IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 August 2022, accepted 14 September 2022, date of publication 19 September 2022,
date of current version 26 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3207812

RESEARCH ARTICLE

Toward Accountable and Explainable Artificial Intelligence Part One: Theory and Examples

MASOOD M. KHAN, (Member, IEEE), AND JORDAN VICE
Faculty of Science and Engineering, Curtin University, Perth, WA 6102, Australia
Corresponding author: Masood M. Khan (masood.khan@curtin.edu.au)

This work was supported by the Curtin University under Grant HDR-APA-18336230/2019.

ABSTRACT Like other Artificial Intelligence (AI) systems, Machine Learning (ML) applications cannot explain decisions, are marred with training-caused biases, and suffer from algorithmic limitations. Their eXplainable Artificial Intelligence (XAI) capabilities are typically measured in a two-dimensional space of explainability and accuracy ignoring the accountability aspects. During system evaluations, measures of comprehensibility, predictive accuracy and accountability remain inseparable. We propose an Accountable eXplainable Artificial Intelligence (AXAI) capability framework for facilitating separation and measurement of predictive accuracy, comprehensibility and accountability. The proposed framework, in its current form, allows assessing embedded levels of AXAI for delineating ML systems in a three-dimensional space. The AXAI framework quantifies comprehensibility in terms of the readiness of users to apply the acquired knowledge and assesses predictive accuracy in terms of the ratio of test and training data, training data size and the number of false-positive inferences. For establishing a chain of responsibility, accountability is measured in terms of the inspectability of input cues, data being processed and the output information. We demonstrate applying the framework for assessing the AXAI capabilities of three ML systems. The reported work provides bases for building AXAI capability frameworks for other genres of AI systems.

INDEX TERMS Explainable artificial intelligence, accountable XAI, machine learning system design, interactive graphical user interface.

I. INTRODUCTION

Experts from the domains of logic programming, automated reasoning and software engineering are believed to lead Artificial Intelligence (AI) and Machine Learning (ML) system design efforts [1], [2], [3]. Practitioners, usually less involved in these efforts, find the prevailing eXplainable Artificial Intelligence (XAI) frameworks algorithm-centric, neglecting domain-specific needs and, missing practical explanations [4]. Contemporary literature highlights several gaps in computing experts' view of eXplainable Artificial Intelligence and practitioners' explainability requirements [5], [6], [7]. From practitioners' perspectives, these gaps result in (a) no or little utility of the system explainability features and (b) users' inability to interpret the given reasoning. Such gaps inhibit automation of tedious practices and impede adoption of AI systems [8], [9], [11], [12]. Statistical and probabilistic explanations are considered limited and less effective [13], [14]. The relevant literature suggests that the prevailing XAI frameworks do not fully comply with the norms of regulatory bodies and industry [5], [6]. A proven method of measuring the non-explainability of an AI or ML system is not available yet [15]. As availability of better XAI frameworks would boost user confidence in ML and AI systems, attempts are underway to develop holistic XAI frameworks [8], [9], [11], [12]. Since AI systems are still regarded as difficult to understand, adopt and trust [16], several groups and are engaged in holistic XAI framework development efforts [17], [18], [19], [20].

This work posits that perceiving XAI in a two-dimensional space of predictive accuracy and comprehensibility results in mixing factors of accuracy, explainability and accountability [1]. Such a convoluted representation does not help practitioners, cannot fulfil regulators' expectations and, offers limited transparency for establishing a chain of responsibility [8], [9], [10], [11], [12]. In order to formulate a better XAI

The associate editor coordinating the review of this manuscript and approving it for publication was Varuna De Silva.

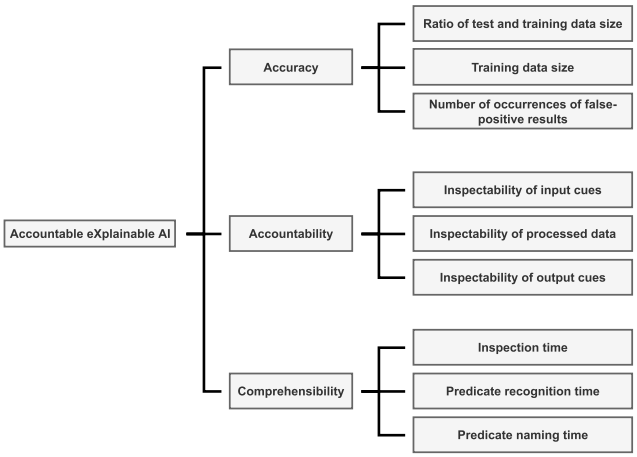


FIGURE 1. The three-level Galois-lattices structure leading to development of a holistic representation of explainability in ML and AI system.

framework, we formulated a three-level (narrow and shallow) Galois-lattices structure [21], shown in Fig. 1.

The Galois-lattices structure contains nine important elements that allow separating the convoluted factors of XAI. This separation is achieved by constructing a three-dimensional (3D) space using the nine terminal elements of the Galois-lattices structure. The perceived 3D space comprises of three mutually perpendicular vectors: accuracy, comprehensibility and accountability, each having the same units of length [22]. Each of the three axes of this Accountable eXplainable Artificial Intelligence (AXAI) space is an independent vector in a Cartesian coordinate system. Hence, each vector would be of the form: $A = \sqrt{a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}}$ where $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are unit vectors. In this 3D AXAI space, quantitatively separable vectors would allow for the deconvolution of predictive accuracy, comprehensibility and accountability. In Section III, Table 1 we report the data type and numerical values assigned to each element of the three vectors. These assigned values would determine the length of each vector in the 3D space. Displaying these system attributes would make it easy to quantitatively delineate various ML systems. This novel approach was built upon prevailing XAI paradigms [1], [3], [4], [6], [7], [8], [9], [11], [12], [13], [14] to propose an AXAI capability framework to:

1) Provide an easy to incorporate AXAI capability framework, mainly for ML systems;
2) Enable incorporating and measuring predictive accuracy;
3) Enable incorporating and quantifying the level of comprehensibility of the system;
4) Enable incorporating and quantifying the level of accountability of the system and;
5) Allow practitioners to visually and quantitatively examine various pieces of information and easily assess the system AXAI capability.

The AXAI capability framework, in its current form, is applicable to the ML systems. Henceforth, any reference to the framework's application would mean design and/or assessment of ML aspects of AI systems. In the following sections, we demonstrate the framework application by assessing and comparing three affective state classification systems. As shown in Fig. 2, the AXAI capability framework would allow for incorporating theoretical guarantees, empirical evidences and statistical assurances in AI systems.

In order to present the theoretical foundations of the AXAI and demonstrate its utility, this paper is organized in seven sections. After introducing this work in Section I, Section II provides a brief overview of the XAI related issues citing relevant works. We establish theoretical foundations of the proposed AXAI framework in Section III. The following Section IV demonstrates application of the proposed framework in designing and assessing AXAI capabilities of three ML systems. The three systems' assessment results are presented in Section V. The proposed framework and its applications are analysed and discussed in Section VI. Finally, Section VII identifies the possible directions of future work and concludes this work.

II. ISSUES IN EXPLAINABLE ARTIFICIAL INTELLIGENCE

Issues pertaining to algorithmic biases embedded in ML systems were first realized in the late 1970s [23]. Initial ML systems had nothing but predictive accuracy to offer as explanations. Later, it was realized that predictive accuracy alone would not suffice dealing with biases. It was understood that several embedded factors like the historical background, political constraints, and institutional context of ML systems also induce biases in ML system [17]. Such realizations are still valid for all genres of AI systems including supervised learning-supported classifiers, regression systems, unsupervised learning-supported clustering and labelling systems, reinforcement learning systems and deep neural networks. With time, the importance of explaining inferences, proving system accuracies, addressing accountability in the context of AI systems has increased [25], [26]. Recently, governments and business entities have also started to emphasize the need to account for the ethical implications of using AI systems [5], [6]. A recent report jointly published by the Ada Lovelace Institute, AI Now Institute and the Open Government partnership lists some forty algorithmic accountability mechanisms and their respective jurisdictions [28]. Hence, XAI has emerged as a topic of interest for computer scientists, AI theorists and practitioners across various domains [8], [18].

Though rule-based expert systems and ML systems were traditionally assessed on the basis of their predictive accuracy alone [29], recent developments made it possible to delineate them in a two-dimensional space of orthogonal axes viz., predictive accuracy and comprehensibility [30]. Consequently, ML systems are becoming relevant in solving both routine and complex problems [31] and in some domains they outperform humans and are becoming inevitable assets [24]. Thus, ML systems are now being used in critical tasks like disease diagnosis, psychological and psychiatric

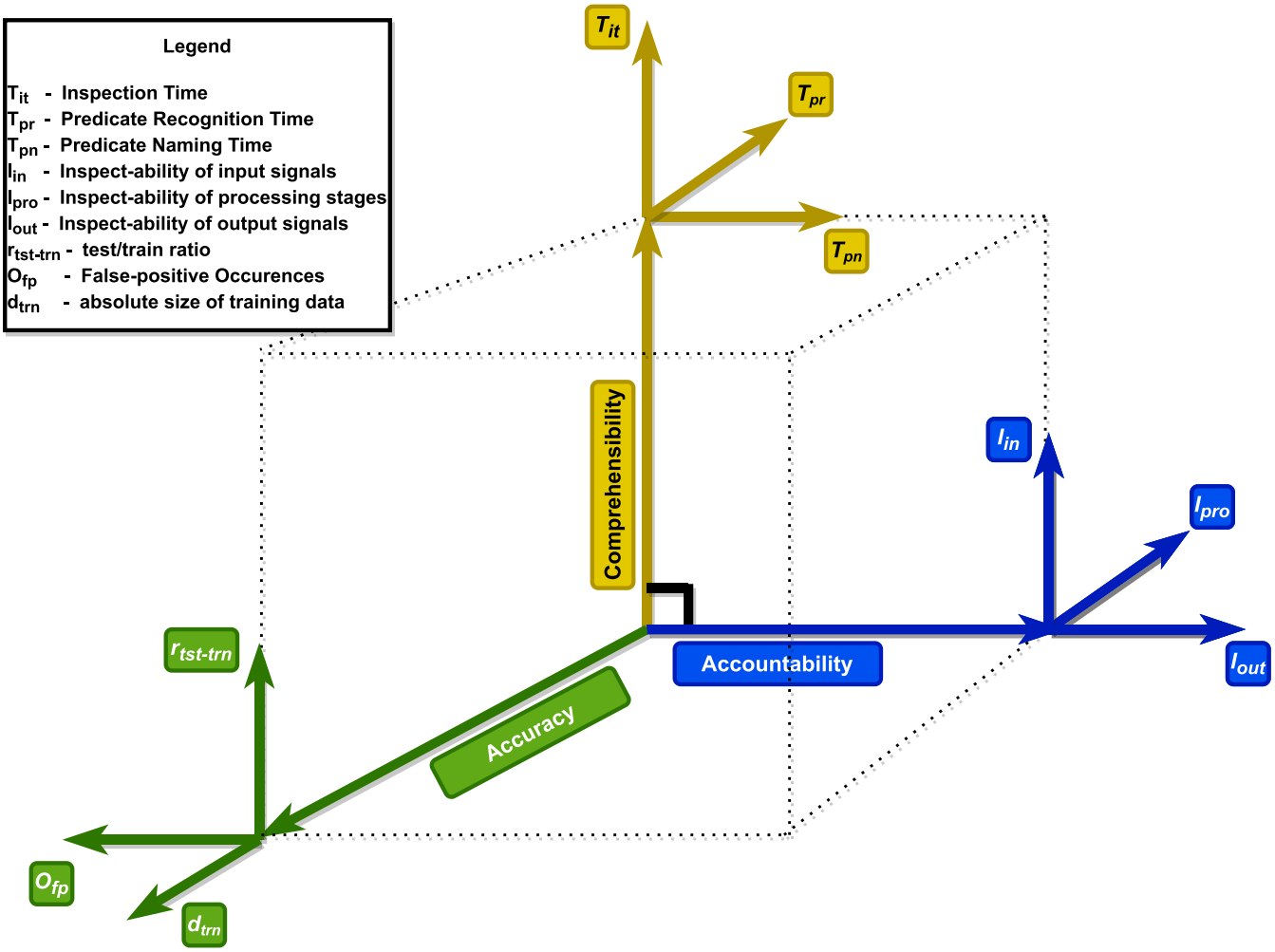


FIGURE 2. A three-dimensional representation of the AXAI capability framework, showing all quantifiable elements of system accountability, comprehensibility and predictive accuracy in the three vectors. Each vector comprises of three unique elements. The proposed framework allows for the quantitative assessment and delineation of ML and AI systems in the three-dimensional AXAI space.

