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

Masood Khan ¹ and Jordan Vice ²

¹Curtin University
²Affiliation not available

October 30, 2023

Abstract

After reviewing the current state of explainable Artificial Intelligence (XAI) capabilities in artificial Intelligence (AI) systems developed for critical domains like criminology, engineering, governance, health, law and psychology, this paper proposes a domain-independent Accountable explainable Artificial Intelligence (AXAI) capability framework. The proposed AXAI framework extends the XAI capability to let AI systems share their decisions and adequately explain the underlying reasoning processes. The idea is to help AI system developers overcome algorithmic biases and system limitations through incorporation of domain independent AXAI capabilities. Moreover, existing XAI methods would neither separate nor quantify measures of comprehensibility, accuracy and accountability so incorporating and assessing XAI capabilities remains difficult. Assessment of the AXAI capabilities of two AI systems in this paper demonstrates that the proposed AXAI framework facilitates separation and measurement of comprehensibility, predictive accuracy and accountability. The AXAI framework allows for the delineation of AI systems in a three-dimensional AXAI space. It measures comprehensibility as the readiness of a human to apply the acquired knowledge. The system accuracy is measured as the ratio of the test and training data, training data size and the observed number of false-positive inferences. Finally, the AXAI framework measures accountability in terms of the inspectability of the input cues, the processed data and the output information, for addressing any legal and ethical issues.
Toward Accountable and Explainable Artificial Intelligence Part one: Theory and Examples

Masood Khan  
Jordan Vice  
Faculty of Science and Engineering, \textit{Curtin University, Perth, Australia}

Abstract

This paper proposes a domain-independent Accountable eXplainable Artificial Intelligence (AXAI) capability framework as an extension to the prevailing eXplainable Artificial Intelligence (XAI) capability frameworks. As artificial Intelligence (AI) systems are generally unable to explain inferences and decisions and are marred with built-in biases and algorithmic limitations, system designers and users seek incorporation of XAI in AI systems. Non-availability of domain-independent XAI capability frameworks encourages system developers to propose domain-specific methods of incorporating XAI in AI systems. Moreover, existing XAI methods neither separate nor quantify measures of comprehensibility, accuracy and accountability. This makes incorporating and assessing XAI capability complex and tedious. The proposed AXAI capability framework responds to these issues, facilitating separation and measurement of predictive accuracy, comprehensibility and accountability, allowing for the delineation of AI systems in a three-dimensional AXAI space. Our proposed AXAI capability framework quantifies comprehensibility in terms of the readiness of a human to apply the acquired knowledge. It assesses system accuracy in terms of the ratio of test and training data, training data size and the observed number of false-positive inferences. The AXAI capability framework measures accountability in terms of the inspectability of input cues, processed data and output information for establishing a chain of responsibility and addressing important legal and ethical issues.

1. Background

Issues pertaining to built-in algorithmic biases in Artificial Intelligence (AI) systems were first realized in the late 1970s (Liu, Lin & Chen, 2019). Recent works suggest predictive accuracy alone would not suffice when addressing biases in AI systems. Factors like the historical background, political constraints, and institutional context of AI systems also need to be considered while addressing AI system biases (Katell et al., 2020). In the following decades, such realizations were extended to all genres of AI systems i.e.: supervised learning-supported classifiers and regression systems, unsupervised learning-supported clustering and labelling systems and, reinforced learning systems including deep neural networks. With time, the importance of explanations for inferences, accountability for impact of outcomes and the determination of chain of responsibility in the context of AI systems has increased (DARPA, 2016; Doshi-Velez et al., 2017). Recently, government and business entities have also started to emphasize the need to account for the ethical implications of using AI systems (Hagendorff, 2020; Hickok, 2021). 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 (Open Government Partnership, 2021). Hence, eXplainable Artificial Intelligence (XAI) has emerged as a topic of interest
for computer scientists, AI theorists and practitioners across various domains (Adadi & Berrada, 2018; Photopoulos, 2021).

AI systems are generally delineated in a two-dimensional space of orthogonal axes viz., predictive accuracy and comprehensibility (Michie, 1998). However, expert systems and Machine Learning (ML) systems are traditionally assessed on the basis of their predictive accuracy alone (Muggleton et al., 2018). Since AI systems are becoming relevant in solving both routine and complex problems (Ford et al., 2015), in some domains, AI systems can outperform humans and are now becoming inevitable assets (Hagras, 2018). Nonetheless, AI systems need to be accurate, explainable and liable for their impact on users for them to be socially and technically acceptable (Hagras, 2018). Such capabilities are necessary for delegating critical tasks like disease diagnosis, psychological and psychiatric assessments, loan approvals, autonomous driving and threat analysis to AI systems. Since critical roles and important decisions can cause negative consequences (Bostrom, 2020), theorists and practitioners acknowledge that ‘accountability’ needs to be added to AI systems’ ability to ‘explain’ reasoning (Kim & Doshi-Velez, 2021).

