A Cognitive Load Theory (CLT) Analysis of Machine Learning Explainability, Transparency, Interpretability, and Shared Interpretability

Stephen Fox 1,* and Vitor Fortes Rey 2,3

1 VTT Technical Research Centre of Finland, 02150 Espoo, Finland
2 DFKI German Research Center for Artificial Intelligence, 67663 Kaiserslautern, Germany; fortes@dfki.uni-kl.de
3 Department of Computer Science, RPTU University Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
* Correspondence: stephen.fox@vtt.fi

Abstract: Information that is complicated and ambiguous entails high cognitive load. Trying to understand such information can involve a lot of cognitive effort. An alternative to expending a lot of cognitive effort is to engage in motivated cognition, which can involve selective attention to new information that matches existing beliefs. In accordance with principles of least action related to management of cognitive effort, another alternative is to give up trying to understand new information with high cognitive load. In either case, high cognitive load can limit potential for understanding of new information and learning from new information. Cognitive Load Theory (CLT) provides a framework for relating the characteristics of information to human cognitive load. Although CLT has been developed through more than three decades of scientific research, it has not been applied comprehensively to improve the explainability, transparency, interpretability, and shared interpretability (ETISI) of machine learning models and their outputs. Here, in order to illustrate the broad relevance of CLT to ETISI, it is applied to analyze a type of hybrid machine learning called Algebraic Machine Learning (AML). This is the example because AML has characteristics that offer high potential for ETISI. However, application of CLT reveals potential for high cognitive load that can limit ETISI even when AML is used in conjunction with decision trees. Following the AML example, the general relevance of CLT to machine learning ETISI is discussed with the examples of SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and the Contextual Importance and Utility (CIU) method. Overall, it is argued in this Perspective paper that CLT can provide science-based design principles that can contribute to improving the ETISI of all types of machine learning.

Keywords: agreeable AI; algebraic machine learning; cognitive load theory; CIU; decision trees; explainability; interpretability; LIME; SHAP; shared interpretability; transparency; world models

1. Introduction

Improving the explainability, transparency, and interpretability of machine learning are recognized as important goals [1–3]. In short, the explainability of machine learning can be facilitated by it being transparent enough to be interpreted directly by people. However, interpretability will not necessarily lead to different people agreeing a shared interpretation of the same information [4]. For example, there can be different interpretations of one written word [5] and even of a single drawn line [6]. There can be different interpretations of the same topic in research [7] and of the same topic in practice [8]. There can be different interpretations of large amounts of information [9] and of diagrams, such as tree diagrams, that are intended to simplify the presentation of information [10]. Different interpretations of the same information can arise because different people have different internal models of self in the world, which, for brevity, can be referred to as world models [11]. It can
be important to avoid conflicting interpretations of the same information. For example, in healthcare, conflicting interpretations need to be avoided because they may lead to poor treatment outcomes [12]. Accordingly, in addition to explainability, transparency, and interpretability, improving the *shared interpretability* of machine learning models and outputs is an important goal. For brevity, explainability, transparency, interpretability, and shared interpretability are referred to in this paper as ETISI.

ETISI can be hindered if information about machine learning models and information produced by machine learning models is complicated and/or ambiguous. The word complicated is used here for brevity to describe information that contains multiple interacting elements that must be considered simultaneously. Cognitive Load Theory (CLT) research has revealed that high intrinsic *cognitive load* can arise from information that inherently has high element interactivity. For example, equations can have multiple interacting elements that must be considered simultaneously. In addition, extraneous cognitive load can arise from ambiguities in the communication of information, such as ambiguous explanations of equation notation. CLT provides practical recommendations such as designing information so that attention is not split over more than one element. CLT research has revealed that the same information can entail different cognitive load for different people, such as novices and experts. This can involve information that is useful for a novice being redundant for an expert. Hence, CLT can provide insights into potential for the shared interpretability of information. Insights from CLT are already applied in a variety of fields in order to try to reduce the cognitive load of information [13–15]. This is because information with high cognitive load can entail high cognitive effort that does not necessarily lead to understanding or learning [16]. Recently, there has been some limited recognition that consideration of CLT could improve explainability and interpretability [17–20]. However, previous papers have been concerned with focused applications, rather than considering the broad relevance of CLT to the ETISI of all types of machine learning.