assessments, loan approvals, autonomous driving and threat analysis. Their critical roles and decisions might also cause negative consequences [32]. AI theorists and practitioners acknowledge that a measure of 'accountability' needs to be added to AI-supported reasoning [1]. Hence, accountability is becoming a new dimension that would transform the two-dimensional space of predictive accuracy and comprehensibility [30] into a three-dimensional space.

Until now, incorporation of accountability features in AI systems' XAI capabilities is not common. For example, [29] reported that accuracy of predictions and comprehensibility of knowledge provided bases for proposing a set of criteria for delineating ML systems [2]. A weak criterion was used to identify ML systems whose predictive performance could improve using larger amounts of training data. A strong criterion would identify systems that symbolically provided reasons. An ultrastrong criterion was able to delineate ML systems that would teach reasoning [2]. Building upon these works, the description and scope of AI system comprehensibility was further refined in [3]. As evident in IEEE standard P2840, researchers are trying to go beyond the current XAI capabilities for building responsible AI systems [33].

Accountability, in the context of AI systems, connotes compliance with ethical, procedural and legal norms while processing information, invoking rules and making decisions [34]. A widely adopted definition of accountability defines it as a relationship between an actor and a forum, in which the actor has an obligation to explain and justify the conduct. Also, the actor may face consequences [35] for the impact of actions. Therefore, accountability is perceived as a multi-factor issue that deals with transparency, interpretability, post hoc inspection of outputs, pre- and post-market empirical performances and system design processes [1]. The 2019 Algorithmic Accountability Act discussed in the US senate required businesses to assess AI and decision support system for risks associated with privacy and security of personal information. The act also emphasized on assesing

risks of ''inaccurate, unfair, biased, or discriminatory decisions.'' The act further asked businesses to address the outcomes of AI systems' assessments [36]. Several applicable elements of AI system accountability could be extracted from the bill. Furthermore, as AI systems now perform highly critical tasks, they are also considered liable to adjudication, legislation and litigation [1]. For example, the use of COMPAS, a system that assesses offenders' criminogenic needs and risks of recidivism instigated legal debate and deliberations on accountability and lack of transparency in AI systems [23].

Philosophically, explanation is the act of making something intelligible or understandable [36]. Explainability in the context of AI-supported systems has been treated as a core software engineering issue but *accountability* is typically assessed in the context of application. Thus, an intelligent tutor would be deemed responsible for coaching and an autonomous vehicle would be held accountable for safety-centred issues. Typical ML algorithms rely on robust and accurate models based on the given data. Nonetheless, during their application, these ML systems fail to provide user-centred descriptions of how models were developed and how inferences or predictions were made. Incorporating acceptable, trustworthy and explainable artificial intelligence poses many challenges, mainly for application-related sensitivities and domain-specific requirements of various professions. It has been realized that practitioners' inputs had been minimal in an almost four-decade long journey - from relying on the one-dimensional predictive accuracy to the integration of explainability and accountability in AI-supported systems [1], [2], [3], [33]. An added issue that complicates incorporation of explainability in ML systems is that any two people would see the relevance and quality of explanations differently. In recent literature, evaluation of explanations is connected with data visualization techniques [37] as stated in [13]. The Palo Alto Research Centre (PARC) proposed incorporation of an interactive system for explaining the capabilities of an XAI system that controls a simulated unmanned aerial system. The PARC posited that system explanations should reveal all information used in decision-making by showing that the system understood how things worked and was aware of its goals. In order to achieve these capabilities, the PARC's Common Ground Learning and Explanation (COGLE) initiative established the terms to use in explanations and their meaning. The PARC used an introspective discourse model, which interleaves learning and explaining processes [13].

Based on the cited works, it could be argued that assessing AI-supported systems in a two-dimensional space of explainability and accuracy provides limited information on systems' capabilities. Given the emerging roles and applications of ML systems, a three-dimensional framework of AXAI capability assessment that includes accountability is required. The following section is dedicated to establishing the theoretical foundations of a three-dimensional AXAI capability space.

III. THEORETICAL FOUNDATIONS OF THE ACCOUNTABLE EXPLAINABILITY (AXAI) CAPABILITY FRAMEWORK

Although several recent works discuss incorporating measurable parameters of accountability and explainability [17], [27], little work has been done for developing a holistic framework and providing a set of quantifiable features to assess the AXAI capabilities of a system. A framework for assessing the AXAI capabilities must be built upon considerations pertaining to personal, social, moral and legal factors used to hold an individual accountable and liable for explaining personal actions and decisions [41]. Significant moral and legal factors that make a decision system liable to explain decisions are [39]:

1) Significance of the impact (effect) of a decision on others excluding the decision maker;
2) Possibility of contesting or overturning a decision;
3) Possibility of seeking compensation for damages caused by the decision, and
4) Existence of doubts about any one or a combination of: the provided information, the produced information and the process of making inferences and decisions.

While suggesting enhancements to the prevailing XAI capabilities, need for algorithmic accountability has been highlighted in the recent works [19], [39], [40]. Accountability of an AI system would depend on the context of the confronted issue [19]. For example, how a medical AI system chooses which one of two patients should be treated first or how a search and rescue robot would pick one of several injured victims [41]. Hence, an ML system should be aware of the context of ethical values and should have the capacity to understand the moral consequences of its actions and decisions [42]. Accountability should therefore be derived from both information/data and the algorithmic approach [36]. The employed algorithmic approach and data must be sensitive to the context while making inferences and decisions [22], [39], [41]. It is also argued that a system should be operated in such a manner that the chain of responsibility is clear and identifiable [25], [38].

In order to address such needs, our proposed AXAI framework includes a system accountability vector comprising of three components viz., inspectability of input data models or cues, inspectability of data being processed, and inspectability of output models or cues. In order to hold either a system developer or a user accountable for the impact of system decisions, relevant information must be presented to them in a meaningful manner [42]. We posit that inspectability, in the context of XAI, must allow users to examine the relevant system details and let them determine if the system is able to fulfil the decision-making requirements. Inspectability is also referred to as verifiability and traceability in the literature and is considered as one of the core features that ensure system transparency [43], [44].

The proposed AXAI framework posits that system developers and system users should be able to inspect the input data, important details on data being processed and the output

information. Both developer and user would be expected to understand, analyse and interpret the inspected data.

In the AXAI framework, an explanation is viewed as a deductive argument containing universal laws. Following this premise, the explainability vector comprises of the inspection time, the predicate recognition time and the time required to recognize or connect with a situation. The three factors of explainability improve user understanding of the situation (contextual inference) - from superficial knowing to a deeper knowing. In explainability, the inspection time serves as a substitution for incomprehension [28]. The predicate recognition time is grounded in the idea that humans understand an encountered situation by mapping the situation to those situations they would have encountered in the past. The last factor, the predicate naming time represents the time required to recognize or connect with a situation reflecting on system's ability to provide readily understandable explanations.

The predictive accuracy of a system in our AXAI framework includes three factors viz., ratio of the test data size and the training data size), the training data size and the number of occurrences of false-positive results. The ratio of test and training data informs how well a model performs on new data that were not used during the model development and system training [45]. The size of training data is important, as sufficient data are required for both developing an ML model and evaluating the model with a high degree of confidence. Without an adequately sized dataset, it will be dangerous and difficult to generalize results. Several good practices have been recommended for determining the adequecy of the validation dataset. For example, power calculations can be helpful for determining the sample data size that would be required to confidently evaluate the ML model performance and compare the model with a pre-determined baseline [46]. In addition, in the context of ML systems, the cross-validation approaches need a particular minimum size of the training data. In the absence of sufficient data for training and evaluating a model, making meaningful forecasts would not be possible. However, the required minimum size of the training data varies with the complexity of the model [45]. The last factor, occurrences of false positive results, helps in estimating the risks associated with a model [47]. The three components of accuracy vector work together to inform system developers and users on the perceived accuracy of the model and various inferences made using the model.

A. THE THREE-DIMENSIONAL SPACE OF ACCOUNTABLE EXPLAINABILITY (AXAI)

Building upon previous works [1], [3], [13], [14], [28], [38], we propose an AXAI capability framework for effectively incorporating accuracy, comprehensibility and accountability in ML systems. Following subsections discuss assumptions, definitions and hypotheses leading to the design of our proposed AXAI capability framework. These assumptions and definitions were inspired by and adopted from the relevant literature [3], [10], [28], [29], [49], [50].

We assume that an ML system is a definite program $\mathcal{P}$. Our definition of a definite program considers $\mathcal{P}$ as having a set of stages or series of steps that help in transforming a set of inputs into some desired outputs [28]. This definition of $\mathcal{P}$ also considers a system as a holistic system comprising of one or multiple systems, sub-systems or algorithms, capable of producing the desired outputs that enable making inferences and decisions [28]. Such systems include: supervised learning-supported classifiers and regression systems, unsupervised learning-supported clustering and labelling systems and, reinforcement learning systems including deep neural networks. Therefore, such a system would include definite symbols, definite functions, definite propositions, definite predicates, logical symbols, object variables and propositional variables [48], [50].

In the following sections, $\mathcal{C}$ denotes a constant, p represents a predicate symbol and $\mathbb{S}$ shows a human population having an individual human represented as 's'. In this paper, $\mathbb{V}$ shows a first-order variable and $\mathbb{B}$ is the background knowledge. A human possessing the background knowledge $\mathbb{B}$ is considered tantamount to a definite program $\mathcal{P}$. D_n denotes a definition D having a number n and $\mathbb{D}$ in this paper denotes a domain. Having these notations borrowed fromm the previous works [7], [38], [41], [48], the following subsections describe all measurable parameters belonging to each of the three vectors forming the 3D AXAI measurement space shown in Fig. 2.

B. DEFINITIONS

D1: A predicate symbol, usually called in queries, is such that $p \in \mathbb{P}$. Declared in a ML system ($\mathcal{P}$) is p, which is *public* with respect to a human population $\mathbb{S}$ if p forms part of the background knowledge $\mathbb{B}$ of each human s ($s \in \mathbb{S}$). Otherwise, p is a *private* predicate symbol contained in $\mathcal{P}$.