As reported by Muggleton et al. (2018), accuracy of predictions and comprehensibility of knowledge provided a basis for proposing a set of criteria for delineating ML systems as per Michie (1988). 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 (Michie, 1988). Building upon these works, the description and scope of AI system comprehensibility was further refined by Schmid et al. (2017). Recent efforts directed toward building responsible AI systems are also evident in IEEE standard P2840 and similar initiatives (Kwan & Spohrer, 2021).

Accountability, in the context of AI systems, connotes compliance with ethical, procedural and legal norms while processing information, invoking rules and making decisions (Kroll & Felten, 2015). 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 when the forum poses questions and pass judgement. Also, the actor may face consequences (Bovens, 2007).

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 (Kim & Doshi-Velez, 2021). 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 the assessment risks of “inaccurate, unfair, biased, or discriminatory decisions.” The act further asked businesses to address the outcomes of AI systems’ assessments (MacCarthy, 2020). Several palpable elements of AI system accountability could be extracted from the bill.

Explainability of AI systems has been treated as a core and domain-independent software engineering issue but accountability has been 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. Since AI systems now perform highly critical tasks, they are also considered liable to adjudication, legislation and litigation (Kim & Doshi-Velez, 2021). 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 (Liu, Lin & Chen, 2019).
Incorporating acceptable, trustworthy and explainable Artificial Intelligence in intelligent systems poses many challenges, mainly for application-related sensitivities and domain-specific requirements of various professions. It was realized that practitioners’ input had been minimal in an almost four-decade long journey from relying on the one-dimensional predictive accuracy models of the past, to the integration of explainability and accountability in AI systems now (Kim & Doshi-Velez, 2021; Michie, 1988; Schmid et al., 2017; Kroll & Felten, 2015).

As logic programming, automated reasoning and software engineering experts lead AI system design efforts (Kim & Doshi-Velez, 2021; Michie, 1988; Schmid et al., 2017), practitioners find the prevailing XAI frameworks algorithm-centric, neglecting domain-specific needs and, missing practical explanations (Miller, 2019). Contemporary literature (Vu et al., 2018; Pawar et al., 2020; Liao, Gruen & Miller, 2020) highlights several gaps in a computing experts’ view of XAI and a practitioners’ requirements. From practitioners’ perspectives, these gaps result in (a) no or little utility of the available explainability and (b) an inability to interpret the given reasoning. In medical, neurosciences and related domains, such gaps inhibit automation of tedious practices and impede the adoption of AI systems (Adadi & Berrada, 2018; Spinner et al., 2020; Mohseni, Zarei & Ragan, 2021; Palacio et al., 2021; Jin et al., 2021). Statistical and probabilistic explanations provided by generic XAI frameworks are considered to be limited and less effective (Vu et al., 2018; Pawar et al., 2020). It is also acknowledged that the prevailing XAI frameworks do not fully comply with the norms of regulatory bodies and industry (Gunning & Aha, 2019; Longo et al., 2020). Furthermore, till now we have no proven method of measuring the non-explainability of an AI system (Ai et al., 2021).

2. Introduction

As addressing practitioners’ requirements and developing holistic XAI frameworks is expected to boost user confidence in AI systems, attempts are being made to develop more holistic and domain-independent accountability and explainability frameworks for AI systems (Adadi & Berrada, 2018; Spinner et al., 2020; Mohseni, Zarei & Ragan, 2021; Palacio et al., 2021; Jin et al., 2021). Several national and international organizations are engaged in efforts to develop and recommend better norms for incorporating explainability and accountability in AI systems (Katell et al., 2020; Pothopoulos, 2021; Goodman & Flaxman, 2017). Meanwhile, several recent works propose methods of measuring and assessing explainability of machine learning and AI systems (Burkart & Huber, 2021). Nonetheless, recent literature cites AI systems as difficult to understand, adopt and trust (Murray et al., 2020).