In the remaining four sections of this Perspective paper, CLT is related to machine learning ETISI. In Section 2, CLT is described in more detail and related to existing literature concerned with ETISI. In Section 3, CLT is applied to analysis of a type of machine learning called Algebraic Machine Learning (AML). This is the example here because AML has been developed in an effort to go beyond older types of ML that have been included within previous studies concerned with explainability, transparency, and interpretability. In particular, AML is a type of hybrid machine learning. That is, ML that encompasses predefinition of rules and ongoing learning from data [21,22]. Generally, the predefinition of rules could contribute to ETISI. AML has three characteristics that offer higher potential for ETISI than older types of ML: a fundamentally simple shallow structure comprising only three layers; AML algorithms to check that human-defined constraints are maintained; and AML rules that can be traced back to human-defined policies [23,24]. In Section 4, the general relevance of CLT to machine learning ETISI is explained with the examples of SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and the Contextual Importance and Utility (CIU) method [25]. In Section 5, principal findings are stated, practical implications of the study are discussed, and directions for further research are proposed.

### 2. Cognitive Load

In this section, cognitive load from complicated ambiguous information is explained in terms of CLT. Subsequently, CLT is related to previous literature concerned with explainability, transparency, interpretability, and shared interpretability (ETISI).

#### 2.1. Cognitive Load

High cognitive load can come from the addition of intrinsic load and extraneous load [26]. With regard to intrinsic cognitive load, within CLT it is posited that human working memory can handle only a very limited number—possibly as few as two or three—of novel interacting elements [27]. Cognitive load from the same information...
can depend upon the extent of prior knowledge, and more prior knowledge may not necessarily reduce cognitive load [28]. Some people may give up trying to understand information if it comprises too many elements [29]. As cognitive overload is detrimental, cognitive task analysis techniques have been developed in order to assess cognitive loads and how they can be reduced [30]. Cognitive task analysis can encompass different types of cognitive distance: for example, issue distance and semantic distance. Issue distance represents the cognitive effort required to understand that a shift in goal is needed to achieve some tasks [31]. Semantic distance can be considered in terms of path length comprising the number of steps needed to traverse from one word to another word in the same path [32]. Extraneous cognitive load can arise from ambiguities in the format of information. Ambiguities can be conceptual, presentational, and/or linguistic [33]. Terms that are widely used can be conceptually ambiguous [34]. For example, it has been argued that the term “medically unexplained symptoms” is a barrier to improved care because it is conceptually ambiguous [35]. Presentational ambiguities include color and layout [36,37]. Presentational ambiguities can also arise from the formats in which written text is presented [38]. Linguistic issues can be lexical, syntactic, semantic, and/or phonological. Lexical ambiguity comes from a word having more than one meaning. There can be syntactic ambiguities from sentences if they can be parsed in more than one way. There can be semantic ambiguity when the meaning of a sentence could be determined only through reference to wider knowledge sources [39]. Phonological ambiguity can arise when spoken sounds can be interpreted in more than one way [40].

Trying to understand information with high cognitive load can entail high cognitive effort that does not necessarily lead to understanding or learning [16]. An alternative to expending a lot of cognitive effort is to engage in motivated cognition that pays attention to new information if it matches existing beliefs [41]. Motivated reasoning and motivated numeracy may influence the extent to which people will make an interpretation of new information based on their existing beliefs rather than on the content of the new information. Within motivated reasoning, different political beliefs can influence what different people consider to be credible evidence [42]. Within motivated numeracy, people may align interpretations of data with their existing political attitudes even if they are highly numerate people [43]. Accordingly, motivated social cognition [44] may lead to interpretations of machine learning models and outputs being based on what is already believed rather than on what is currently seen. Other social issues, such as potential for stigma, can affect people’s willingness to agree with new information. For example, the term “patient label avoidance” refers to patients’ efforts to distance themselves from a label, i.e., a diagnosis, because it can be perceived as being socially unacceptable [45]. Apropos, some people who have functional gait disorders may not believe a diagnosis due to concerns about stigma [46]. Such reluctance to believe a diagnosis can be framed within the general tendency for people to want to continue to survive in preferred states [47]. In this example, that means preferring to survive in the preferred state of good health rather than having to try to survive in a “dis-preferred state” [48], such as having a functional gait disorder.