D2: Let $\mathbb{H}$ be a system. If the background knowledge $\mathbb{B}$ of $\mathcal{P}$ is extended such that $\mathbb{B} \cup \mathbb{H}$ is formed, then the predicate symbol $p \in \mathcal{P}$ becomes a predicate invention since p was originally defined in $\mathbb{H}$ but not in $\mathbb{B}$.

D3: The AXAI capability denoted by C_{AXAI} is a representation in a three-dimensional space. We posit that C_{AXAI} comprises of three independent vectors: $\mathbb{C}$ (comprehensibility), P_A (predictive accuracy) and S_A (system accountability). Also, each one of the three vectors $\mathbb{C}$, P_A and S_A comprises of three independent components whose details are given in the following definitions ***D4 – D6C***.

D4: The comprehensibility $\mathbb{C}$ of $\mathcal{P}$ in the context of a human population $\mathbb{S}$ is represented as $\mathbb{C}(\mathbb{S}, \mathcal{P})$ where $\mathbb{C}$ is a vector comprising of three components: the inspection time (T_{it}), the predicate recognition time (T_{pr}) and the time required to name a predicate (T_{pn}) such that:

$$\mathbb{C}(\mathbb{S}, \mathcal{P}) = \sqrt{(T_{it}^2 + T_{pr}^2 + T_{pn}^2)} \tag{1}$$

Here naming, an important goal of learning, means expressing the ''object-property'' relation, and naming object and/or groups of objects. Hence, the comprehensibility of $\mathcal{P}$ in the

context of AXAI refers to the mean readiness ($\mathbb{R}$) of a human s ($s \in \mathbb{S}$) for applying the knowledge given in program $\mathcal{P}$ to assign a public name to a new definition q with respect to the domain $\mathbb{D}$ after inspecting $\mathcal{P}$ for times T_{it}, T_{pr} and T_{pn}. It is worth mentioning that clustering, primarily an unsupervised method can also be combined with supervised learning in a system such that each cluster can be given an intended definition [49].

D4A: The inspection time (T_{it}) is the mean time that a human s ($s \in \mathbb{S}$) requires for inspecting the information presented by $\mathcal{P}$ before using the knowledge provided by $\mathcal{P}$ for solving a new problem within the domain $\mathbb{D}$.

D4B: The predicate recognition time (T_{pr}) is the mean time that a human s ($s \in \mathbb{S}$) requires for assigning a correct public name to a predicate symbol p within the domain $\mathbb{D}$.

D4C: The predicate naming time (T_{pn}) is the mean time that a human s ($s \in \mathbb{S}$) requires for naming a predicate symbol p presented as a privately named definition q within the domain $\mathbb{D}$ for correctly assigning a public name to the predicate symbol p after inspecting $\mathcal{P}$.

D5: The predictive accuracy P_A of a system $\mathcal{P}$ with respect to a human population $\mathbb{S}$ and a domain $\mathbb{D}$ is represented as $P_A(\mathbb{S}, \mathcal{P})$ where P_A is a vector comprising of three components; $r_{tst-trn}$ (the ratio of test data size and training data size), d_{trn} (the training data size) and O_{fp} (number of occurrences of false-positive results) such that:

$$P_A(\mathbb{S}, \mathcal{P}) = \sqrt{r_{tst-trn}^2 + d_{trn}^2 + O_{fp}^2} \tag{2}$$

The predictive accuracy in the context of AXAI refers to the mean ability ($\mathbb{A}$) of a human s from a population $\mathbb{S}$ to correctly name a predicate symbol p presented as a privately named description q with respect to the domain $\mathbb{D}$.

D5A: The ratio of the size of data used for testing and training the system $\mathcal{P}$, expressed as $r_{tst-trn}$ with respect to a domain $\mathbb{D}$ is an indicator of the level of rigour $\mathbb{L}_{Rig}$ applied in training and testing the program $\mathcal{P}$ for enabling correct naming of a predicate symbol p represented as a privately named definition q within a domain $\mathbb{D}$.

D5B: The absolute size of data (d_{trn}) used in training an AI system $\mathcal{P}$, with respect to a domain $\mathbb{D}$ is an indicator of the exposure of the program $\mathcal{P}$ for correctly naming a predicate symbol p represented as a privately named definition q within the domain $\mathbb{D}$. The value associated with d_{trn} indicates the ability of $\mathcal{P}$ to identify variations in new samples of data belonging to the domain $\mathbb{D}$.

D5C: Occurrences of false-positive naming of predicate symbols $p_n(n = 1, 2, 3, \ldots, n)$ presented as named definitions $q_n(n = 1, 2, 3, \ldots, n)$ observed while testing a system $\mathcal{P}$ is expressed as O_{fp}. In the context of AXAI, O_{fp} is an indicator of the ability of the system $\mathcal{P}$ to compare the models used in its training with models of new and unknown symbols belonging to the same domain. The magnitude of O_{fp} therefore indicates the level of errors built into the system $\mathcal{P}$ for accurately naming a predicate symbol p represented as a named inference q with respect to a domain $\mathbb{D}$. Please note that, to the best of authors' understanding, the cited literature highly recommends integration of a human component in assessing the system accuracy [27], [29], [30], [48].

TABLE 1. The AXAI capability assessment parameters and their data types proposed for estimating the explainability attributes and determining the overall AXAI capability of definite programs.

S No.	Parameter	Measure
1.	The norm of the vector of comprehensibility $\|\|\mathbb{C}\|\|$	Integer
2.	The norm of the vector of predictive accuracy $\|\|P_A\|\|$	Integer
3.	The norm of the vector of system accountability $\|\|S_A\|\|$	Integer
4.	The inspection time (T_{it})	Score (integer)
5.	The predicate recognition time (T_{pr})	Score (integer)
6.	The predicate naming time (T_{pn})	Score (integer)
7.	The ratio of the test data and the training data($r_{tst-trn}$)	Score (integer)
8.	The absolute size of the training data (d_{trn})	Score (integer)
9.	The number occurrences of the false-positive naming (O_{fp})	Score (integer)
10.	The mean score of inspect-ability of input signals (I_{in})	Score (integer)
11.	The mean score of inspect-ability of the processed data (I_{pro})	Score (integer)
12.	The mean score of inspect-ability output cues (I_{out})	Score (integer)

D6: The system accountability S_A of a system $\mathcal{P}$ with respect to a human population $\mathbb{S}$ is represented as $S_A(\mathbb{S}, \mathcal{P})$ where S_A is a vector comprising of three components: I_{in} (inspectability of input models or cues), I_{pro} (inspectability of data being processed) and I_{out} (inspectability of output models or cues) such that:

$$S_A(\mathbb{S}, \mathcal{P}) = \sqrt{I_{in}^2 + I_{pro}^2 + I_{out}^2} \tag{3}$$

The system accountability in the context of AXAI refers to the mean accuracy with which a human s ($s \in \mathbb{S}$) can realize any occurrences of constants $\mathcal{C}$, predicate symbols $\mathbb{P}$ and variable $\mathbb{V}$ to correctly recognize a new definition with respect to the domain $\mathbb{D}$.

D6A: The mean score of inspectability I_{in} of input models/cues, supplied as named definitions $q_n(n = 1, 2, 3, \ldots, n)$ to a program $\mathcal{P}$ is an indicator of the mean clarity observed by a human s ($s \in \mathbb{S}$) with which s would inspect the definition q before q is named as a predicate symbol p with respect to the domain $\mathbb{D}$. Therefore I_{in} reflects on the form and format of the input models/cues with definitions $q_i(i = 1, 2, 3, \ldots, i)$ and predicate symbols $p_j(j = 1, 2, 3, \ldots, j)$.

D6B: The mean score of inspectability of data after being processed, I_{pro} in a program/system $\mathcal{P}$ is an indicator of the mean clarity of the processed (or conditioned) definition q as observed by a human from a population $\mathbb{S}(s \in \mathbb{S})$. Hence mean I_{pro} is the mean clarity with which a human s inspects the processed form of definition of q before q is named as a predicate symbol p with respect to a domain $\mathbb{D}$. Therefore I_{pro} reflects on the form and format of the intermediary models of definitions $q_n(n = 1, 2, 3, \ldots, n)$ while any q_n is being transformed into a predicate symbol p.

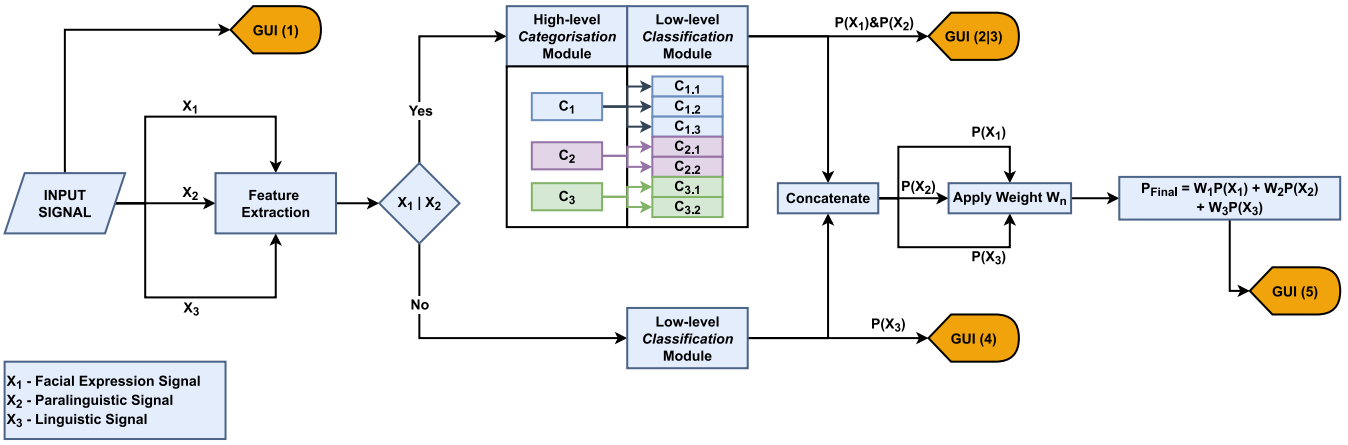


FIGURE 3. The high-level system architecture of the ASAM highlighting how signals $\{X_1, X_2, X_3\}$ traverse through the system in the processing and output stages. The mechanism of displaying various information and the nature of data displayed to users via the graphical user interface (GUI) are shown in Fig. 4.

D6C: The mean inspectability score of output signals I_{out} provided by $\mathcal{P}$ is an indicator of the mean clarity of the definition q as observed by a human s ($s \epsilon \mathbb{S}$) with which s would inspect the output definition of q for naming it as a predicate symbol p with respect to a domain $\mathbb{D}$. Therefore, this parameter reflects on the form and format of the output models/cues of definitions $q_n(n = 1, 2, 3, \ldots, n)$ after q is processed by the program $\mathcal{P}$.

C. HYPOTHESES

We now present the set of hypotheses that enable the assessment of an ML program $\mathcal{P}$ in terms of its AXAI capability.

