



Human-centric algebraic machine learning

AML based world models - Final

Deliverable D5.8

WP5 - Learning World Models Ethics
and Culture with AML



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952091.

Consortium



ALGEBRAIC AI



Champalimaud
Foundation



EPROSIMA
The Middleware Experts



German
Research Center
for Artificial
Intelligence



TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN



Project

Title: ALMA: Human Centric Algebraic Machine Learning
 Acronym: ALMA
 Coordinator: Proyectos y Sistemas de Mantenimiento SL (eProsima)
 Grant agreement: 952091
 Call: H2020-EIC-FETPROACT-2019
 Program: Horizon 2020
 Start: 01 September 2020
 Duration: 48 months
 Website: <https://alma-ai.eu>
 E-mail: alma@eprosima.com
 Consortium: **Proyectos y Sistemas de Mantenimiento SL (eProsima) (EPROS)**, Spain
Champalimaud Foundation (CF), Portugal
ALGEBRAIC AI (ALG), Spain
German Research Center for Artificial Intelligence (DFKI), Germany
Inria (Inria), France
TU Kaiserslautern (TUK), Germany
FIWARE Foundation e.V. (FF), Germany
Technical Research Centre of Finland (VTT), Finland
Carlos III University (UC3M), Spain

Deliverable

Number: **D5.8**
 Title: **AML based world models - Final**
 Month: 18
 Work Package: WP5 - Learning World Models, Ethics and Culture with AML
 Work Package leader: DFKI
 Deliverable leader: Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
 Deliverable type: R: Document, report
 Dissemination level: Public (PU)
 Date of submission: 2022-08-31
 Version: v1.6
 Status: Finished

Version history

Version	Date	Responsible	Author/Reviewer	Comments
v1.0	2022-06-07	VTT	Stephen Fox	Partial Draft
v1.1	2022-07-19	VTT	Stephen Fox	Expanded Draft
v1.2	2022-08-01	VTT	Stephen Fox	Complete Draft
v1.4	2022-09-02	DFKI	Vitor Fortes Rey	AML
v1.5	2022-09-07	DFKI, VTT, Inria	Vitor Fortes Rey, Stephen Fox, Janin Koch	Finalized Version
v2.0	2023-11-21	DFKI, VTT	Vitor Fortes Rey, Stephen Fox	Review Revision Version

Executive summary

As stated in the project summary, the ALMA project is concerned with interactive human-centric machine learning systems. There are three types of world models in interactive human-centric machine learning systems: the world models of individual people, of machine learning, and of human organizations. In this document, Deliverable (D) 5.8 - AML based world models - Final, we report findings from our continued research into the potential of Algebraic Machine Learning (AML) to support world models encompassing culture and ethics. This is the third and concluding deliverable in the AML based world models series: Initial (D5.4), Revised (D5.5), and Final (D5.8). In D5.4, we related AML-enabled analyses to interfaces between individuals' embodied models of themselves in the world and organizations' documented models of themselves in the world (i.e. world model interfaces). In D5.5, we considered the interpretability in world model interfaces of outputs from AML-enabled analyses. Our research reported in D5.5 indicated that AML has some potential to improve the interpretability of machine learning, but that AML would need to be developed further to bring about realization of that potential.

Here we report our research and development work to improve the interpretability of AML. In particular, we have evaluated AML-enabled gait analyses in terms of important constructs related to explainable artificial intelligence (xAI): explainability, transparency, and interpretability. Furthermore, we have evaluated in terms of shared interpretability of AML-enabled analyses outputs. Shared interpretation is a fundamental characteristic of culture and ethics, which are the focus of ALMA Work Package 5.

For example, shared interpretation is important because different models of self in the world can lead to different people making different, and even opposing, interpretations of the same information. Apropos, there is an ethical requirement to avoid conflicting interpretations of analyses related to functional disorders, including gait disorders. This is because different interpretations can lead to high treatment costs with poor treatment outcomes. For example, functional disorder patients who do not agree with a diagnosis are less likely to respond to treatment. In addition to evaluating AML in terms of important xAI constructs, we have developed AML application in order to improve its potential to facilitate shared interpretation. This development work has focused on maintaining the accuracy of outputs from AML-enabled analyses while reducing the complexity of their presentation. Development work has been informed by reference to hermeneutics, which is the study of interpretation processes.

Results of development work have also been evaluated in terms of explainability, transparency, interpretability, and shared interpretability. Evaluation indicates improvement, but highlights that barriers to shared interpretability are deeply-rooted in human embodied psychomotor functioning and will remain challenging.