As well as motivated reasoning, the extent to which different people make different interpretations of the same information may also be influenced by lack of reasoning [49]. This could be framed as laziness. However, human beings have evolved to be economical in the expenditure of brain energy [50]. As energy is required to change information that is already neurologically encoded [51], less energy may be required for people to make reference to their existing internal world models than is required for reasoning about new information. Apropos, the probability of a system adopting a new configuration is inversely proportional to the energy associated with moving to that new configuration [48], and acceptance of some new information could have to involve extensive neural reconfiguration in what can be described as rewiring the connectome of many neural connections [52]. For example, people who accept diagnoses that they have functional gait disorders may need to update many beliefs about themselves and consider how they will survive in the new
dis-preferred state of chronic ill health. As human beings have evolved to conserve brain energy, it should not be assumed that people will want to expend a lot of cognitive effort trying to understand high cognitive load information if acceptance of that information would subsequently lead to the need for energy-consuming rewiring of the connectome over months and even years. Rather, research suggests that the brain seeks the most economic trade-off between neural connectivity and wiring costs [53] and has a mechanism for balancing benefits and risks in the connectome [54]. As summarized in Figure 1, consideration of cognitive load highlights that ETISI can involve interactions between social issues and biological issues as well as technical issues.

Figure 1 illustrates that people’s reference to their own existing internal world models, rather than to ML models and outputs, can be a least action biological response to minimize cognitive effort. The biological importance of minimizing energy expenditure in trying to understand information may be reflected in least action principles such as pragmatic principles of least effort [55] and the principle of least collaborative effort [56]. By contrast, trying to understand ML can involve cognitive effort across biosocial–technical actions if people choose to try to progress from explanation of ML to shared interpretation of ML. For example, action to resist biological preference for economical expenditure of brain energy, action to allay concerns about potential negative social consequences of accepting new information, and action to try to understand constructs, technical jargon, etc., related to ML and its applications.

2.2. Explainability, Transparency, and Interpretability

The term “explainability” can be associated with explanations of the internal workings of artificial intelligence (AI), including ML, during training and when making decisions. Such explainability can be summarized with the abbreviation XAI from eXplainable artificial intelligence [57]. There is also need for AI to be explainable to humans via the interfaces that they use. Apropos, there is need for human-centered AI i.e., HCAI [58]. This is because AI is deployed within tools that are used by people [59]. Different people can have different requirements for explanations [60], and even experts may not agree about explainability [61]. Hence, to use a colloquial automotive analogy, there is need for explanations for different people about what is “under the hood” of ML (i.e., ML models) and explanation of what is “above the hood” (i.e., of ML interfaces and outputs). It has been argued that causability is an antecedent of explainability [62]. In this context, causability has been defined “as the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use” [63]. Interrelationships between causation and explanation have been debated since at least the time of Aristotle. It has been argued that causation and explanation cannot be understood separately. However, it has also been questioned.
whether all causes are explanatory [64,65]. Generally, XAI and HCAI literature includes little consideration of the potential for explanations to entail high cognitive load. For example, cognitive load was not considered in a comparative study of three explainability techniques [25]. Also, there has been little consideration of CLT. For example, a method has been proposed for measuring and moderating cognitive load in charts and graphs [17], but without reference to recommendations within CLT.