Hypothesis 1: The comprehensibility ($\mathbb{C}$) in the context of AXAI capability refers to the mean readiness ($\mathbb{R}$) of a human s to use the knowledge gained after understanding the program $\mathcal{P}$ to accurately solve new problems in a domain $\mathbb{D}$. We hypothesize that comprehensibility is directly proportional to the mean readiness of s $(s \epsilon \mathbb{S}) : \mathbb{C} \infty \mathbb{R}$.

Hypothesis 2: The larger the norm of the comprehensibility vector $||\mathbb{C}||$, the more comprehensible an ML program $\mathcal{P}$ is.

Hypothesis 3: The predictive accuracy (P_A) of a human s from a population $\mathbb{S}$ to correctly name a predicate symbol p given as a privately named definition q is directly proportional to the mean ability ($\mathbb{A}$) of an individual s $(s \epsilon \mathbb{S}) : P_A \infty \mathbb{A}$.

Hypothesis 4: The larger the norm of the predictive accuracy vector $||P_A||$ of a definite program $\mathcal{P}$ the better predictive accuracy $\mathcal{P}$ offers.

Hypothesis 5: The system accountability (S_A) refers to the mean accuracy ($\overline{\mathbb{A}_{cc}}$) with which a human s from a population $\mathbb{S}$ $(s \epsilon \mathbb{S})$ recognizes any occurrences of constant $\mathbb{C}$, predicate symbol $\mathbb{P}$ and variable $\mathbb{V}$ to correctly recognize a new model belonging to the domain $\mathbb{D} : S_A \infty \overline{\mathbb{A}_{cc}}$.

Hypothesis 6: The larger the norm of the system accountability vector $||S_A||$ of a definite program $\mathcal{P}$ the better the level of incorporated accountability in the program $\mathcal{P}$.

Hypothesis 7: The inverse of the mean time $\frac{1}{T_{it}}$ that a human s $(s \epsilon \mathbb{S})$ requires for inspecting the information presented by the program $\mathcal{P}$ before using the knowledge provided by that $\mathcal{P}$ for solving a new problem in domain $\mathbb{D}$, is directly proportional to the presentation quality ($\mathbb{Q}_p$) of $\mathcal{P}$ given as $\frac{1}{T_{it}} \infty \mathbb{Q}_p$.

Hypothesis 8: The inverse of the mean predicate recognition time $\frac{1}{T_{pr}}$ that a human s $(s \epsilon \mathbb{S})$ requires to assign a correct public name to a predicate symbol p in a system is proportional to the ability ($\mathbb{A}_p$) of recognizing and accurately assigning a public name to a predicate symbol p. Hence, $\frac{1}{T_{pr}} \infty \mathbb{A}_p$. Note that an incorrect assignment of a public name to a predicate symbol should not be counted and considered in assessing a system.

Hypothesis 9: The ratio of the size of test data and the size of the training data ($r_{tst-trn}$) of a program $\mathcal{P}$ is directly proportional to the level of rigour ($\mathbb{L}_{Rig}$) applied in training and testing $\mathcal{P}$ with respect to a domain $\mathbb{D}$, hence, $r_{tst-trn} \infty \mathbb{L}_{Rig}$.

Hypothesis 10: The mean score of inspectability of data after being processed (I_{pro}) shows how understandable the intermediary data representation/models ($\mathbb{F}_{mod}$) in a definite program $\mathcal{P}$ are. Thus, $I_{pro} \infty \mathbb{F}_{mod}$.

Hypothesis 11: The instances of the false-positive naming of predicate symbols with privately named definitions O_{fp} indicate the level of errors E_{bp} built into the program $\mathcal{P}$ with respect to a domain $\mathbb{D}$, hence, $O_{fp} \infty E_{bp}$.

Table 1 presents a complete list of measurable parameters used to determine the overall AXAI capability of a definite program $\mathcal{P}$.

IV. ASSESSING AXAI CAPABILITIES OF THREE ML SYSTEMS

In order to test the relevance of the AXAI capability framework, the AXAI scores of three ML systems were calculated. The following subsections present details of the three ML systems whose C_{AXAI} scores were estimated using the proposed AXAI framework.

TABLE 2. Guidelines for assessing, scoring and determining the AXAI capabilities of the three definite programs. The ASAM and ASAM-2 use multimodal input (facial expressions and speech cues) and the DAS uses facial thermal variations to analyse and recognize affective states.

		Scoring Criteria		
S No.	**Parameter**	**0.0 - 1.0**	**2.0 - 3.0**	**4.0 - 5.0**
4	Inspection Time	Information appears to be very difficult and takes a long time to understand	Information takes some efforts and time to understand	Information is easy to understand with minimal efforts
5	Predicate Recognition Time	A human would take very long time to interpret the output	A human would take some time to interpret the output	A human would quickly interpret the output
6	Predicate Naming Time	A human would take very long time to use the inferences for naming another predicate within the domain	A human would take some time to use the inferences for naming another predicate within the domain	A human would quickly name another predicate within the domain using the inferences
7	Test/Training Data Ratio	$0 : 0.1 \leq r_{tst-trn} \leq 0.9$ $1 : 1.0 \leq r_{tst-trn} \leq 2.0$	$2 : 2.1 \leq r_{tst-trn} \leq 3.0$ $3 : 3.1 \leq r_{tst-trn} \leq 4.5$	$4 : 4.6 \leq r_{tst-trn} \leq 4.9$ $5 : r_{tst-trn} = 5.0$
8	Training Data Absolute Size	$0 : d_{trn} \leq 5n_{names}$ $1 : 5n_{names} < d_{trn} \leq 10n_{names}$	$2 : 10n_{names} < d_{trn} \leq 50n_{names}$ $3 : 50n_{names} < d_{trn} \leq 100n_{names}$	$4 : 100n_{names} < d_{trn} \leq 1000n_{names}$ $5 : d_{trn} \geq 1000n_{names}$
9	False-positive naming occurrences	$0 : O_{fp} \geq 50\%$ of the time $1 : 40\% < O_{fp} \leq 50\%$ of the time	$2 : 30\% < O_{fp} \leq 40\%$ of the time $3 : 25\% < O_{fp} \leq 30\%$ of the time	$4 : 10\% < O_{fp} \leq 25\%$ of the time $5 : O_{fp} \leq 10\%$ of the time
10	Inspect-ability of input signals	Explanations are not clear to users	Explanations are somewhat clear to users	Explanations are clear to users
11	Inspect-ability of intermediate data stages	Intermediate data cannot be seen or cannot be interpreted	Some of the intermediate data cannot be seen or cannot be interpreted	The intermediate data can be seen or interpreted
12	Inspect-ability of output signals	Output information is nondescript and hard to understand/interpret	Output information is some-what descriptive and takes some time to understand/interpret	Output information is descriptive and easy to understand/interpret

A. AN AFFECTIVE STATE ASSESSMENT MODULE (ASAM)

The first assessed ML system was designed to have the proposed AXAI framework built into it. The Affective State Assessment Module (ASAM) is a multimodal definite system [53] implemented as a portable affective state assessment sub-system for integration into robotic systems. The tested version of the ASAM is an improved system of our previously developed and published system [53]. The ASAM was developed for real time multimodal analysis of facial expressions and speech for assessing affective states. The design and implementation of the ASAM is detailed in an accompanying paper entitled ''Toward Accountable Explainable Artificial Intelligence Part two: The Framework Implementation'' published in this journal [55]. The provisions of explainability and accountability in the ASAM were ensured by adding the AXAI features listed in Table 1. Figure 3 shows the high-level architecture of the ASAM.

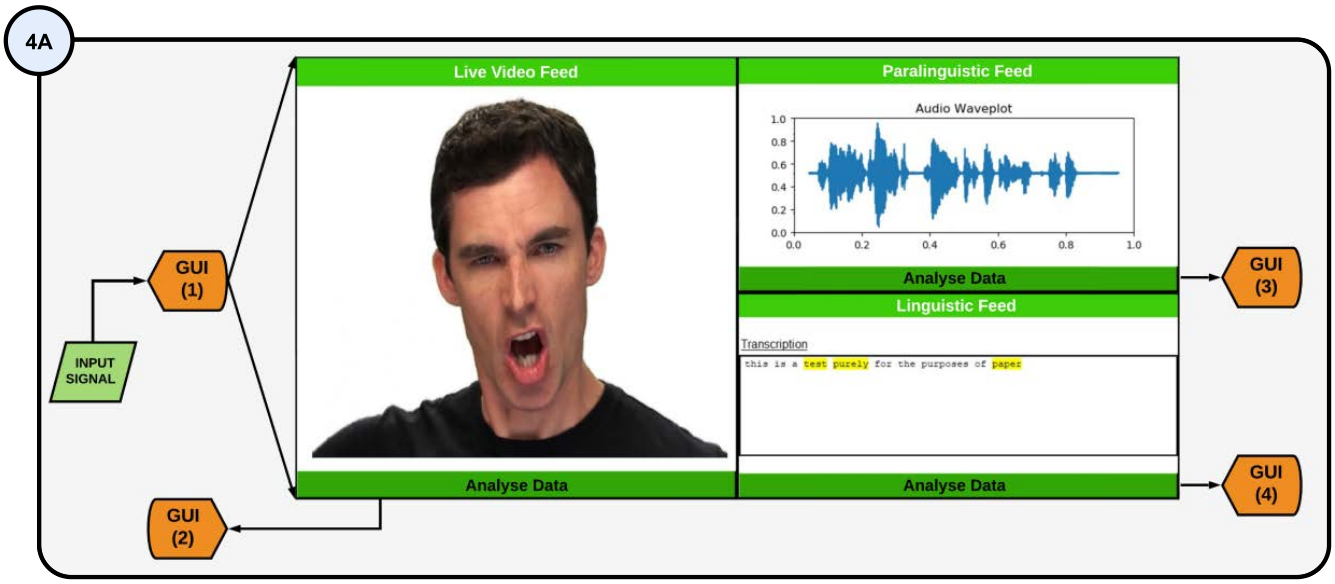
As the ASAM was designed to showcase AXAI capabilities in an affective state assessment system, it provides transparency by showing the input and feedback data and graphical and tabular information. As such, it is capable of providing users with scrutiny and debugging opportunities. The explanations are made available through display of Bayesian probability measures and high-level feature attributions. The three components of accountability viz., inspectability of input cues, inspectability of data being processed and inspectability of output cues were built into an intuitive and user-centered Graphical User Interface (GUI). Figure 4 shows input, data under processing and output information that ASAM presents to users through its GUI.