As stated by Kim & Doshi-Velez (2021), perceiving XAI in a two-dimensional space of predictive accuracy and comprehensibility results in mixing explainability-related factors with factors of accountability. Such a convoluted representation does not help practitioners, cannot fulfill regulators’ expectations and, offers limited transparency, making it difficult to establish a chain of responsibility (Adadi & Berrada, 2018; Spinner et al., 2020; Mohseni, Zarei & Ragan, 2021; Palacio et al., 2021; Jin et al., 2021). In order to better represent, describe and report the explainability of an AI system, we formulated a three-level (narrow and shallow) Galois-lattices structure (Andor, Joó & Méró, 1985), shown in Figure 1.

The Galois-lattices structure contains the nine most important elements of AI system explainability. For separating the convoluted factors of XAI, we propose constructing a three-dimensional (3D) space using the level-1 elements of the Galois-lattices structure such that AI systems’ account-
able explainability would comprise of three mutually perpendicular axes: accuracy, comprehensibility and accountability with same unit of length (Davis & Snider, 1995). Each of the three axes of the accountable explainable artificial intelligence (AXAI) space is an independent vector resulting in a Cartesian coordinate system. Hence, each vector would be of form: \[ A = \sqrt{a_1 i + a_2 j + a_3 k} \] where \( i, j \) and \( k \) are unit vectors. In this 3D AXAI space, quantitatively separable and globally applicable representations of vectors would allow for the deconvolution of predictive accuracy, comprehensibility and accountability.

This novel approach builds upon prevailing XAI paradigms to propose a new AXAI framework for analysing and assessing AI systems. The new AXAI capability framework is being proposed to:

1. Provide a globally applicable, domain-independent and easy to incorporate AXAI capability framework;
2. Incorporate decision accuracy measurements in the AXAI capability framework;
3. Incorporate and quantify the level of comprehensibility of an AI system;
4. Incorporate and quantify the level of accountability of an AI system and;
5. Allow practitioners to visually and quantitatively examine various pieces of information and assess their relevance.

We also demonstrate the applicability of the AXAI capability framework by comparing and delineating two affective state classification systems. As shown in Figure 2, the AXAI capability framework would allow for the incorporation of theoretical guarantees, empirical evidences and statistical assurances in AI systems. Each axis of the 3D space comprises of three components and each component can be assigned a numerical value depending on a system’s attributes, making it easy to quantitatively delineate AI systems.

Figure 1: Three-level Galois-lattices structure used to represent the explainability of an AI system.
This paper is organized in seven sections. After providing the background information in Section 1 and introducing this work in Section 2, Section 3 establishes theoretical foundations of the proposed AXAI framework. Section 4 demonstrates application of the proposed framework in designing and assessing AXAI capabilities of AI systems. The assessment results are presented in Section 5. The proposed framework and its applications are analysed and discussed in Section 6. Finally, Section 7 identifies the possible directions of future work and concludes this work.

3. Theoretical foundations of the accountable explainability (AXAI) capability framework

Although several recent works provide foundations for incorporating measurable parameters of accountability and explainability in AI systems (Katell et al., 2020; Open Government Partnership, 2021), little work has been done for developing a holistic AXAI framework and providing a set of quantifiable features for assessing the AXAI capability of systems. In order to develop an AXAI framework, personal, social, moral and legal factors used to hold an individual accountable and liable to explain personal actions and decisions must be understood. Gradually, it is becoming important to include these factors in the design and assessment of AI systems (Dignum, 2017). Sig-
nificant moral and legal factors that make a decision maker (an AI system in this case) liable to explain decisions are (Doshi-Velez et al., 2017):

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.

Accountability of an AI system would depend on the context of the confronted issue (Goodman & Flaxman, 2017). 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 (Dignum, 2017). Hence, an AI system should be aware of contextual ethical values and should have some built-in capacity to understand the moral consequences of its actions and decisions (Tai, 2018). Accountability should therefore be derived from both information/data and the algorithmic approach (MacCarthy, 2020). The employed algorithmic approach and data must be sensitive to the context while making inferences and decisions (Liu, Lin & Chen, 2019; Doshi-Velez et al., 2017; Tai, 2018). Hence, for an AI system to be accountable, social and moral values and legal norms must be represented in the underlying models. It is also argued that an AI system should be operated in such a manner that the chain of responsibility is clear and identifiable (Doshi-Velez et al., 2017; Dignum, 2017).