The term transparency is associated with the observability of system behavior [66]. This is summarized in well-established comparisons of metaphorical opaque “black box” computing and transparent “glass box” computing [67], which have been applied to AI [68]. It has been argued that transparent techniques may work well if problems have already been described as a set of abstract facts, but they may not be so good for processes involved in extracting facts from raw data. [69]. Accordingly, there are efforts to improve transparency [70]. They can be focused on improving the transparency of a model or on improving the transparency of data outputs from models. These can be considered in terms of global characteristics or local characteristics, where global transparency encompasses an entire model and local transparency encompasses individual data samples. In some cases, models may provide local explanations but may not be globally interpretable [61]. It can be possible to bridge the global and local through feature engineering that involves designing a model around features that are not opaque to potential end-users. For example, a gait analysis model can be designed around features that both healthcare providers and patients can understand, such as head posture, left/right stride length, left-arm/right-arm swing, walking speed, and body swaying [71]. However, the display of important features is not inherent in some types of AI, such as deep neural networks. Accordingly, some types of AI models may only ever be “black boxes” and attempts to make them explainable may not be successful [2]. Meanwhile, advocates of transparent models have given little consideration to the potential for transparent models to entail high cognitive load.

Within research concerned with AI and ML, there is no one universally accepted definition of interpretability [72]. However, a distinction can be drawn between post hoc explanations of “black box” models and directly interpretable/self-explanatory models [61]. It has been argued that, for example, deep neural networks require post hoc explanation, but other forms of machine learning and AI, such as decision trees, can be directly interpretable. That is, provided there is sufficient domain knowledge and feature engineering involved in their formulation [61]. However, a survey of advances in decision trees revealed that not all decision trees are equally interpretable, and some decision trees may not be interpretable [73]. More broadly, a survey of interpretability has yielded a taxonomy of interpretability methods [72]. However, this does not include consideration of shared interpretation. Notably, neither the word cognition nor the word cognitive are included in this extensive survey. Hence, the survey reveals a lack of consideration of the need to reduce potential high cognitive load. Nonetheless, recently, there has been some studies that relate cognitive load to interpretability [18–20]. These have focused on consideration of intrinsic cognitive load in interpretation of algorithms [18], element interactivity in evaluating the interpretability of generative models [19], and cognitive load of contextual information for human–AI collaboration [20].

3. CLT Assessment of a Machine Learning That Could Have High Potential for ETISI

Here, CLT is related to gait analysis enabled by AML, which is a type of machine learning that could have high potential for ETISI. First, AML has a fundamentally simple shallow structure comprising only three layers. Second, AML algorithms check that human-defined constraints are maintained in relationships between inputs and outputs that are learned. Third, AML rules can be traced back to human-defined policies [23,24].
In the example, human-interpretable gait features are used to predict if depression was detected or not. The data set is from a previous study conducted by [71]. The features are arm swing left, arm swing right, body sway, head posture, stride height right, stride height left, stride length left, stride length right, and vertical head movement. The gait analysis examples involve two important elements of AML: constants and atoms. Within AML, constants define features of the machine learning problem, such as different walking speeds. Metaphorically, constants can be considered to be the vocabulary for the machine learning problem. Also metaphorically, atoms are components in descriptions of particular examples of the problems that are learned by the AML algorithm. In Figure 2, one atom is shown. This atom provides a description component that includes the vocabulary terms of vertical head movement and walking speed, which describes the gait of a particular person. As shown by Figure 2, each atom is connected to one or more gait features. Figure 2 illustrates that it can be difficult to see atoms and constants even when there is only one atom connected to a few constants. Thus, even a shallow structure of inputs–atoms–outputs does not necessarily provide a practical basis for ETISI.

As shown in Figure 3 below, interactions between elements (atoms and constants) can be made clearer by enlarging the view shown in Figure 2. However, this can introduce new cognitive load when only part of the information in Figure 2 is visible if attention has to be split over two or more enlarged partial views of a diagram.