B. ASSESSMENT OF THE AXAI CAPABILITIES OF ASAM

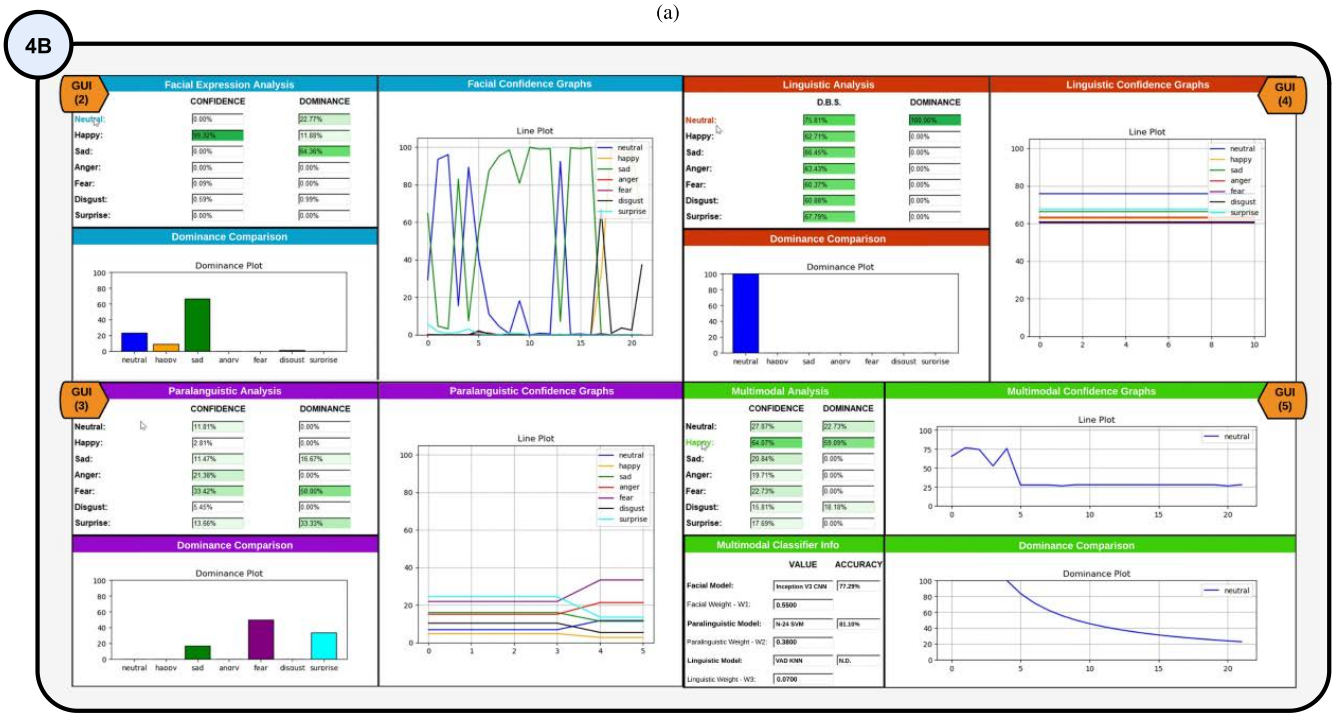
Ten qualified industry professionals and postgraduate students who were well-versed with ML and other AI-supported systems had volunteered to assess the AXAI capabilities of the ASAM. The parameters outlined in Table 1 were used for assessment of the AXAI capabilities. During an introduction session, these assessors who were educated in the fields of engineering, social science and psychology were briefed and informed on the objectives and outcomes of the assessment. After the briefing, participants were given ASAM's system user manual. Assessors had the opportunity to use the ASAM before starting to assess its functionality. The ASAM assessors tested the ASAM for an average time of twenty minutes. While testing, assessors awarded scores for parameters 4-6 and 10-12 on a 0-to-5 scale detailed in Table 2. Assessing the ASAM on parameters 7-9 was not required as these scores were supposed to be provided by the team of system designers. The scores were normalised and converted to unit vector forms (in the range of 0 to 1) allowing to delineate the AXAI-capabilities of the ASAM in a 3D space as discussed in previous sections and visualised in Fig. 2.

C. AN ENHANCED AFFECTIVE STATE ASSESSMENT MODULE ASAM-2

The second system tested for its AXAI capabilities was a modified and enhanced version of the ASAM called ASAM-2. We designed ASAM-2 as a continuous assessment tool capable of classifying 114 unique states across affective speech and facial expression signals using a hierarchical classification approach. In ASAM-2, a combination of 42 ternary/binary models was used. Similar to its predecessor,



(a)



(b)

FIGURE 4. **(A) The ASAM GUI. The home screen shows the mechanism of displaying the input information. The GUI shows data pertaining to all three input signals. This figure also shows how the lower-level functions of the system can be accessed from the home window (shown in Figure 4B). This GUI window is shown to users upon execution of the software. The image shown in the frame was taken from the RAVDESS dataset [56]. (B) The shown windows help in monitoring the ASAM's and allow inspecting the processing and output information. Note: CYAN = Facial Expression Analysis, ORANGE = Linguistic Analysis, PURPLE = Paralinguistic Analysis and LIME = Multimodal Analysis. All windows are executed on separate threads allowing for parallel processing and viewing of information.**

ASAM-2 is a real-time embedded system capable of being added to an existing robotic system for affective state assessment of humans.

At each level of classification, ASAM-2 uses different decision-making protocols to discern between the affective states. ASAM-2 uses data on: affective state groups, temporal phases, affective state intensities and discrete affective state models. As shown in the flowchart in Fig. 5, all classification results and intermediate information are displayed to the user via the GUI. The GUI in ASAM-2 was improved and redeveloped from the ground-up and was different to those shown in Figs. 4A and 4B.

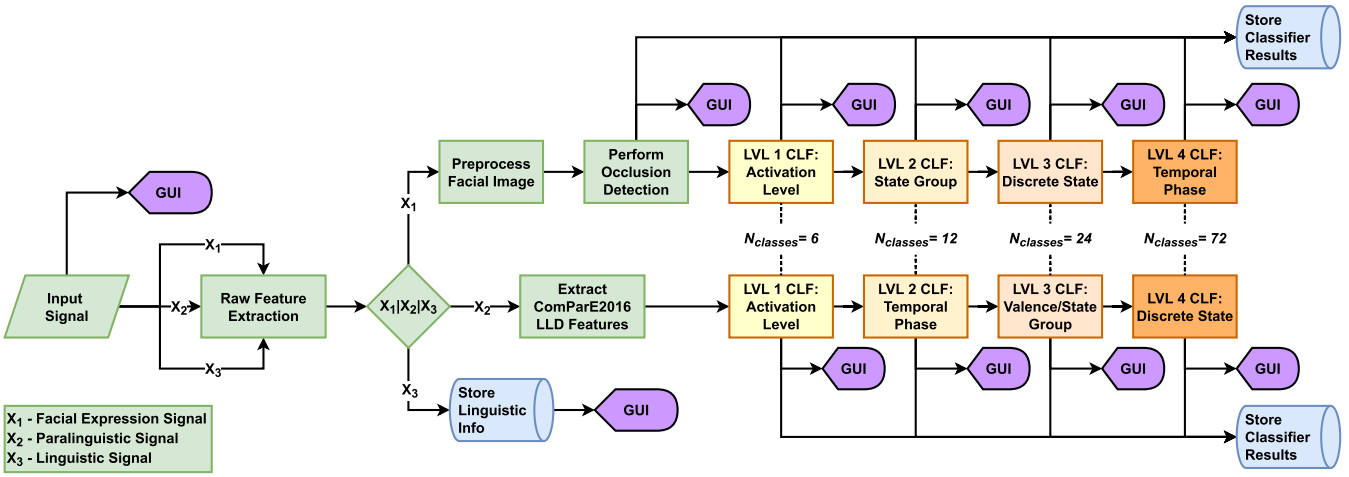


FIGURE 5. Flow of execution and a high-level description of ASAM-2. The hierarchical architecture shows various steps leading to classification of affective states in ASAM-2. During the process the originally invisible, intermediate and processed data stages are gradually revealed to users through the GUI. Each stage of the classification process is shown to users.

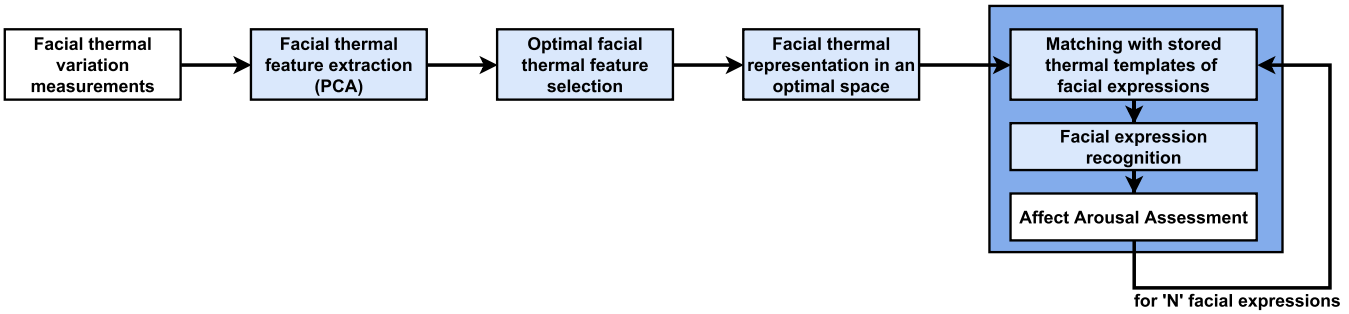


FIGURE 6. The flow of execution is DAS and a visual description of its functional architecture. The white boxes show visible and inspectable information. The blue boxes show the back-end, invisible processes and information.

The decision to expand the classification capabilities and reveal ASAM-2's decision-making processes was at the core of its design process. Thus, ASAM-2 has enhanced levels of accountability and comprehensibility vis-à-vis maintaining a high degree of classification accuracy. Through an iterative design process and feedback received during the ASAM assessment, ASAM-2 was equipped with improved AXAI capabilities. Results are discussed in section V to inform readers about ASAM-2's AXAI capabilities.

D. ASSESSMENT OF ASAM-2's AXAI CAPABILITIES

ASAM-2 was assessed by eight trained assessors. The assessors who volunteered for ASAM-2's assessment were well-trained and had professional background in applied science, engineering, and behavioural studies. All assessors were given a brief to introduce ASAM-2 before providing them access to ASAM-2. On average, assessors spent approximately 17 minutes in assessing ASAM-2 AXAI capabilities. Table 5 highlights the 5-point scores given to ASAM-2 by the users while assessing its AXAI capabilities.

E. A SYSTEM FOR DYNAMIC ASSESSMENT OF AFFECTIVE STATES AND AROUSAL LEVELS

A third ML system tested for its AXAI capabilities was also a definite program that was designed to work as a two-step system of dynamic assessment of affective states and arousal levels called DAS [54]. It uses thermal infrared images (TIRI's) of facial expressions and was not designed to have AXAI capabilities built into it. Hence, post-production assessment of AXAI capabilities was performed in this case.

The DAS would first analyse TIRI's for examining the hæmodynamic variations caused by changes in affective states. The algorithmic execution of DAS starts by analysing the hæmodynamic variations along the facial muscles. The observed variations are used to estimate the affect induced facial thermal variations. In the first step, 'between-affect' and 'between-arousal-level' variations are subject to Principal Components Analysis (PCA). The most influential principal components are then used to cluster the features belonging to different affective states. Subsequently each set of thermal features is assigned to an affective state cluster. In the second step, the affective state clusters are partitioned into high, medium and mild arousal levels. The distance between a test TIRI and centroids of sub-clusters at three arousal levels belonging to a single affective state, identified from the first step, is used to determine the arousal level of the identified affective state. Figure 5 shows the flow of execution in DAS - white boxes show the visible and inspectable information and blue boxes show the information hidden in the program.

F. ASSESSING AXAI CAPABILITIES OF THE DAS

A postdoctoral fellow and seven postgraduate students who were trained in AI and ML volunteered to assess the AXAI capabilities of the DAS. As in the previous cases, the AXAI capabilities were assessed using parameters listed in Table 1. All volunteers were informed on the objectives and outcomes of the assessment and were also given the program code, the executable program, relevant data and publications. The average time each assessor spent on running and testing DAS was recorded to be 27 minutes. As in the previous cases, assessors awarded scores for parameters 4-6 and 10-12 on a 0-5 scale as highlighted in Table 4.