3.1 The three-dimensional space of accountable explainability (AXAI)

As stated by Gunning & Aha (2019), the Palo Alto Research Center (PARC) proposed the incorporation of an interactive system for explaining the learned capabilities of an XAI system to control a simulated unmanned aerial system. The PARC posited that AI system explanations should reveal all information used in decision making by showing that the AI 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 (Gunning & Aha, 2019). We attempt to build upon such previous works (Muggleton et al., 2018; Kim & Doshi-Velez, 2021; Schmid et al., 2017; Gunning & Aha, 2019; Longo et al., 2020; Dignum, 2017) and propose an AXAI capability framework for effectively incorporating accuracy, comprehensibility and accountability in AI systems. The AXAI framework being proposed in this paper allows for the quantitative assessment of system capabilities and the delineation of AI systems within a 3D space. The following subsections discuss assumptions, definitions and hypotheses leading to the design of the proposed AXAI capability framework. These assumptions and definitions were inspired by and adopted from the relevant literature (Michie, 1998; Muggleton et al., 2018; Schmid et al., 2017; Mohseni, Zarei & Ragan, 2021; Naidenova, 2009; Poria et al., 2014; Mueller, 2014).

We assume that an AI 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 (Open Government Partnership, 2021). This definition of $\mathcal{P}$ also considers
an AI system as a holistic system comprising of one or many AI systems or algorithms, capable of producing the desired outputs that enable inferences and decision making (Open Government Partnership, 2021). The term ‘AI system’ in this paper encompasses: supervised learning-supported classifiers and regression systems, unsupervised learning-supported clustering and labelling systems and, reinforced learning systems including deep neural networks. Therefore, an AI system would include definite symbols, definite functions, definite propositions, definite predicates, logical symbols, object variables and propositional variables (Naidenova, 2009; Mueller, 2014).

In the following sections, $C$ denotes a constant, $P$ represents a predicate symbol and $S$ shows a human population having each individual human represented as ‘$s$’. In this paper, $V$ shows a first-order variable and $B$ is the background knowledge. A human possessing the background knowledge $B$ is also considered as a definite program $P$. Here, $D_n$ denotes a definition $D$ having a number $n$. $D$ in this paper denotes a domain. Having these notations and based on the previous works (Liao, Gruen & Miller, 2020; Dignum, 2017; Tai, 2018; Naidenova, 2009), the following subsections describe all measurable parameters belonging to each of the three vectors forming the 3D AXAI measurement space shown in Figure 2.

### 3.2 Definitions

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

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

**D3**: The Accountable Explainable Artificial Intelligence (AXAI) capability denoted by $X_{\text{AAI}}$ is a representation in a three-dimensional space. We posit that $X_{\text{AAI}}$ comprises of three independent vectors: $C$ (comprehensibility), $P_A$ (predictive accuracy) and $S_A$ (system accountability). Also, each one of the three vectors $C$, $P_A$ and $S_A$ comprises of three independent components whose details are given in the following definitions $D4 – D6$.

**D4**: The comprehensibility $C$ in the context of a human population $S$ is represented as $C(S, \mathcal{P})$ where $C$ 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:

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

(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 (R) of a human $s (s \in S)$ for applying the knowledge given in program $\mathcal{P}$ to assign a public name to a new definition $q$ with respect to the domain $D$ 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 integrated with supervised learning in AI systems such that each cluster can be given an intended definition (Naidenova, 2009).

**D4A**: The inspection time ($T_{it}$) is the mean time that a human $s (s \in S)$ 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 $D$. 
**D4B**: The predicate recognition time ($T_{pr}$) is the mean time that a human $s$ ($s \in S$) requires for assigning a correct public name to a predicate symbol $p$ within the domain $D$.

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

**D5**: The predictive accuracy $P_A$ of an AI system $P$ with respect to a human population $S$ and a domain $D$ is represented as $P_A(S, P)$ 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:

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

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

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

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

**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 an AI system $P$ is expressed as $O_{fp}$. In the context of AXAI, $O_{fp}$ is an indicator of the ability of the system $P$ 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 $P$ for accurately naming a predicate symbol $p$ represented as a named inference $q$ with respect to a domain $D$.

**D6**: The system accountability $S_A$ of an AI system $P$ with respect to a human population $S$ is represented as $S_A(S, P)$ 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:

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

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

**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 an AI program $P$ is an indicator of the mean clarity observed by a human $s$ ($s \in S$) with which $s$ would inspect the definition $q$ before $q$ is named as a predicate symbol $p$ with respect to the domain $D$. 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 an AI system $P$ is an indicator of the mean clarity of the processed (or conditioned) definition $q$ as observed by a human from a population $S(s \in S)$. 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 \( \mathcal{D} \). Therefore, \( I_{\text{pro}} \) 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 \( p \).

**D6C:** The mean inspectability score of output signals \( I_{\text{out}} \) provided by \( \mathcal{P} \) is an indicator of the mean clarity of the definition \( q \) as observed by a human \( s \ (s \in \mathcal{S}) \) with which \( s \) would inspect the output definition of \( q \) for naming it as a predicate symbol \( p \) with respect to a domain \( \mathcal{D} \). 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 \( \mathcal{P} \).