This document comprises seven sections. First, the introduction section. Second, review of literature concerned with shared interpretation. Third, challenges in achieving shared interpretations are related to explainability, transparency, and interpretability. Fourth, evaluation of AML-enabled analyses before AML development is reported. Fifth, development work is described and re-evaluation is presented. Sixth, the challenges of shared interpretability are related to individuals' embodied models of themselves in the world, to organizations' documented models of themselves in the world, and to machine learning world models with an AML example. Seeking to achieve shared interpretability of world models is important when it is not possible to reach shared interpretation of machine learning outputs, and reference to world models could reveal sources of opposing interpretations. In conclusion, principal findings are stated and directions for further research are proposed.

Following this deliverable, research in Work Package 5 will deliver D5.6 Embedding of ethics and cultural concerns into AML based world models - Initial, and D5.7 Embedding of ethics and cultural concerns into AML based world models - Final. In those future deliverables, we shall report our further research into the potential of Algebraic Machine Learning (AML) to improve interactive human-centric machine learning systems in complex settings.

Contents

Executive Summary	3
Contents	4
1 Introduction	5
2 Shared Interpretation	6
2.1 Hermeneutics	6
2.2 Risks and Ambiguities	7
2.3 Referents as Boundary Objects	7
2.4 Compressed (Simple) versus Expressive (Meaningful) Referents	9
2.5 Minimizing risk	10
2.6 Minimizing ambiguities	11
3 Explainability, transparency, and interpretability	11
3.1 Explainability	12
3.2 Transparency	12
3.3 Interpretability	13
4 Evaluation of AML-enabled analyses report	13
5 Development work to improve the shared interpretability of AML	18
6 Shared Interpretability of World Models	21
6.1 World models in interactive human-centric machine learning systems	21
6.2 Shared Interpretability of World Models: An AML example	23
6.3 Shared interpretability for culture and ethics in AML world models	27
7 Conclusions	28
References	31

1 Introduction

As stated in the project summary, the ALMA project is concerned with interactive human-centric machine learning systems. Human trust in artificial intelligence (AI) can depend upon AI being explainable [1]. It is preferable that AI models are transparent so that they can be interpreted directly by people. In particular, direct interpretability is preferable to there having to be an additional model generated by the AI that is intended explain its outputs to humans [2].

However, irrespective of whether or not AI is used, interpretability does not necessarily lead to different people agreeing a shared interpretation of the same information. For example, there can be conflicting interpretations of the same topic in research [3] and of the same topic in practice [4]. There can be conflicting interpretations of large amounts of information [5] and of diagrams, such as trees, that are intended to simplify the presentation of information [6]. There can be conflicting interpretations of one written word [7] and even of a single drawn line [8].

Conflicting interpretations of the same information can arise from different models in self in the world. For example, a person in need of healthcare support has an embodied of self in the world [9], while a healthcare organization has a documented model of itself in the world [10]. There can be ethical requirements to avoid conflicting interpretations. For example, there is an ethical requirement to avoid conflicting interpretations of analyses concerning healthcare related to functional disorders. This is because different interpretations can lead to high treatment costs with poor treatment outcomes [11]. For example, functional disorder patients who do not agree with a diagnosis are less likely to respond to treatment [12].

Interpretation processes are studied in hermeneutics [13], [14]. It has been recognized for some years that consideration of hermeneutics can improve the implementation of information systems. Importantly, hermeneutics can encompass both subjective meanings for individual people and the social structures that shape shared meanings [15]. The hermeneutics of shared interpretability has been addressed by neuroscience research [16], [17], which has subsequently been applied to social cognition [18] and to human-robot interaction [19], [20]. Here, we report findings from a case study informed by hermeneutics, which has been carried out with the intention of improving the shared interpretability of results from analyses carried out by machine learning (ML).

In particular, our work sought to improve the shared interpretability of algebraic machine learning (AML). This is the focus of the case study because AML has potential to facilitate shared interpretability. Here, we carry on work reported in D5.5 concerned with the interpretability of AML-enabled gait analyses. Specifically, AML-models have a fundamentally simple shallow structure. They comprise only three layers: inputs – atoms – outputs. Second, AML algorithms check that human-defined constraints are maintained in relationships between inputs and outputs that are learnt as new atoms are generated during training. Third, AML rules can be traced back to human-defined policies. However, the potential for shared interpretability of AML models and outputs from AML-enabled analyses is reduced by the number of atoms increasing from tens to thousands during training. Thus, AML has novel potential to improve upon existing machine learning options in theory, but in practice entails familiar challenges to interpretability, such as a lightly constrained model form expanding exponentially during operation [21].