V. RESULTS

A. AXAI CAPABILITY ASSESSMENT OF THE ASAM

The predictive accuracy components given in Table 2 were known to the system developers as they were designing the ASAM for having the AXAI capabilities. The test/training data ratio ($r_{tst-trn}$) of the ASAM was kept as 80 : 20. A similar ratio had been used in some previous works [53], [54]. The ASAM's $r_{tst-trn}$ score therefore resulted in a normalized value of 1.0.

The ASAM used 700 facial images from the extended Cohn-Kanade (CK+) dataset [59] and 1400 speech samples from the Toronto Emotional Speech Set (TESS) [60] to train the facial expression and paralinguistic speech classifiers. The facial expression classifier contained approximately 100 samples per label (7 labels/classes) giving a score of $4/5 = 0.8$. In comparison, the paralinguistic speech classifier contained approximately 200 samples per class (7 labels/class) hence it would score a $4.11/5 = 0.822$, calculated by mapping the range of $100 - 1000n_{names}$ per class to a score range of 4-5 as per Table 2. The *average* score for the d_{trn} parameter for the ASAM was therefore $4.055/5$ resulting in a normalised value of $d_{trn} = 0.8111$.

The false-positive naming occurrences 'O_{fp}' could be determined during validation tasks. The ASAM's paralinguistic speech and facial expression classifiers were validated on the Ryerson Audio-visual Database of Emotional Speech and Song (RAVDESS), a multimodal dataset containing affective speech and facial expression data [56]. The data were unknown at the time of training, thus the validation experiments provided test results on the ASAM's ability to assess foreign and real-life data. Through validation tasks, the ASAM achieved predicate naming errors of 22.71% and 18.90% respectively for the facial expression and paralinguistic classifiers. Hence, an average naming error of 20.805% resulted in a score close to 4 being observed. Specifically, the O_{fp} was calculated to be $4.2797 = 0.8559$. Given $r_{tst-trn} = 1.0$, $d_{trn} = 0.811$ and $O_{fp} = 0.8559$, as per (2) the norm of the predictive accuracy vector was:

$$\begin{aligned} P_A(\mathbb{S}, \mathcal{P}) &= \sqrt{r_{tst-trn}^2 + d_{trn}^2 + O_{fp}^2} \\ &= \sqrt{1^2 + 0.811^2 + 0.8559^2} \\ &= \sqrt{1 + 0.657721 + 0.732565} \\ &= 1.54606 \end{aligned}$$

The score parameters 4-6 and 10-12 in Table 1 were respectively used to determine the system accountability and comprehensibility vector norms. The mean values were determined through user experiences and surveys of the system. The system comprehensibility was found to be greater than system accountability as reported in Table 3. The $||\mathbb{C}||$ and $||S_A||$ values were calculated using equations (1) and (3).

The data in Table 3 highlights very good comprehensibility results for the ASAM, with inspection time 'T_{it}' being the highest, (average score $T_{it} = 3.95$). The lowest component in terms of comprehensibility was the predicate naming time (average score $T_{pn} = 3.05$). Given user responses, the general feedback suggested that predicate naming was more difficult and time consuming for assessors when compared to other comprehensibility factors and should be addressed for future works.

We found the system accountability scores to be comparatively lower than the scores for comprehensibility, specifically in regard to the inspectability of the data processing stages 'I_{pro}'. User feedback suggested that while the ASAM's rule-based expert system output showed how a combination of signals could be used to report a multimodal output, the ASAM could be improved by providing a better display of the processed data for the facial expression, paralinguistic and linguistic channels. In comparison, the inspectability of inputs and outputs were received positively, highlighting the ASAM's ability to report the system's initial and final states.

The ASAM's GUI, shown in Fig. 4, was designed to display some processed data stage information in the form of associated weights of the rule-based expert system output. Applying weight numbers to facial expression, paralinguistic and linguistic speech classification results allows for the display of tabular and graphical rule-based system outputs i.e., the transformation of data from input, to processed, to output.

Using the reported scores and the consequential location of the ASAM within the 3D space of $\mathbb{C}$, P_A and S_A, we concluded that improving I_{pro}-related features would greatly enhance the user experience and AXAI capabilities of the system. In summary, the ASAM's scores for comprehensibility, predictive accuracy and system accountability were respectively: $\mathbb{C} = 1.203$, $P_A = 1.546$, and $S_A = 1.139$. Thus, the three vector norms provide an estimate of the ASAM's AXAI capabilities, allowing us to visualise the ASAM's position within the 3D axes as shown in Fig. 7.

Using these results, we could compare the AXAI capabilities of the three systems in terms of their levels of explainability, predictive accuracy and comprehensibility. However, the accountability score suggests that more attention should be paid to the ASAM's accountability components. The estimated S_A score suggested that the information being processed I_{pro} will not suffice user requirements. Overall, the proposed framework provided a practical and easy to follow method of assessing the AXAI capabilities of the ASAM.

TABLE 3. The ASAM users' AXAI capability scores on a 0-5- scale and the normalised scores. These user scores were used to determine the ASAM's accountability 'S_A' and comprehensibility '$||\mathbb{C}||$' capabilities.

S No.	Symbol	User ID 1	2	3	4	5	6	7	8	9	10	Average using 5 Point Score	Normalised Average Score
1	$\|\|\mathbb{C}\|\|$												1.203
3	$\|\|S_A\|\|$												1.139
4	T_{it}	4	4	4	4	5	3.5	4	5	4	2	3.95	
5	T_{pr}	5	3	3	4	3	3	4	4	1	2	3.2	
6	T_{pn}	3	3	3	2	4	3.5	3	5	2	2	3.05	
10	I_{in}	4	5	3	3	2.5	3	3	5	2	5	3.55	
11	I_{pro}	2	1	2	3	0	2	2	4	2	1	1.9	
12	I_{out}	3	4	4	3	5	4	4	5	3	3	3.8	

TABLE 4. Users' experience scores and their normalised scores for ASAM-2 on a 0-5 scale. These scores were used to determine ASAM-2's system accountability 'S_A' and comprehensibility '$||\mathbb{C}||$' capabilities.

S No.	Symbol	User ID 1	2	3	4	5	6	7	8	Average using 5 Point Score	Normalised Average Score
1	$\|\|\mathbb{C}\|\|$										1.275
3	$\|\|S_A\|\|$										1.453
4	T_{it}	3	5	3	5	5	4	3	4	4.0	
5	T_{pr}	5	3	3	3	4	3	1	4	3.25	
6	T_{pn}	5	5	2	4	4	4	2	4	3.75	
10	I_{in}	5	5	4	4	4	4	4	5	4.38	
11	I_{pro}	5	3	2	4	5	3	3	3	3.5	
12	I_{out}	5	5	4	4	5	5	4	5	4.63	

TABLE 5. Users' experience scores for DAS on a 0-5 scale. Normalised scores are also reported in the table. The reported scores were used to assess accountability 'S_A' and comprehensibility '$||\mathbb{C}||$' capabilities of the DAS.

S No.	Symbol	User ID 1	2	3	4	5	6	7	8	Average using 5 Point Score	Normalised Average Score
1	$\|\|\mathbb{C}\|\|$										0.333
3	$\|\|S_A\|\|$										0.489
4	T_{it}	0.5	1	1	0.5	1	1	1	1	0.87	
5	T_{pr}	0.2	0.2	0	0.1	0	0	0	0	0.06	
6	T_{pn}	0.5	1.5	0.5	1	2	1.5	2	2	1.37	
10	I_{in}	1.5	2	2	2	2	2	2	2	1.93	
11	I_{pro}	0	0	0	0	0	0	0	0	0	
12	I_{out}	1.5	2	2	1	1	2	1	1	1.43	

B. AXAI CAPABILITY ASSESSMENT OF ASAM-2

Deriving the predictive accuracy vector components: Firstly, the test/train data ratio '$r_{tst-trn}$' was kept at 80 : 20 similar to the ASAM, resulting in a normalized score of 1. The RAVDESS dataset was used for both training and validation of the facial expression and paralinguistic classification subsystems. The facial expression classifiers were trained using approximately 1500-2500 samples per class (total = 76270 samples) giving it a d_{trn} score of 5/5 = 1.0. The paralinguistic speech classifiers in comparison, were trained using 96 samples per class (total = 3744 samples) thus achieving a score of 3.92/5 = 0.784, which was calculated by mapping the range of 50-100n_{names} per class to a score range of 3-4. Thus, the average d_{trn} score for the ASAM-2 was 4.46/5 = 0.892. Finally, the O_{fp} metric can be derived through validation tasks, with ASAM-2 achieving respective naming errors of 16.93% and 4.10% for facial expression and paralinguistic speech classifiers, resulting in an average naming error of 10.52% across the systems classifiers, which equates to a score of 4.965/5 = 0.993. Given: $r_{tst-trn} = 1.0$, $d_{trn} = 0.892$ and $O_{fp} = 0.993$, ASAM-2's predictive accuracy is calculated using (2) as:

$$\begin{aligned} P_A(\mathbb{S}, \mathcal{P}) &= \sqrt{r_{tst-trn}^2 + d_{trn}^2 + O_{fp}^2} \\ &= \sqrt{1^2 + 0.892^2 + 0.993^2} \\ &= \sqrt{1 + 0.795664 + 0.986049} \\ &= 1.66785 \end{aligned}$$

The scores derived in Table 5 determine ASAM-2's comprehensibility and system accountability scores i.e.: $\mathbb{C} = 1.275$ and $S_A = 1.453$, we can see that the changes made throughout the design process using feedback from the ASAM shows significant improvements in all three vectors when we compare their scores. Most significant, is the improvement in the predicate naming time

'T_{pn}' (3.05 → 3.75) and the inspect-ability of data processing stages 'I_{pro}' (1.9 → 3.5), which significantly enhanced the user experience, and ultimately showed how the AXAI framework could be used to improve the usability, transparency and explainability of AI and ML systems.

Deriving the S_A, P_A and $\mathbb{C}$ scores for the ASAM, ASAM-2 and DAS discussed earlier allow how to plot them within a three-dimensional AXAI space and compare their AXAI capabilities as visualised in Fig. 7. Analysing this figure, we see that ASAM-2 has the highest level of AXAI capability compared with the other two systems. We could see how ASAM was improved in terms of the nine factors of the proposed AXAI capability framework. It could be argued that the proposed AXAI capability framework provided a systematic method of assessing and comparing ML systems for their respective levels of accuracy, accountability and explainability.

C. AXAI CAPABILITY ASSESSMENT OF THE DAS

The AXAI capability assessment results for DAS are given in Table 4. The predictive accuracy parameters were estimated using the system training and testing data. The test/training data ratio ($r_{tst-trn}$) of the DAS was 1:1 [54] resulting in a score of 5.0 (normalized value of 1). The training data size d_{trn} parameter score for the DAS was given as 2.0 and the score for occurrences of false positive results, O_{fp} was 4.0. Based on these parameter values, the predictive accuracy (P_A) of DAS was:

$$\begin{aligned} P_A(\mathbb{S}, \mathcal{P}) &= \sqrt{r_{tst-trn}^2 + d_{trn}^2 + O_{fp}^2} \\ &= \sqrt{1^2 + 0.4^2 + 0.8^2} \\ &= \sqrt{1 + 0.16 + 0.64} \\ &= 1.3416 \end{aligned}$$

The data in Table 4 suggest that DAS had a low level of comprehensibility and a less than average level of accountability. However, being a statistical classifier, it was able to offer a high level of predictive accuracy. Specifically, through the DAS scores, we report comprehensibility, predictive accuracy and system accountability values of: $\mathbb{C} = 0.333$, $P_A = 1.342$, and $S_A = 0.489$.