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

### 3.3 Hypotheses

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

**Hypothesis 2:** The larger the norm of the comprehensibility vector \( ||\mathcal{C}|| \), the more comprehensible an AI program \( \mathcal{P} \) is.

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

**Hypothesis 4:** The larger the norm of the predictive accuracy vector \( ||P_{A}|| \) of a definite AI program \( \mathcal{P} \) the better predictive accuracy \( \mathcal{P} \) offers.

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

**Hypothesis 6:** The larger the norm of the system accountability vector \( ||S_{A}|| \) of a definite AI program \( \mathcal{P} \) the better the level of incorporated accountability in the program \( \mathcal{P} \).

**Hypothesis 7:** The inverse of the mean time \( \frac{1}{t_{u}} \) that a human \( s \ (s \in \mathcal{S}) \) requires for inspecting the information presented by the AI program \( \mathcal{P} \) before using the knowledge provided by that \( \mathcal{P} \) for solving a new problem in domain \( \mathcal{D} \), is directly proportional to the presentation quality (\( Q_{p} \)) of \( \mathcal{P} \) given as \( \frac{1}{t_{u}} \propto Q_{p} \).

**Hypothesis 8:** The inverse of the mean predicate recognition time \( \frac{1}{t_{p}} \) that a human \( s \ (s \in \mathcal{S}) \) requires to assign a correct public name to a predicate symbol \( p \) in an AI system is proportional to the ability (\( \mathcal{A}_{p} \)) of recognizing and accurately assigning a public name to a predicate symbol \( p \). Hence, \( \frac{1}{t_{p}} \propto \mathcal{A}_{p} \).

**Hypothesis 9:** The ratio of the size of test data and the size of the training data \( (r_{\text{tst-\text{trn}}}) \) of an AI program \( \mathcal{P} \) is directly proportional to the level of rigour (\( L_{\text{Rig}} \)) applied in training and testing \( \mathcal{P} \) with respect to a domain \( \mathcal{D} \), hence, \( r_{\text{tst-\text{trn}}} \propto L_{\text{Rig}} \).

**Hypothesis 10:** The mean score of inspectability of data after being processed (\( I_{\text{pro}} \)) shows how understandable the intermediary data representation/models (\( F_{\text{mod}} \)) in a definite AI program \( \mathcal{P} \) are. Thus, \( I_{\text{pro}} \propto F_{\text{mod}} \).
**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 \( P \) with respect to a domain \( D \), hence, \( O_{fp} \propto E_{bp} \).

Table 1 presents a complete list of measurable parameters used to determine the overall AXAI capability of a definite program \( P \).

<table>
<thead>
<tr>
<th>S No.</th>
<th>Parameter</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>The norm of the vector of comprehensibility (</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>The norm of the vector of predictive accuracy (</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>The norm of the vector of system accountability (</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>The inspection time ( (T_{it}) )</td>
<td>Score (integer)</td>
</tr>
<tr>
<td>5.</td>
<td>The predicate recognition time ( (T_{pr}) )</td>
<td>Score (integer)</td>
</tr>
<tr>
<td>6.</td>
<td>The predicate naming time ( (T_{pn}) )</td>
<td>Score (integer)</td>
</tr>
<tr>
<td>7.</td>
<td>The ratio of the test data and the training data ( (T_{test-trn}) )</td>
<td>Score (integer)</td>
</tr>
<tr>
<td>8.</td>
<td>The absolute size of the training data ( (d_{trn}) )</td>
<td>Score (integer)</td>
</tr>
<tr>
<td>9.</td>
<td>The number occurrences of the false-positive naming ( (O_{fp}) )</td>
<td>Score (integer)</td>
</tr>
<tr>
<td>10.</td>
<td>The mean score of inspect-ability of input signals ( (I_{in}) )</td>
<td>Score (integer)</td>
</tr>
<tr>
<td>11.</td>
<td>The mean score of inspect-ability of the processed data ( (I_{pro}) )</td>
<td>Score (integer)</td>
</tr>
<tr>
<td>12.</td>
<td>The mean score of inspect-ability output cues ( (I_{out}) )</td>
<td>Score (integer)</td>
</tr>
</tbody>
</table>

4. Examples of assessing the AXAI capabilities of AI systems

To test the relevance of the AXAI framework proposed in this work, the \( X_{AAI} \) scores of two AI systems were calculated. The following subsections present details of the two AI system in which \( X_{AAI} \) scores were estimated using the AXAI framework.

