negative weights: "*The negative values on the bar charts were somewhat confusing because by "increasing" meant an increase in the negative direction*" [P119]. In contrast, some users found GAMs easier to use for the counterfactual task, e.g., *"I think that the line charts were easier for me to understand,*

because I felt like I could find outcome changes easier. I was able to see the trend going up or down easier and understand it better" [P176], *"The line charts were easy to understand. They clearly show how a home value is affected by the factor"* [P127]. Given the need of explanations for counterfactual reasoning [41], our results highlight the importance of employing line graphs as the appropriate representation, instead of the common bar charts.

In summary, our comparison study found that users could a) read COGAM as quickly as sLM, b) found it as easy to understand as sLM, and c) could recall the explanation of COGAM as well as sLM, but d) had counterfactual understanding in-between sLM and GAM. Surprisingly, sLM was least easy to understand subjectively and objectively for deeper counterfactual reasoning.

DESIGN IMPLICATIONS AND FUTURE WORK

We have developed COGAM to moderate cognitive load in explainable AI as visual chunks, which goes beyond only considering the sparsity and faithfulness of explanations. We discuss how our methods and results open opportunities for further research in studying various human factors for XAI.

Moderating cognitive load can maintain user's learned mental model from explanations. Our results show that quantifying cognitive load to moderately reduce it in explanations helps users to better understand the model and its predictions. Explanations with lower cognitive load improve the accuracy of the user's mental model reflected in improved recall and shorter reading times. This indicates that simplifying explanations to manage cognitive load does not necessarily reduce the knowledge gained about the model when users read the explanations.

Math and visual literacy requirements for XAI. Our rigorous graph literacy tests with low pass rate suggests most lay users are not sufficiently qualified to effectively interpret model explanations. We target early AI users who are typically more tech savvy and more likely to pass our VLAT test. Some users had difficulty to apply lollipop (bar) charts for counterfactual reasoning. Thus, there is a strong need to simplify explanations for lay users. Less technical explanations (e.g. simple text) could be provided which reduces selection bias, but this may compromise accuracy.

Tunable explanations for varying trade-offs. Our design of COGAM mixed two explanation representations, simple linear model and GAMs, that are typically considered separately. By introducing a tuning parameter for their combination, COGAM considers a spectrum between linear and nonlinear explanations and enables tunable explanations to trade off between cognitive load and accuracy. This provides more flexibility for XAI developers to tune their explanation complexity based on the domain requirements. For example, music listeners may prefer simpler but less accurate explanations for automatic song recommendations, while medical doctors may require more detailed and accurate explanations with higher cognitive load for an AI-assisted medical decision-making task. In general, with our hybrid and tunable approach, explanations can be explored on a continuum instead of limited binary or discrete alternatives. Other trade-offs to study include privacy-interpretability, plausibility-faithfulness, etc.

Generalizing to other sources of cognitive load in XAI. In our study of moderating and measuring cognitive load, we focused on visual explanations to allow us to formally define cognitive load in terms of quantifiable visual chunks. We studied simple GAM visualizations with line charts, but GAM can also include confidence bands. This adds two more lines, which will increase cognitive load due to perceiving more chunks and reasoning with uncertainty. GA^2M [9] is an extension of GAM that visualizes using interaction plots with colored saliency heatmaps; modeling cognitive load for this has to consider perception of 2D areas and small color differences. Furthermore, while visual explanations are widely used [29, 40, 42, 47], there is growing interest in verbal explanations using text [18, 27, 55]. Our formal approach can be extended to other explanation modalities, such as text by drawing from text comprehension theories [19, 30, 52] to quantify textual chunks based on sentence or semantic boundaries and incorporate appropriate controls such as the Verbal Ability Test and Verbal Span Test.

Theory-driven integration of human factors in XAI. Our grounding in cognitive psychology with visual chunks allowed us to parameterize, optimize, calibrate and measure cognitive load in explanations of AI. This is an example of drawing from understanding human reasoning to inform XAI design [65]. First, by identifying a human factor or usability issue (in our case, cognitive load) and studying its constituent components as relevant to explanations (graph and visual chunks), we can explicitly model a new requirement for XAI. Second, our explicit mathematical definition allowed us to objectively simulate the outcome of providing for the new requirement. Third, beyond using standard measures of explanation effectiveness, we define new measures to specifically evaluate explanations with respect to the targeted explanation quality (in our case, cognitive load). Our three-step method – 1) mathematical modeling, 2) simulation and 3) evaluation with users – provides a theoretical framework to help researchers to develop and evaluate human-centric XAI. Our method can be applied to other factors, such as, plausibility, memorability [20, 37, 63].

CONCLUSION

Explanations with lower cognitive load help users form more accurate mental models. We formalize cognitive load for XAI with visual chunks and present Cognitive-GAM (COGAM) to generate explanations for desired cognitive load and accuracy. Specifically, we combine the expressive nonlinear generalized additive models (GAM) with simpler sparse linear models, and moderate sparsity and nonlinearity. In doing so, we characterize the trade-off between cognitive load and accuracy, and find that COGAM improves cognitive load without sacrificing accuracy and vice versa.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April. https://doi.org/10.1145/3173574.3174156

[2] Nadia Ali and David Peebles. 2013. The effect of Gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Human Factors* 55, 1: 183–203. https://doi.org/10.1177/0018720812452592

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300233

[4] Niels V A N Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning : A Recidivism Case Study. 1, 1.