VI. DISCUSSION

The three definite (ML) programs assessed in the preceding sections were fundamentally different. The first system (ASAM) and its enhanced version ASAM-2 were designed to have the AXAI capability incorporated in them. ASAM-2, being an improved version of the ASAM, had improvements leading to better levels of accountability and comprehensibility. The third program (DAS) was a basic classification and clustering system that was not designed to have the AXAI capability incorporated. Despite their fundamental differences, the proposed AXAI capability framework allowed for assessing the three programs in terms of predictive accuracy, comprehensibility and accountability. Delineating the three

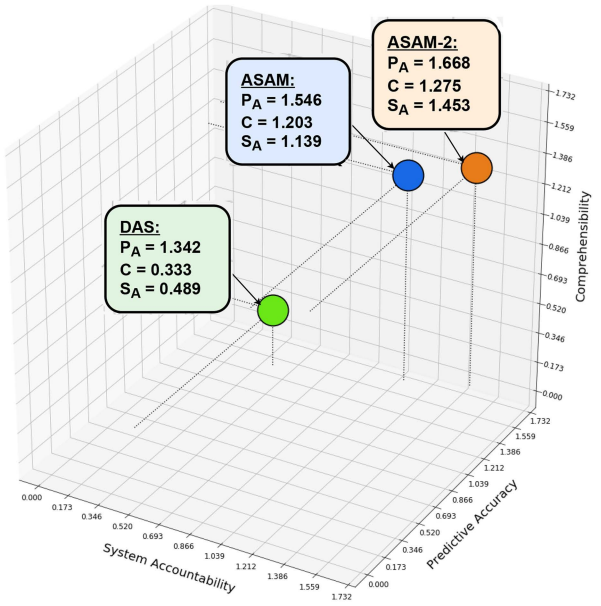


FIGURE 7. The AXAI capabilities of the ASAM (BLUE), ASAM-2 (ORANGE) and DAS (GREEN) systems are compared to show how the AXAI framework helps in assessing various ML systems. The three system are plotted in the tree-dimensional AXAI space. The placement of each circle shows system scores along the axes of comprehensibility, accountability and predictive accuracy making it easy for the system developer and system users to compare various AXAI aspects of the same system or multiple systems.

ML systems in a 3D AXAI capability space demonstrated that the proposed framework was helpful in system design and assessment of ML systems. Furthermore, the AXAI capability framework also provided an opportunity to systematically address ethical and professional issues, such as those highlighted in [40], and [55] while building ML systems. As evident in the above comparison, the nine measurable components of the AXAI capability framework ensured paying attention to system details, ethical responsibilities and moral duties during the conceptual design and functional analysis stages. Such manifestations have been desired in AI and ML systems for quite some time [41], [50]. However, the AXAI framework does not work as a purpose-built forensic framework would in tracing and combating any deviations from the expected system norms.

Building upon the XAI capability centred philosophical discussions in the literature [23], [42], [61], our proposed AXAI capability framework provides three sets of quantifiable parameters, each having three variables, for assessing levels of comprehensibility, accuracy and accountability. Through these parameters, the AXAI capability framework ensures incorporating important ethical, moral and legal safeguards in AI systems. This makes the proposed AXAI capability framework relevant and contemporary. The accuracy, comprehensibility and accountability measures also provide the required breadth and depth for designing, comparing and assessing AI systems in a domain-agnostic manner. Hence, incorporating the AXAI capability framework would not limit the system developer to follow a particular domain-specific method [6], [9], [12].

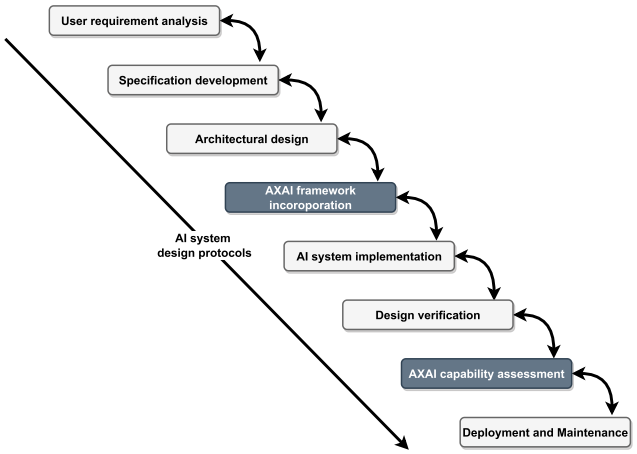


FIGURE 8. Mapping the process of incorporating the AXAI capability onto a typical software system design model for helping developers efficiently build and test an ML. The model would be beneficial for embedding AXAI capabilities while determining the user requirements and testing the system at various stages of the system life cycle. Dark boxes highlight stages where the AXAI framework design protocols would be added.

It is important to mention that the three components of the AXAI's comprehensibility vector rely on users' ability to inspect and understand information as we did not discuss or recommend any method of measuring the data inspectability in this part of the work. However, part two of our work [55] addresses the issue and recommends a collaborative system building approach that requires system developers and users to agree on the quality of inspectable information [64].

It would be safe to suggest that the proposed AXAI capability framework is step towards meeting DARPA's perceived goal of developing human-centred AI systems [64] as it provides such learning models and decision-making processes that would be shared, understood and trusted by the relevant communities [13].

Through the aforementioned AXAI capability assessments, we have demonstrated that system comprehensibility can be seen in terms of the mean readiness of a human to apply the knowledge acquired from an AI program and interpreting unknown problems within the domain.

We have modelled the predictive accuracy of an AI program in terms of the ratio of the test and training data, training data size and the number of false-positive results. Thus, predictive accuracy features would allow for estimating the ability of a human to correctly name a predicate symbol presented as a privately named description in a domain. It is important to signify that the predictive accuracy in the AXAI framework is domain-bound.

Finally, system accountability in the AXAI framework is reflected in the level of accuracy of a human's realization of occurrences of logical elements in an ML or AI system and would use them to solve a problem in a particular domain. The accountability, manifested through its three components (inspectability of input cues, processed data and, output cues) facilitates establishing a chain of responsibility. If any one or more of the three accountability components were not inspectable by users then the system design team could be held responsible for the shortcomings. However, if these components were inspectable then the user could be considered responsible for any negative consequences. Hence, accountability in our AXAI capability framework is assessed in an appropriate context [1], [34].

Because of the time limitations and the scope of this work, we could not test hypothesis 3 given in sub-section C of Section III. However, our inability to test the predictive accuracy (P_A) of a human s to correctly name a predicate symbol p given as a privately named definition q does not reflect on the applicability of the AXAI framework. Testing this hypothesis would require identifying and approaching domain experts to confirm if the hypothesis is verifiable and useful in the context of the affective computing systems assessed for this work.

VII. CONCLUSION

This work proposes a novel and easy to implement AXAI capability framework for designing, analysing and assessing machine learning systems. The proposed framework, as demonstrated through examples, was easy to incorporate, application-agnostic and useful in comparing and delineating various ML systems. While measuring AXAI capabilities, the proposed framework also provides a measure of non-explainability and addressed an issue raised in [15]. The measure of assessing the non-explainability is given as: $\textit{non-explainability} = 1 - \textit{explainability}$. Through the proposed AXAI framework, automated matching of 'levels of abstraction' [11] was also made possible as interpretations were connected with interpretations and explanans were aligned with explanans.

The proposed AXAI capability framework is based on the realization that 'fundamentally complex' prediction tasks would be influenced by developments in domain-specific tools and techniques. Hence, the AXAI framework provides an application-agnostic XAI capability incorporation mechanism. It operates at a higher-level and is not affected or influenced by developments in tools and techniques or domain-specific changes in professional practices.

As explicit in this paper and part two of this paper [55], the AXAI framework also provides design guidelines and encourages provision of separable and quantifiable parameters of accuracy, comprehensibility and accountability. This makes the proposed AXAI capability framework different from existing XAI incorporation methods. Part two of this paper shows how developers and practitioners would engage in the process of incorporating and evaluating the efficacy of the proposed framework. Also, translating the AXAI capabilities into a set of system design requirements is demonstrated in part two of this paper [55]. Together, the two papers will be useful in developing the system requirements and producing a design process model as shown in Fig. 8. The AXAI capability framework related stages of the ML and AI system design are explicitly shown in Fig. 8.

For building upon the initial success, the ML-centred AXAI capability framework can be extended to others AI systems. The framework needs to be tested on a larger set of existing systems. We anticipate that parts one and two of this work will initiate works on building more acceptable and accountable intelligent systems.

We do not claim that the nine elements used for measuring AXAI capabilities provide the best set of measurable elements. However, these nine elements provide a set of parsimonious, swift and effective AXAI capability measurements. Though the list of our proposed AXAI elements is not exhaustive, it would suffice the common comprehensibility, accuracy and accountability measurement requirements. Nonetheless, this list of AXAI elements needs more input from legal practitioners, AI experts, software developers and cognition scientists. Also, the AXAI framework is unable to specify if a system would require root-cause analysis or forensic tracing. Despite these limitations, the proposed AXAI capability framework, in its current state, provides foundations for moving toward accountable and explainable AI solutions. It would be innocuous to conclude that the AXAI capability framework promises an era beyond hypothesis-driven XAI capability frameworks.