4.1 Assessment of the affective state assessment module

The first AI system assessed for this work was designed to have the proposed AXAI framework built into it. It is a multimodal definite AI system (Vice, Khan & Yanushkevich, 2019) implemented to be a portable affective state assessment module (ASAM). The current version of the ASAM is an enhanced and improved form of our previously published system (Vice, Khan & Yanushkevich, 2019). 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” being submitted for publication in this journal. The provisions of explainability and accountability in the ASAM were ensured by adding the AXAI features listed in Table 1. Figure 3 shows the overall, high-level architecture of the ASAM.

The ASAM was designed as an independent and portable module. It can be added to any existing robotic system and enables real time affective state assessment, useful in systems like authentication machines or social robots. Affective states in the ASAM are classified through analyses of facial expressions and speech cues with the latter being split into paralinguistic and linguistic.
Toward AXAI Part 1: Theory and Examples

Figure 3: ASAM high-level system architecture, highlighting how signals \{X_1, X_2, X_3\} traverse throughout the system through processing and output stages and how/what information is displayed to the user via the graphical user interface (GUI) shown in Figure 4.

Through a combination of machine and deep learning techniques, the ASAM can perform robust affective state assessment. Its inputs, outputs, and data being processed remain observable to users. Various pieces of information at the input, processing and output stages are presented in graphical, tabular and other visual forms to assist humans in inspecting, understanding and interpreting the decision-making process and inferences.

As the ASAM was designed to showcase AXAI capabilities in an affective state assessment system, it provides transparency via input feedback, graphical and tabular information, while providing users with scrutiny and debugging opportunities. ASAM explanations are provided through displays 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 interface. Figure 4 shows the input, processing and output information presented to users by the ASAM, highlighting the system’s graphical user interface (GUI).

4.2 Assessment of the ASAM’s AXAI capabilities

Ten qualified industry professionals and postgraduate students well-versed in machine learning and AI systems volunteered to assess the AXAI capabilities of the ASAM using parameters outlined in Table 1. These assessors, educated in Engineering, Science and Psychology were thoroughly informed on the objectives and outcomes of the assessment during an introduction session. After the briefing, participants were given a system user manual. Assessors had the opportunity to use the ASAM before assessing its functionality. Users tested the ASAM for an average time of twenty minutes. While testing, assessors awarded scores for parameters 4-6 and 10-12 using the 0-5 point scale detailed in Table 2. Parameters 7-9 did not require user input as these metrics could 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 us to delineate the AXAI-capabilities of the ASAM in a 3D space as discussed in previous sections and visualised in Figure 2.
Figure 4: Overview of the ASAM GUI windows, highlighting the input, processing and output information presented to the user. The upper window is shown to users upon execution of the software with the image in the frame being an example from the RAVDESS dataset (Livingstone & Russo, 2018). This window presents all input information to the user for the three signals. The four coloured windows in the lower half of the figure are the ASAM monitoring windows, which are used to investigate processing and output information, where: CYAN = Facial Expression Analysis, ORANGE = Linguistic Analysis, PURPLE = Paralinguistic Analysis and LIME = Multimodal Analysis. All windows are executed on separate threads thus allowing for parallel viewing.
Table 2: Guidelines for scoring and determining the AXAI capabilities of the ASAM, a definite program that uses multimodal input cues to assess user affective states using facial expression, paralinguistic and linguistic cues.

<table>
<thead>
<tr>
<th>S No.</th>
<th>Parameter</th>
<th>Scoring Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.0 - 1.0</td>
</tr>
<tr>
<td>4</td>
<td>Inspection Time</td>
<td>Information appears to be very difficult and takes a long time to understand</td>
</tr>
<tr>
<td>5</td>
<td>Predicate Recognition Time</td>
<td>A human would take very long time to interpret the output</td>
</tr>
<tr>
<td>6</td>
<td>Predicate Naming Time</td>
<td>A human would take very long time to use the inferences for naming another predicate within the domain</td>
</tr>
<tr>
<td>7</td>
<td>Test/Training Data Ratio</td>
<td>$0 : 0.1 \leq r_{tst-trn} \leq 0.9$</td>
</tr>
<tr>
<td>8</td>
<td>Training Data Absolute Size</td>
<td>$0 : d_{trn} \leq 5n_{names}$</td>
</tr>
<tr>
<td>9</td>
<td>False-positive naming occurrences</td>
<td>$0 : O_{fp} \geq 50%$ of the time</td>
</tr>
<tr>
<td>10</td>
<td>Inspectability of input signals</td>
<td>Explanations are not clear to users</td>
</tr>
<tr>
<td>11</td>
<td>Inspectability of intermediate data stages</td>
<td>Intermediate data cannot be seen or cannot be interpreted</td>
</tr>
<tr>
<td>12</td>
<td>Inspectability of output signals</td>
<td>Output information is nondescript and hard to understand/interpret</td>
</tr>
</tbody>
</table>
4.3 A system for dynamic assessment of affective states and arousal levels