[5] Harald Binder and Gerhard Tutz. 2008. A comparison of methods for the fitting of generalized additive models. *Statistics and Computing* 18, 1: 87–99. https://doi.org/10.1007/s11222-007-9040-0

[6] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. https://doi.org/10.1145/3301275.3302289

[7] C. Melody Carswell, Cathy Emery, and Andrea M. Lonon. 1993. Stimulus complexity and information integration in the spontaneous interpretations of line graphs. *Applied Cognitive Psychology* 7, 4: 341–357. https://doi.org/10.1002/acp.2350070407

[8] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare. https://doi.org/10.1145/2783258.2788613

[9] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noémie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2783258.2788613

[10] Alexandra Chouldechova and Trevor Hastie. 2015. Generalized Additive Model Selection. 1–24. Retrieved from http://arxiv.org/abs/1506.03850

[11] Dean De Cock. 2011. Ames , Iowa : Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education*.

[12] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2009.05.016

[13] Andy Cotgreave. Lollipop Charts. Retrieved May 1, 2019 from https://gravyanecdote.com/andy-cotgreave/lollipop-charts/

[14] Meredyth Daneman and Patricia A. Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*. https://doi.org/10.1016/s0022-5371(80)90312-6

[15] Robyn M. Dawes. 1979. The robust beauty of improper linear models in decision making. *American Psychologist*. https://doi.org/10.1037/0003-066X.34.7.571

[16] Finale Doshi-velez and Been Kim. 2017. A Roadmap for a Rigorous Science of Interpretability. 1–13. https://doi.org/10.1016/j.intell.2013.05.008

[17] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Christopher Bavitz, Samuel J. Gershman, David O'Brien, Stuart Shieber, Jim Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3064761

[18] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. https://doi.org/10.1145/3301275.3302316

[19] Charles R. Fletcher. 1981. Short-term memory processes in text comprehension. *Journal of Verbal Learning and Verbal Behavior*. https://doi.org/10.1016/S0022-5371(81)90183-3

[20] Meadhbh I. Foster and Mark T. Keane. 2019. The Role of Surprise in Learning: Different Surprising Outcomes Affect Memorability Differentially. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12392

[21] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*.

[22] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*. https://doi.org/10.1609/aimag.v40i2.2850

[23] David Harrison and Daniel L. Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*. https://doi.org/10.1016/0095-0696(78)90006-2

[24] Trevor Hastie and Robert Tibshirani. 1986. Generalized additive models. *Statistical Science* 1, 3: 297–310. https://doi.org/10.1214/ss/1177013604

[25] Trevor Hastie and Robert Tibshirani. 1986. Generalized additive models. *Statistical Science*. https://doi.org/10.1214/ss/1177013604

[26] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300809

[27] Fred Hohman, Arjun Srinivasan, and Steven M Drucker. 2019. TELEGAM: Combining Visualization and Verbalization for Interpretable Machine Learning. In *IEEE Visualization Conference (VIS)*.

[28] Weidong Huang, Peter Eades, and Seok Hee Hong. 2009. Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization* 8, 3: 139–152. https://doi.org/10.1057/ivs.2009.10

[29] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2018. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1: 88–97. https://doi.org/10.1109/TVCG.2017.2744718

[30] Walter Kintsch and Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*. https://doi.org/10.1037/0033-295X.85.5.363

[31] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2858036.2858529

[32] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces. ACM*, 126–137.

[33] Sukwon Lee, Sung Hee Kim, and Bum Chul Kwon. 2017. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2016.2598920

[34] Brian Y. Lim and Anind K. Dey. 2012. Weights of evidence for intelligible smart environments. In *UbiComp'12 - Proceedings of the 2012 ACM Conference on Ubiquitous Computing*.

[35] Zachary C. Lipton. 2018. The mythos of model interpretability. *Communications of the ACM 61*, 35–43. https://doi.org/10.1145/3233231

[36] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences (Elsevier)* 10, 10: 464–470.

[37] Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive Psychology*. https://doi.org/10.1016/j.cogpsych.2006.09.006

[38] Yin Lou, Jacob Bien, Rich Caruana, and Johannes Gehrke. 2016. Sparse Partially Linear Additive Models. *Journal of Computational and Graphical Statistics* 25, 4: 1126–1140. https://doi.org/10.1080/10618600.2015.1089775

[39] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2339530.2339556

[40] Scott M. Lundberg and Su In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*.

[41] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[42] Yao Ming, Huamin Qu, and Enrico Bertini. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2018.2864812

[43] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. 1–21. Retrieved from http://arxiv.org/abs/1802.00682

[44] Rich Nori, Harsha and Jenkins, Samuel and Koch, Paul and Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223*.