Decisions are made based on the number of atoms present (i.e., not missing). In terms of explainability, some everyday words are used, such as speed and head movement. Thus, as shown in Figure 3, there is potential for human reading of the diagram when it is enlarged to be legible. However, there is intrinsic cognitive load as the overall structure, its components, and interrelationships between them are not necessarily self-explanatory. For example, the top atom is present if and only if at least one of the constants is present in the instance, which in this case is the subject’s particular recorded gait. In this example, one constant can be vertical head movement being too high or vertical head movement being too low. Another example of one constant is the walking speed being too high. Moreover, interval values are not succinct, with the same one variable, for example, speed, appearing many times instead of being summarized in one entry. This illustrates that showing the overall raw structure of the atoms, even for one example, can be ambiguous unless further preprocessing is carried out. Figure 4 shows an AML description for gait analysis using AML Description Language, which is an example of a development intended to facilitate increased use of machine learning by people who do not have advanced education in machine learning.
However, regarding interpretability, it requires not only knowledge of programming language words, but also of the AML Description Language. Thus, it is not easily interpretable for non-experts. Regarding transparency, the definitions are clear for experts, and the final trained model can possibly be a “glass box” for experts. In terms of shared interpretability, we have observed confusion about the possible solutions the model can generate, even among people who have backgrounds in computer science. Hence, it is questionable whether AML Description Language, or similar languages intended to bridge everyday terminology and machine learning languages, can be used to facilitate shared understanding of descriptions of everyday lived experience, such as gait, which involve specifying interpretable features as discretized numbers; for example, the speed of walking. Rather, bridging languages can introduce another layer between people and machine learning, which needs a major design effort to prevent introduction of new unintended barriers to explainability, transparency, interpretability, and shared interpretability. This layer increases the number of interacting elements. Another challenge for shared interpretability is cognitive overload. If one has to analyze many atoms, it soon becomes not feasible to interpret the model, even if each atom alone may be transparent. This problem is illustrated in Figure 5, which comprises less than 30 atoms. As the reader can see, an AML model can expand to become transparent but uninterpretable even when enlarged as in Figure 6. Accordingly, it can be expected that people would refer more to their existing beliefs about gait than to the new information from AML-enabled analysis.

Considering that final AML models can have thousands of atoms, there is high intrinsic cognitive load from high element interactivity and, hence, low practical potential for direct interpretability. Given the amount of information present, it is challenging for participants to achieve a shared interpretation. Instead, participants could pick and choose interpretations that support their existing beliefs [10]. As it is not feasible to look at all atoms, one may simplify the model by looking at only the most relevant ones. One way to do this is to use atom presence as a feature and then build a decision tree. For example, a tree diagram can be generated by using classification and regression algorithm (CART) from the AML-enabled gait analysis. A tree diagram generated using a standard technique such CART can reduce intrinsic cognitive load by reducing the number of elements and interactivity between them. Furthermore, potential extraneous cognitive load from ambiguities could be reduced. In particular, everyday English language words can be used within a tree diagram, such as head posture, walking speed, left-arm/right-arm swing, left/right stride length,
and body swaying. As shown in Figure 7 below, a tree diagram can have a consistent rationale in its descriptive progression of 222 participants in gait analysis. At the top of the tree, which is the beginning of the tree, gait indicators of depression were not detected in 109 participants, but were detected in 113 participants. Hence, the classification stated in the first tree node is detected. In this first node, the analyzed feature is head posture. The results of subsequent feature analyses are shown in subsequent nodes. In particular, the next level of the tree diagram in the left node shows that analyses of walking speed, arm swing left, body sway, head posture, and stride length indicate that 34 of 40 participants have indicators of depression. By contrast, the right node shows that indicators of depression were not detected in the majority of the other 182 participants (i.e., 222 participants minus 40 participants). The majority being 103 participants. Presentation of the analysis results proceeds with this rationale from the top to the bottom of the tree.

**Figure 7.** Tree diagram representation of results of AML-enabled gait analysis.

The tree diagram is transparent. However, the extent to which the tree is interpretable is questionable. This is because the tree provides a post hoc explanation of a model rather than a direct interpretation of a model. Typically, it is argued that post hoc explanation is needed for “black box models”, and it can be argued that AML does not entail black box models. This is because AML has a fundamentally simple shallow structure of only three layers: inputs, atoms and outputs. Also, AML algorithms check that human-defined constraints are maintained in relationships between inputs and outputs that are learnt as new atoms are generated during training. In addition, AML rules can be traced back to human-defined policies. Nonetheless, as illustrated above in Figure 5, the potential for direct interpretation of AML models and outputs from AML-enabled analyses is reduced by the number of atoms increasing from tens to thousands during training. Hence, post
hoc explanations can be required for models that are theoretically “glass box” models such as AML models, in addition to other types of models that are generally regarded as “black box models”.