Findings from the case study are reported in the remaining **six** sections. In section 2, review of literature concerned with shared interpretation is provided. In particular, literature from neuroscience and behavioral sciences that provide a first principles account of shared interpretation. In section 3, challenges in achieving shared interpretations are related to explainability, transparency, and interpretability. In section 4, evaluation of AML-enabled analyses is reported in terms of explainability, transparency, interpretability, and shared interpretability. In section 5, development work to improve the shared interpretability of AML is described and subsequent re-evaluation of AML is reported. **In section 6, the challenges of shared interpretability are related to individuals' embodied models of themselves in the**

world, to organizations' documented models of themselves in the world, and to machine learning world models with an AML example. In conclusion, principal findings are stated and directions for further research are proposed in section 7. Overall, a first principles account of shared interpretability for agreement among potentially opposing individuals and organizations is applied to machine learning. In doing so, the paper goes beyond previous studies that have focused upon explainability rather than agreeability [22], [23], [1].

2 Shared Interpretation

2.1 Hermeneutics

It has been recognized for some years that reference to hermeneutics can improve the implementation of information systems [15], and recently that hermeneutics could contribute to interpretable AI [24]. Hermeneutics involves study of interpretation processes. It encompasses communication for interpretation and understanding from interpretation [25]. Single hermeneutic is concerned with one-way interpretation. For example, natural scientists try to understand the natural world through their interpretation of their observations of the natural world. Yet, the natural world is typically not involved in reciprocal exchanges with natural scientists to try to understand them. For example, a geologist seeks to understand rocks through observations of rocks, but as far as we know rocks do not have reciprocal exchanges with geologist to try to understand them.

By contrast, double hermeneutic is concerned with two-way interpretation. For example, social scientists interpret social realities that have already been interpreted by individuals and institutions. Then, the reinterpretations of social scientist are fed back into society and can shape subsequent interpretations of social reality by individuals and groups [26]. An example of this is the social science interpretation, social class, having entered into common usage in society by individuals and groups. It has been argued that people can engage in exchanges that lead to shared interpretations in so-called "hermeneutic niches" [27]. It has been argued that people evolved to do this when securing resources necessary for survival first depended on cooperation with others. For example, when individuals needed to align their mental states to obtain a high risk, but high reward, food item, such as a large wild animal [28]. Exchanges that lead to shared interpretations can involve individuals' initially having divergent rich mental representations, which converge eventually to shared succinct representations that are accessed automatically [29]. Shared interpretation is illustrated in Figure 1. This shows that two participants have the same interpretation of the same document. Thus, there is little, if any, cognitive distance between the participants regarding the contents of the document [30].

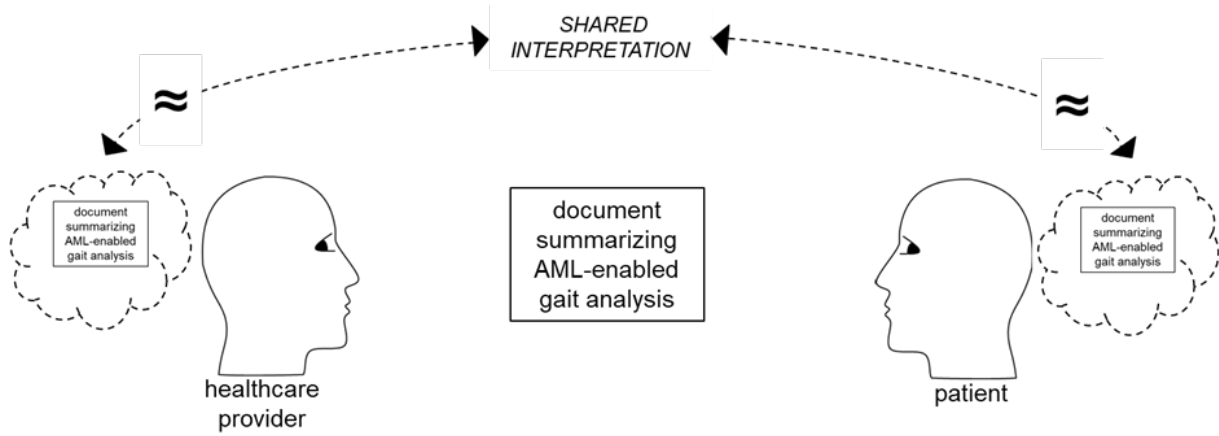


Figure 1: Shared interpretation of a document

2.2 Risks and Ambiguities

Reciprocal exchanges leading to shared interpretations for joint learning and development have been addressed in neuroscience research. In particular, it has been argued that reciprocal exchanges involve trying to reduce participants' prediction errors about what other participants have in their minds. In this framing, what is in mind is described in terms of internal generative models of self in the world [16], [17]. These internal generative models of self in the world should resemble environments in which the self intends to survive [31]. This entails interrelationships between the risks to survival and internal generative models complexity. It also entails interrelationships between environmental ambiguity and the accuracy of predictions from internal generative models [32], [33].