[45] Unaizah Hanum Binti Obaidellah. 2012. The role of chunking and schemas in learning and drawing. *PQDT - UK & Ireland*: 1. Retrieved from http://search.proquest.com/docview/1442498898?accountid=10673%5Cnhttp://openurl.ac.uk/athens:_edu?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:ProQuest+Dissertations+%26+Theses+Global&atitle=&ti

[46] Steven Pinker. 1990. A theory of graph comprehension. *Artificial intelligence and the future of testing*. https://doi.org/10.1145/2046684.2046699

[47] M T Ribeiro and S.Singh and C. Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier,. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*: 1527–1535.

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.

[50] Sergio Della Sala, Colin Gray, Alan Baddeley, Nadia Allamano, and Lindsey Wilson. 1999. Pattern span: A tool for unwelding visuo-spatial memory. *Neuropsychologia*. https://doi.org/10.1016/S0028-3932(98)00159-6

[51] Annett Schmeck, Maria Opfermann, Tamara van Gog, Fred Paas, and Detlev Leutner. 2015. Measuring cognitive load with subjective rating scales during problem solving: differences between immediate and delayed ratings. *Instructional Science* 43, 1: 93–114. https://doi.org/10.1007/s11251-014-9328-3

[52] Kilian G. Seeber. 2011. Cognitive load in simultaneous interpreting: Existing theories — new models. *InterpretingInterpreting International Journal of Research and Practice in Interpreting*. https://doi.org/10.1075/intp.13.2.02see

[53] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2017.74

[54] Daniel Servén, Charlie Brummitt, and Hassan Abedi. 2018. pyGAM: Generalized Additive Models in Python. *Zenodo*. https://doi.org/10.5281/zenodo.1476122

[55] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel Keim, and Mennatallah El-Assady. 2018. Going beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models. In *Proc. of IEEE VIS Workshop on Visualization for AI Explainability (VISxAI)*.

[56] Priti Shah and Patricia A. Carpenter. 1995. Conceptual Limitations in Comprehending Line Graphs. *Journal of Experimental Psychology: General* 124, 1: 43–61. https://doi.org/10.1037/0096-3445.124.1.43

[57] Priti Shah and Eric G. Freedman. 2011. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science* 3, 3: 560–578. https://doi.org/10.1111/j.1756-8765.2009.01066.x

[58] Shah Priti and Hoeffner James. 2002. Review of Graph Comprehension Research:\nImplications for Instruction\n. *Educational Psychology Review* 14, 1: 47–69. Retrieved from http://www.springerlink.com/content/v2581778612k5432/?MUD=MP

[59] Benjamin Strobel, Marlit Annalena Lindner, Steffani Saß, and Olaf Köller. 2016. Do graph readers prefer the graph type most suited to a given task? Insights from eye tracking. *Journal of Eye Movement Research* 9, 4. https://doi.org/10.16910/jemr.9.4.4

[60] Justin Talbot, Vidya Setlur, and Anushka Anand. 2014. Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2014.2346320

[61] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3278721.3278725

[62] Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. In *Machine Learning*. 349–391. https://doi.org/10.1007/s10994-015-5528-6

[63] Nadya Vasilyeva, Daniel Wilkenfeld, and Tania Lombrozo. 2017. Contextual utility affects the perceived quality of explanations. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-017-1275-y

[64] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*: 1–52. https://doi.org/10.2139/ssrn.3063289

[65] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300831

[66] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*: 1–15. https://doi.org/10.1145/3290605.3300831

[67] Andrzej Wodecki, Gregory C Allen, Michael C Horowitz, Elsa B Kania, Paul Scharre, H James Wilson, Paul R Daugherty, Nicola Morini-Bianzino, Rishie Sharma, Johan Boye, Chris Reed, David Gunning, Darpa Io, A I System, Andrew J Fawkes, Jacques Bughin, Eric Hazan, Sree Ramaswamy, Michael Chui, Tera Allas, Peter Dahlstrom, Nicolas Henke, Monica Trench, Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, Luciano Floridi, Gregory C Allen, T. Chan, Bo-hu Li, Bao-cun Hou,

Wen-tao Yu, Xiao-bing Lu, Chun-wei Yang, Derwin Suhartono Budihartono, widodo, and Jatin Borana. 2017. Explainable Artificial Intelligence (XAI) The Need for Explainable AI. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 58, 2: 4. https://doi.org/10.1111/fct.12208

[68] Simon N. Wood. 2017. *Generalized additive models: An introduction with R, second edition*. https://doi.org/10.1201/9781315370279

[69] Mike Wu, Michael C. Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. 2018. Beyond sparsity: Tree regularization of deep models for interpretability. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.

[70] Xiaoming Xi. 2010. Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing* 27, 1: 73–100. https://doi.org/10.1177/0265532209346454

[71] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. 2017. Scalable Bayesian rule lists. In *34th International Conference on Machine Learning, ICML 2017*.

[72] I. C. Yeh. 1998. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*. https://doi.org/10.1016/S0008-8846(98)00165-3

[73] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. https://doi.org/10.1111/rssa.12227