With regard to shared interpretability, the tree diagram shown in Figure 7 includes many information elements and interactions that need to be processed in order to understand why categorizations are made. Furthermore, there can be both types of cognitive distance to overcome. For example, there can be issue distance between healthcare providers and people with gait disorders because the goal of the tree diagram presentation is not stated clearly. Also, there can be semantic distance because spatiotemporal terms are presented in different sequences in different nodes. In addition, there are ambiguities that can entail extraneous cognitive load. For example, the top node of the tree contains the information: \((\text{headPosture} > 1.18) \leq 0.5\). In this case, \((\text{headPosture} > 1.18)\) represents an atom that is only present when the head posture feature is bigger than 1.18. Since an atom can either be present (value of 1) or not present (value of 0), the top node is split using \(\leq 0.5\). This means the reader should follow the left arrow if the atom is present or the reader should follow the right arrow if the atom is not present. For example, if indeed the headPosture is \(>1.18\), the reader would be at the first node on the left for the second level of the tree and, at this point, have a prediction of detected because there are 34 cases of detected and only six not detected. By contrast, in the fifth level, the atom is more complex, and is represented as \((\text{headPosture} \leq 1.21 \text{headPosture} > 1.22 \text{strideHightRight} > -0.7)\). In this case, the atom is linked to three conditions, namely, \(\text{headPosture} \leq 1.21\), \(\text{headPosture} > 1.22\), and \(\text{strideHightRight} > -0.7\). This means the atom is present if at least one of these three conditions is true. This is related to the nature of AML, where atoms are present if any of their constants are found. Thus, although AML has a simple structure, cognitive load can arise when analyzing a single node within a tree, because several novel interacting elements need to be considered simultaneously. In this case, the reader needs to consider two intervals for headPosture as well as a second variable, stride-HightRight. The evaluation of the expression also requires the reader to perform a logical OR, which is less intuitive than simply evaluating a single variable, as one would do in a decision tree trained on the raw data. Coming back to the prediction related to this node, \((\text{headPosture} \leq 1.21 \text{headPosture} > 1.22 \text{strideHightRight} > -0.7) \leq 0.5\) means that depression is not detected if the atom is not present. This means that if none of the listed conditions is true at this point in the analysis, then the final prediction is not detected. Thus, although there is correspondence between numbers and words, there is a lack of obvious consistency about how measurements directly relate to categorization. When there are ambiguities within nodes, cognitive load may be increased by readers having to split their attention across nodes as they try to understand how the information within one node leads to information in subsequent nodes.

Generally, there are ongoing efforts to improve tree diagrams. However, these efforts can involve implementation of more technical steps to provide explanations for tree diagrams, i.e., there is recognition that tree diagrams are not necessarily directly interpretable. For example, SHapley Additive exPlanation (SHAP) values can be applied to tree diagrams [74]. Application of (SHAP) values can involve a wide variety of diagrams, including bar charts that have both positive and negative SHAP values for multiple variables represented by horizontal bars, which can be colored to indicate the extent to which each variable has a high or low feature value [74]. Hence, tree diagrams can be replaced by different diagrams that also have potential for cognitive load.

4. Broad Relevance of CLT to Machine Learning ETISI

CLT analysis of AML reveals that even a type of ML that, in theory, could have high potential for ETISI can have high potential for high cognitive load in practice. In this section, the broad relevance of CLT to other types of machine learning is discussed.
There is potential for CLT principles to be applied to machine learning from the outset of model formulation through to the choice of formats for communicating model outputs. For example, CLT research findings indicate that modular structuring can entail lower intrinsic cognitive load [75]. Accordingly, CLT could be applied to inform the structuring of ML models with the aim of reducing their intrinsic cognitive load. With regard to reduction of extraneous load in the communication of model outputs, the split-attention effect and the redundancy effect are important to consider. In the following paragraphs, these are related to comparison of three types of ML explanation methods: SHAP, LIME, and CIU.