REFERENCES

[1] B. Kim and F. Doshi-Velez, "Machine learning techniques for accountability," *AI Mag.*, vol. 42, no. 1, pp. 47–52, Apr. 2021.
[2] D. Michie, "Machine learning in the next five years," in *Proc. 3rd Eur. Conf. Eur. Work. Session Learn.*, Glasgow, U.K., 1988, pp. 107–122.
[3] U. Schmid, C. Zeller, T. Besold, A. Tamaddoni-Nezhad, and S. Muggleton, "How does predicate invention affect human comprehensibility?" in *Proc. 26th Int. Conf. Log. Program.*, London, U.K., 2017, pp. 52–67.
[4] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.
[5] M.-A.-T. Vu, T. Adalı, D. Ba, G. Buzsáki, D. Carlson, K. Heller, C. Liston, C. Rudin, V. S. Sohal, A. S. Widge, H. S. Mayberg, G. Sapiro, and K. Dzirasa, "A shared vision for machine learning in neuroscience," *J. Neurosci.*, vol. 38, no. 7, pp. 1601–1607, Feb. 2018.
[6] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in healthcare," in *Proc. Int. Conf. Cyber Situational Awareness, Data Anal. Assessment (CyberSA)*, Dublin, Ireland, Jun. 2020, pp. 1–2.
[7] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing design practices for explainable AI user experiences," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Honolulu, HI, USA, 2020, pp. 1–15.
[8] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[9] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, "ExplAIner: A visual analytics framework for interactive and explainable machine learning," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 1064–1074, Jan. 2020.
[10] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, Dec. 2021.
[11] S. Palacio, A. Lucieri, M. Munir, J. Hees, S. Ahmed, and A. Dengel, "XAI handbook: Towards a unified framework for explainable AI," 2021, *arXiv:2105.06677*.
[12] W. Jin, J. Fan, D. Gromala, P. Pasquier, and G. Hamarneh, "EUCA: A practical prototyping framework towards end-user-centered explainable artificial intelligence," 2021, *arXiv:2102.02437*.
[13] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.
[14] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, T. A. Min, and E. Weippl, Eds. Cham, Switzerland: Springer, 2020, pp. 1–16.
[15] L. Ai, S. H. Muggleton, C. Hocquette, M. Gromowski, and U. Schmid, "Beneficial and harmful explanatory machine learning," *Mach. Learn.*, vol. 110, no. 4, pp. 695–721, Apr. 2021.
[16] B. J. Murray, M. A. Islam, A. J. Pinar, D. T. Anderson, G. J. Scott, T. C. Havens, and J. M. Keller, "Explainable AI for the Choquet integral," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 4, pp. 520–529, Aug. 2020.
[17] M. Katell, M. Young, D. Dailey, B. Herman, V. Guetler, A. Tam, C. Bintz, D. Raz, and P. M. Krafft, "Toward situated interventions for algorithmic equity: Lessons from the field," in *Proc. Conf. Fairness, Accountability, Transparency*, Barcelona, Spain, Jan. 2020, pp. 45–55.
[18] J. Photopoulos, "Fighting algorithmic bias," *Phys. World*, vol. 34, no. 5, pp. 42–47, Jul. 2021.
[19] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a 'Right to Explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.
[20] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021.
[21] C. Andor, A. Joó, and L. Mérö, "Galois-lattices: A possible representation of knowledge structures," *Eval. Educ.*, vol. 9, no. 2, pp. 207–215, 1985.
[22] H. F. Davis and A. D. Snider, *Introduction to Vector Analysis*. Charlottesville, VA, USA: Wm. C. Brown, 1995.
[23] H.-W. Liu, C.-F. Lin, and Y.-J. Chen, "Beyond *State v Loomis*: Artificial intelligence, government algorithmization and accountability," *Int. J. Law Inf. Technol.*, vol. 27, no. 2, pp. 122–141, Jun. 2019.
[24] H. Hagras, "Toward human-understandable, explainable AI," *Computer*, vol. 51, no. 9, pp. 28–36, Sep. 2018.
[25] D.A.R.P. Agency. (2016). *Broad Agency Announcement: Explainable Artificial Intelligence (XAI)*. Accessed: Dec. 21, 2021. [Online]. Available: https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf
[26] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds Mach.*, vol. 30, no. 1, pp. 99–120, Mar. 2020.
[27] M. Hickok, "Lessons learned from AI ethics principles for future actions," *AI Ethics*, vol. 1, no. 1, pp. 41–47, Feb. 2021.
[28] (2021). *Algorithmic Accountability for the Public Sector Learning From the First Wave of Policy Implementation*. Ada Lovelace Institute. AI Now Institute. and Open Government Partnership. Accessed: Dec. 21, 2021. [Online]. Available: https://www.adalovelaceinstitute.org/report/algorithmic-accountability-public-sector/
[29] S. H. Muggleton, U. Schmid, C. Zeller, A. Tamaddoni-Nezhad, and T. Besold, "Ultra-strong machine learning: Comprehensibility of programs learned with ILP," *Mach. Learn.*, vol. 107, no. 7, pp. 1119–1140, 2018.
[30] D. Michie, "Learning concepts from data," *Exp. Syst. Appl.*, vol. 15, nos. 3–4, pp. 193–204. 1998.
[31] K. M. Ford, P. J. Hayes, C. Glymour, and J. Allen, "Cognitive orthoses: Toward human-centered AI," *AI Mag.*, vol. 36, no. 4, pp. 5–8, Dec. 2015.
[32] N. Bostrom, "Ethical issues in advanced artificial intelligence," in *Machine Ethics and Robot Ethics*, W. Wallach, and P. Asaro, Eds. New York, NY, USA: Routledge, 2020, pp. 69–75.
[33] S. K. Kwan and J. Spohrer, "Reducing industry complexity with international standards: Current efforts for services, E-commerce, artificial intelligence," in *Advances in the Human Side of Service Engineering*, vol. 266, C. Leitner, W. Ganz, D. Satterfield, and C. Bassano, Eds. Cham, Switzerland: Springer, 2021, pp. 67–76.
[34] J. A. Kroll and E. W. Felten, "Accountable algorithms," Ph.D. thesis, Dept. Comput. Sci., Princeton Univ., Princeton, NJ, USA, 2015.
[35] M. Bovens, "Analysing and assessing accountability: A conceptual framework," *Eur. Law J.*, vol. 13, no. 4, pp. 447–468, 2007.
[36] M. MacCarthy, "An examination of the algorithmic accountability act of 2019," Inst. Inf. Law, Amsterdam, The Netherlands, Oct. 2019, doi: 10.2139/ssrn.3615731.
[37] R. Audi, *The Cambridge Dictionary of Philosophy*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 1995, doi: 10.1017/CBO9781139057509.
[38] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 9, pp. 1520–1536, Dec. 2011.

[39] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, A. Weller, and A. Wood, "Accountability of AI under the law: The role of explanation," 2017, *arXiv:1711.01134*.
[40] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–18.
[41] V. Dignum, "Responsible autonomy," 2017, *arXiv:1706.02513*.
[42] E. T. Tai, "Liability for (Semi) autonomous systems: Robots and algorithms," in *Research Handbook in Data Science and Law*, V. Mak, E. Tjong, T. Tai, and A. Berlee, Eds. Edward Elgar, 2018, pp. 55–82.
[43] B. Casey, A. Farhangi, and R. Vogl, "Rethinking explainable machines: The GDPR's 'right to explanation' debate and the rise of algorithmic audits in enterprise," *Berkeley Tech. Law J.*, vol. 34, 2019. Accessed: Feb. 19, 2018. [Online]. Available: https://ssrn.com/abstract=3143325
[44] J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan, "Transparency in algorithmic and human decision-making: Is there a double standard?" *Philosophy Technol.*, vol. 32, no. 4, pp. 661–683, 2019.
[45] H. Felzmann, E. Fosch-Villaronga, A. Tamò-Larrieux, and C. Lutz, "Towards transparency by design for artificial intelligence," *Sci. Eng. Ethics*, vol. 26, pp. 3333–3361, Nov. 2020.
[46] R. J. Hyndman, "Measuring forecast accuracy," in *Business Forecasting, Practical Problems and Solutions*. Hoboken, NJ, USA: Wiley, 2014, pp. 177–183.
[47] P.-H. C. Chen, Y. Liu, and L. Peng, "How to develop machine learning models for healthcare," *Nature Mater.*, vol. 18, no. 5, pp. 410–414, 2019.
[48] D. Colquhoun, "The false positive risk: A proposal concerning what to do about *p*-values," *Amer. Statistician*, vol. 73, no. 1, pp. 192–201, 2019.
[49] X. Naidenova, *Machine Learning Methods for Commonsense Reasoning Processes: Interactive Models*. Hershey, NY, USA: IGI Global, 2009.
[50] W. Wallach, C. Allen, and I. Smit, "Machine morality: Bottomup and top-down approaches for modelling human moral faculties," *AI Soc.*, vol. 22, no. 4, pp. 565–582, 2008.
[51] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, and G.-B. Huang, "EmoSenticSpace: A novel framework for affective commonsense reasoning," *Knowl.-Based Syst.*, vol. 69, pp. 108–123, Oct. 2014.
[52] E. T. Mueller, *Commonsense reasoning: An event calculus based approach*, S. Elliot, Ed. Waltham, MA, USA: Elsevier Science, 2014.
[53] J. Vice, M. Mehmood Khan, and S. Yanushkevich, "Multimodal models for contextual affect assessment in real-time," in *Proc. IEEE 1st Int. Conf. Cognit. Mach. Intell. (CogMI)*, Los Angeles, CA, USA, Dec. 2019, pp. 87–92.
[54] M. Mehmood Khan, R. D. Ward, and M. Ingleby, "Toward use of facial thermal features in dynamic assessment of affect and arousal level," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 412–425, Sep. 2017.
[55] J. Vice and M. M. Khan, "Toward accountable and explainable artificial intelligence part two: The framework implementation," *IEEE Access*, vol. 10, pp. 36091–36105, 2022, doi: 10.1109/ACCESS.2022.3163523.
[56] R. S. Livingstone and A. F. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, pp. 1–35, May 2018.
[57] A. Samara, L. Galway, R. Bond, and H. Wang, "Sensing affective states using facial expression analysis," in *Proc. 10th Int. Conf. Ubiquitous Comput. Ambient Intell.*, Gran Canaria, Spain, 2016, pp. 341–352.
[58] S. L. Happy and A. Routray, "Robust facial expression classification using shape and appearance features," in *Proc. 8th Int. Conf. Adv. Pattern Recognit. (ICAPR)*, Kolkata, India, Jan. 2015, pp. 1–5.
[59] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.
[60] K. Dupuis and M. K. Pichora-Fuller, "Toronto emotional speech set," Dept. Psychol., Univ. Toronto. Toronto, ON, Canada, Tech. Rep., Jun. 2010. [Online]. Available: https://tspace.library.utoronto.ca/handle/1807/24487
[61] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren, "Value sensitive design and information systems," in *Early Engagement and New Technologies: Opening Up the Laboratory*, N. Doorn, D. Schuurbiers, I. Van De Poel, and M. E. Gorman, Eds. Dordrecht, The Netherlands: Springer, 2013, pp. 55–95.
[62] D. R. Desia and J. A. Kroll, "Trust but verify: A guide to algorithms and the law," *Harvard J. Law Technol.*, vol. 31, no. 1, pp. 1–64, 2017.
[63] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," *J. Global Health*, vol. 8, no. 2, pp. 1–8, Dec. 2018.
[64] R. Hoffman, T. Miller, S. T. Mueller, G. Klein, and W. J. Clancey, "Explaining explanation, part 4: A deep dive on deep nets," *IEEE Intell. Syst.*, vol. 33, no. 3, pp. 87–95, May 2018.
[65] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa, "Inspectability and control in social recommenders," in *Proc. 6th ACM Conf. Recommender Syst. (RecSys)*, 2012, pp. 43–50.

MASOOD M. KHAN (Member, IEEE) received the B.E. degree in mechanical from the NED University of Engineering and Technology, the M.S.M.E. degree from Colorado State University, and the Ph.D. degree from the University of Huddersfield. He was at the National University of Computer and Emerging Sciences, the Jefri Bolkiah College of Engineering, and the American University of Sharjah. He is with the Faculty of Science and Engineering, Curtin University, Western Australia. He has published more than 55 peer-reviewed articles in his research areas. His research interests include machine learning, affective computing, computer vision and perception, human–computer interaction, and artificial intelligence. He is a fellow of the Higher Education Academy.

JORDAN VICE received the B.Eng. degree (Hons.) in mechatronic engineering from Curtin University, where he is currently pursuing the Ph.D. degree in mechatronic engineering. His research interests include artificial intelligence, explainable artificial intelligence, machine learning, real-time assessment of affective states, and multimodal affective state assessment. He received the 2019 Proxima Consulting Prize for Most Outstanding Final Year Project in mechatronic engineering.