The second AI system tested for its AXAI capabilities was a definite program that was designed to work as a two-step dynamic assessment of affective states and arousal levels system (DAASAS). It uses thermal infrared images (TIRI’s) of facial expressions (Khan, Ward & Ingleby, 2017) and was not designed to have AXAI capabilities built into it. Hence, in this case, post-production assessment of AXAI capabilities was performed.

The DAASAS would first analyse TIRI’s for examining the hæmodynamic variations caused by changes in affective states. The algorithmic execution of DAASAS 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 DAASAS - white boxes show the visible and inspect-able information and blue boxes show the information hidden in the program.

4.4 Assessment of DAASAS AXAI capabilities

A postdoctoral fellow and seven postgraduate students who were trained in AI and machine learning volunteered to assess the AXAI capabilities of the DAASAS. As in the previous example, 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 DAASAS was recorded to be 27 minutes. As in the previous example, assessors awarded scores for parameters 4-6 and 10-12 on a 0-5 point scale as highlighted in Table 4.
Table 3: ASAM user experience scores on a 0-5 scale and their normalised scores. These scores were used to determine the ASAM’s system accountability ‘$S_A$’ and comprehensibility ‘$||C||$’ capabilities.

<table>
<thead>
<tr>
<th>User ID</th>
<th>Symbol</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Average using 5 Point Score</th>
<th>Normalised Average Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$</td>
<td></td>
<td>C</td>
<td></td>
<td>$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3.5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>$</td>
<td></td>
<td>S_A</td>
<td></td>
<td>$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3.5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>$T_{it}$</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3.2</td>
<td>0.5</td>
<td>0.87</td>
</tr>
<tr>
<td>5</td>
<td>$T_{pr}$</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3.2</td>
<td>0.2</td>
<td>0.06</td>
</tr>
<tr>
<td>6</td>
<td>$T_{pn}$</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3.5</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3.05</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>$I_{in}$</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2.5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>3.55</td>
<td>0.87</td>
</tr>
<tr>
<td>11</td>
<td>$I_{pro}$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3.8</td>
<td>0.65</td>
</tr>
<tr>
<td>12</td>
<td>$I_{out}$</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3.8</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 4: User experience scores for DAASAS on a 0-5 scale with normalised scores. These scores were used to determine the DAASAS system’s AXAI capabilities.

<table>
<thead>
<tr>
<th>User ID</th>
<th>Symbol</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Average using 5 Point Score</th>
<th>Normalised Average Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$</td>
<td></td>
<td>C</td>
<td></td>
<td>$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3.5</td>
</tr>
<tr>
<td>3</td>
<td>$</td>
<td></td>
<td>S_A</td>
<td></td>
<td>$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>$T_{it}$</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>5</td>
<td>$T_{pr}$</td>
<td>0.2</td>
<td>0.2</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>6</td>
<td>$T_{pn}$</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>1.5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1.37</td>
</tr>
<tr>
<td>10</td>
<td>$I_{in}$</td>
<td>1.5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1.93</td>
</tr>
<tr>
<td>11</td>
<td>$I_{pro}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>$I_{out}$</td>
<td>1.5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.43</td>
</tr>
</tbody>
</table>

5. Results

5.1 ASAM’s AXAI capability assessment results

The predictive accuracy components given in Table 2 were known to the system developers as they were designing the ASAM. 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 (Samara et al., 2016; Happy & Routray, 2015). 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 (Lucey et al., 2010) and 1400 speech samples from the Toronto Emotional Speech Set (TESS) (Dupuis & Pichora-Fuller, 2010) 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 100–1000 $n_{names}$ 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 $d_{trn} = 0.811$. 

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 (Livingstone & Russo, 2018). The data was 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:

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

$$= \sqrt{1^2 + 0.811^2 + 0.8559^2}$$

$$= \sqrt{1 + 0.657721 + 0.732565}$$

$$= 1.54606$$

The score parameters 4-6 and 10-12 in Table 1 were respectively used to determine the system accountability and comprehensibility vector norms. The means 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 $||C||$ 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 Figure 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 $C$, $P_A$ 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: $C = 1.203$, $P_A = 1.546$, 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 Figure 6.