The study [25] provides a very thorough and well-described comparison of Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and the Contextual Importance and Utility (CIU) method. As explained in more detail in [25], LIME is intended to explain a model’s prediction by approximating the prediction locally. This is done by perturbing the input around the class of interest until it arrives at a linear approximation. This is intended to help justification of the model’s behavior. SHAP aims to describe model predictions by distributing the prediction value among the features, which depends on how each function contributes. This includes global interpretability in terms of the importance of each indicator that has a positive or negative impact on the target variable, and local interpretability in terms of SHAP values for each instance. SHAP values can be calculated for any tree-based model. Both SHAP and LIME approximate the local behavior of a black box system with a linear model. Hence, they provide local fidelity. However, faithfulness to the original model is lost. By contrast, the Contextual Importance and Utility (CIU) method assumes that the importance of a feature depends on the other feature values. In other words, a feature that is important in one context might be irrelevant in another. The feature interaction allows for the provision of high-level explanations, where feature combinations are appropriate or features have interdependent effects on the prediction. CIU encompasses contextual importance (CI), which approximates the overall importance of a feature in the current context, and contextual utility (CU), which provides an estimation of how favorable or unfavorable the current feature value is for a given output class. It is explained in [25] that LIME, SHAP, and CIU each has its own strengths and weaknesses. However, it is also reported in [25] that explanations with CIU were found to be better in several ways than explanations with SHAP and LIME.

Illustrative examples of the three types of explanation methods are shown in Figure 8, which is derived from high-quality color presentations that are shown separately to each other in [25]. In Figure 8, monochrome illustrations of (a) SHAP, (b) LIME, and (c) CIU are shown together. For each explanation method, the illustration represents an image on the left-hand side and related diagram(s) on the right-hand side. Figure 8 is monochrome so there is equal representation of the explanation methods, and to illustrate that the split-attention effect and the redundancy effect can be seen even when relatively rudimentary visual representations are viewed.

With regard to Cognitive Load Theory (CLT), the split-attention effect can occur when two sources of information about one topic, such as an image and some text, are not entirely intelligible when they are separate. Hence, people must try to pay attention to both to try to consider them together in order to understand the topic being presented. This involves cognitive effort that would not be required if there was better integration of information about the topic. Although there is no reference to CLT or cognitive load in [25], it is apparent from viewing the illustrations in Figure 8 that the (a) SHAP explanation entails the split-attention effect far more than (b) LIME and (c) CIU. This is because of the presentation of the horizontal SHAP value scale, which spans negative values and positive values across one image and two related diagrams. Hence, it is necessary to pay attention to one image, two diagrams, and a wide scale at the same time to try to understand the information being presented. Thus, there is high potential for the split-attention effect.
As discussed above, CLT research has proven that the split-attention effect increases cognitive load. Hence, it can be expected that SHAP explanations will require more cognitive effort to understand than the other types of explanation. Apropos, it is reported in [25] that “Users given SHAP explanations also required significantly more time to complete the study compared to those provided with LIME explanation support”, “which may indicate that SHAP explanations required more in-depth concentration and were harder to interpret”.

Also, the redundancy effect should be avoided. This involves avoiding the presentation of information that is redundant because it does not facilitate understanding or learning. For example, the same information content can be presented in more than one format. Unlike the split-attention effect, the information in each format is intelligible on its own [76]. Nonetheless, cognitive effort is involved in processing the same information more than once. This can result in having reduced cognitive resources for understanding the information. Alternatively, additional information can be given with the aim of enhancing or elaborating the first information presented. This can lead to people expending cognitive effort by having to try to coordinate the information. As can be seen in Figure 8b, in the right-hand side of the LIME illustration, additional information is laid on top of the original image to indicate the presence of the phenomenon of interest. By contrast, as can be seen in Figure 8c, there is no redundant repetition of information in the right-hand side of the CIU illustration. Accordingly, by referring to CLT, it could be predicted that CIU explanations would be more effective than SHAP explanations and LIME explanations. Apropos, it is stated in [25] that “users given CIU explanation support answered more questions correctly,
although the difference was not significant”. Also, it is stated in [25] that “When comparing the ability of users to distinguish between correct and incorrect explanations, the users given CIU explanation support showed a better understanding of the explanations than both users given LIME” “and SHAP explanation support” “although the difference was not statistically significant”.