Using these results, we can conclude that the ASAM is an AXAI-capable system with high levels of predictive accuracy and comprehensibility. However, the accountability score suggested that more attention to the ASAM’s accountability components needed to be considered. 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.

5.2 DAASAS’s AXAI capability assessment results

The AXAI capability assessment results for DAASAS 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 DAASAS was 1:1 (Khan, Ward & Ingleby, 2017) resulting in a score of 5.0 (normalized value of 1). The training data size $d_{trn}$ parameter score for the DAASAS 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 DAASAS was:

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

The data in Table 4 suggests that DAASAS 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 DAASAS scores, we report comprehensibility, predictive accuracy and system accountability values of: $C = 0.333$, $P_A = 1.342$, and $S_A = 0.489$. Deriving these scores allows us to plot the DAASAS within the three axes and compare its AXAI capabilities to other AI systems like the ASAM as visualised in Figure 6.

6. Discussion

The two definite AI programs assessed in the preceding sections were fundamentally different. The first system (ASAM) was designed to have the AXAI capability incorporated in it. The second program (DAASAS) was a basic classification and clustering system that was not designed to have the AXAI capability incorporated. Despite this fundamental difference, the proposed AXAI capability framework allowed for assessing the two AI programs in terms of predictive accuracy, comprehensibility and accountability. This example of delineating AI systems demonstrates that the AXAI
capability framework is effective in AI systems’ design and assessment. Furthermore, the AXAI capability framework also provided an opportunity to systematically address ethical and professional issues, such as those highlighted by Dignum (2017) and Friedman et al. (2013), while building an AI system. As evident in the above comparison, the nine measurable components of the AXAI capability framework ensure attention to system details, ethical responsibilities and moral duties during the conceptual design and functional analysis stages. Such manifestations have been desired in AI systems for quite some time (Dignum, 2017; Wallach, Allen & Smit, 2008).

Building upon the XAI capability centred philosophical discussions in the literature (Liu, Lin & Chen, 2019; Tai, 2018; Desai & Kroll, 2017), our proposed AXAI capability framework provides three sets of quantitative parameters 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 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, the proposed AXAI capability framework would potentially free system developers from getting burdened by domain-specific methods and details (Pawar et al., 2020; Spinner et al., 2020; Jin et al., 2021; Panch, Szolovits & Atun, 2018).
It would be safe to suggest that the proposed AXAI capability framework, to a large extent, meets DARPA’s perceived goal of developing human-centred AI systems (Hoffman et al., 2018) as it provides such learning models and decision making processes that would be shared, understood and trusted by the relevant communities (Gunning & Aha, 2019).

Through examples of AXAI capability assessment, 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, though the overall framework can be applied in a domain-independent manner.

Finally, system accountability in the AXAI framework is reflected in the level of accuracy with which a human would realize occurrences of logical elements in an 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 are 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 (Kim & Doshi-Velez, 2021; Kroll & Felten, 2015).

7. Conclusion

This work proposed a novel yet easy to implement AXAI capability framework for designing, analysing and assessing AI systems. The proposed framework, as demonstrated through examples, was easy to incorporate, domain-agnostic and useful in comparing and delineating AI systems. While measuring AXAI capabilities, the proposed framework also a measure of non-explainability raised in Ai et al. (2021). The measure of assessing the non-explainability is given as: non-explainability = 1 – explainability. Through the proposed AXAI framework, automated matching of ‘levels of abstraction’ (Palacio et al., 2021) 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 avoids proposing a domain-specific XAI capability incorporation method. Instead, the proposed AXAI framework 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 and part two of this paper, the AXAI framework also provides design guidelines and encourages provision of separable and quantifiable parameters of accuracy, comprehensibility and accountability in AI systems. This makes the proposed AXAI capability framework different from existing XAI approaches. 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 AI system design requirements is
Figure 7: Translation of the AXAI framework into a set of AI system design protocols which may be beneficial for the development of AI systems. Dark boxes highlight where the AXAI framework design protocols would be integrated.

demonstrated in part two of this paper. The two papers will be useful in developing the AI system requirements and producing a design process model as shown in Figure 7. The AXAI capability framework related stages of the AI system design are explicitly shown in Figure 7.

For building upon the initial success, the AXAI capability 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 AI systems.

Though the list of our proposed AXAI elements is not exhaustive, it seems to suffice the common comprehensibility, accuracy and accountability requirements. Nonetheless, this list of AXAI elements needs more input from legal practitioners, AI experts, software developers and cognition scientists. 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