While CLT principles are broadly applicable, the application of CLT is more important for some applications of machine learning than for others. As shown in Figure 9, the need to reduce cognitive load is lowest when machine learning models and their outputs are well aligned with people’s existing beliefs. This is because it will make no difference to people’s beliefs whether or not people do expend cognitive effort to overcome high cognitive load.

![Figure 9. Importance of applying CLT.](image)

Overall, the less machine learning information is aligned with a person’s or a group of people’s existing world model, the greater the need to minimize cognitive load in machine learning information. This is because the brain gathers sensory data that are consistent with its internal model and avoids sensory inputs that diverge from that internal model [11,48,77,78]. The need to reduce cognitive load is highest when acceptance of information would entail people having to change their core beliefs and when changing core beliefs would involve having to try to survive in a dis-preferred state, such as in a state that is outside of a lifetime’s social group or in a state outside of good health. The need to reduce cognitive load is highest because human beings have evolved to be economical in the expenditure of brain energy [50], and people cannot be expected to expend a lot of cognitive effort trying to understand high cognitive load information if acceptance of that information would subsequently lead to them changing to a dis-preferred state, which involves the need for further increased energy consumption over months or even years.

As shown in the center of Figure 9, in between the two extremes are situations when acceptance of new information would entail changing peripheral beliefs such as which bank offers the best financial services, e.g., current accounts, savings accounts, home loans, etc. Such a belief change could be followed by increased energy consumption over a short time if the person then changes accounts and loans to the bank now believed to offer the best
financial services. However, this would be done with positive expectations of improving potential for survival in preferred state due to improved finances.

5. Conclusions

Cognitive Load Theory (CLT) provides a framework for relating the characteristics of information to human cognitive load. CLT has not been applied comprehensively to improve the explainability, transparency, interpretability, and shared interpretability (ETISI) of machine learning models and their outputs. The principal contribution of this Perspective paper is to illustrate the broad relevance of CLT to ETISI. This has been done through application of CLT to analyze a type of hybrid machine learning that, in theory, has high potential for ETISI. However, application of CLT reveals potential for high cognitive load that can limit ETISI even when used in conjunction with decision trees. The general relevance of CLT to machine learning ETISI is illustrated further with the example of comparing LIME, SHAP, and CIU. In addition, a framework for assessing the importance of applying CLT has been provided (Figure 9).

The implementation of science-based design principles can bring large performance improvements in diverse sectors [79,80]. CLT principles are already being applied in a variety of sectors to improve performance [13–15]. Overall, the practical implication from the research reported here is that CLT could be applied with the aim of improving machine learning ETISI in the many different sectors where machine learning is being used. A challenging direction for future research is to apply CLT with the aim of improving shared interpretability among people who have opposing positions about issues being analyzed with machine learning. Such research is needed to progress from the goal of explainable AI to the goal of agreeable AI.

Author Contributions: Conceptualization, investigation, writing original draft, writing—review and editing, S.F.; software, formal analysis, writing—review and editing, V.F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by European Union (EU) Horizon 2020 project ALMA grant number 952091.

Data Availability Statement: No new data were created for this study.

Acknowledgments: S.F. thanks Karl Friston and Peter Zeidman for discussions of world models belief updating. V.F.R thanks Paul Lukowicz for technical advice.

Conflicts of Interest: The authors declare no conflicts of interest.

References


75. Gerjets, P.; Scheiter, K.; Catrambone, R. Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learn. Instr.* 2006, 16, 104–121. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.