To minimize risk in trying to survive, internal generative models should be sufficiently complex: i.e. neither over-fitted nor under-fitted to the environment [34]. For example, risk can be high if a non-linear model is used to try to survive in a linear environment (i.e. an over-fitted model). Conversely, risk can be high if a linear internal model is used to try to survive in a non-linear environment (i.e. under-fitted model). Hence, survival risk can be considered in terms of mismatches between the complexity of the environment in which a participant intends to survive and the complexity of the participant's internal generative model of self in that environment. Predictions from an over-fitted or under-fitted internal generative model can lead to prediction errors about self in the environment that cannot be surviveds [32], [33].

At the same time, there is a trade-off between environmental ambiguity and the predictive accuracy of internal generative model. In particular, if information from exchanges with the environment are ambiguous, exchanges may not provide sufficient information to verify the accuracy of past predictions from the internal generative model and inform the formulation of accurate predictions from the internal generative model in the future. Accordingly, participants can be expected to prefer exchanges that provide information that is least ambiguous for them [32], [33]. However, this can lead to participants being fixed on exploiting their existing internal models in their familiar environmental niches. Rather than being adaptable in updating their internal generative models as changes in the world come to change their familiar environmental niches [35], [36].

2.3 Referents as Boundary Objects

Reciprocal exchanges involve iteratively attending and attenuating incoming sensory information as participants become attuned to each other. Attending to incoming sensory information can involve listening

and looking. Attenuating incoming sensory information facilitates speaking and showing. During exchanges, participants seek to infer how incoming sensory information are caused by other participants. This can carry on until there is attunement between participants, which involves synchrony between participants' internal brain states that generate predictions about each other. Here, synchrony refers to reciprocal action at a distance among coupled dynamical systems [16], [17]. Research indicates that the meaningfulness of communication is influenced by interbrain synchrony [37], [38].

Such synchrony can be facilitated by participants paying joint attention (i.e. attending together) to the same referent [39], [40]: for example, by paying joint attention to the same gait analysis report document. A referent can act as an ostensive cue that “grabs” the attention and stimulates comment [41], [42]. Joint attention enables, and is enabled by, individuals' capacity to ground their exchanges in the same referents as they endeavour to develop a set of mental states that they can infer to share with each other. Colloquially, such mental states can be called participants' common ground [43], [44]. Regulating exchanges towards increasing common ground involves participants trying to align their exchanges with what they infer to be the needs of other participants. When there is already a lot of common ground among participants, there may be need for only very brief exchanges to align mental states [45]. Thus, the amount of information necessary to align mental states sufficiently to enable cooperative behaviour is inversely proportional to the amount of common ground shared by participants.

However, the development of common ground for shared interpretation of information relevance can be hindered by insufficient collaborative effort. For example, seeking to optimize the relevance of exchanges can involve participants sending the minimal information that could potentially enable intended meaning to be inferred by other participants with least effort. At the same time, the most parsimonious meaning may be inferred from the received information that could potentially explain the speaker's intentions with least effort [46], [47]. This can be described as exchanges conforming to the principle of least collaborative effort [39], [48]. As summarized in Figure 2, the development of common ground could be hindered by people preferring to make least effort by paying more attention to their internal generative models than undertake more effort in paying attention to incoming sensory inputs [49], for example, sensory inputs from paying joint attention to the same referent such as a gait analysis report document. In neurobiology, this is often construed in terms of attentional selection that involves allocating more or less precision to various beliefs at different levels in a hierarchical world model [50]. This has found a particularly powerful explanatory role in many areas of cognitive science and computational psychiatry [51].

As summarized in Figure 2, referents can be considered to be boundary objects between an organization's documented internal model of itself in the world, and an individual's model of self in the world. Boundary objects have different meanings in different social worlds but nonetheless can be meaningful to more than one social world. Boundary objects include written words, spatial diagrams, and any other types of information that can facilitate development and maintenance of coherence across intersecting world models [52], [53]. Preferring least action through reference to existing internal models can lead to organizations' lock-ins [54] and path dependencies [55] coming into opposition with individuals' motivated cognition [56] and wishful seeing [57]. All of which can lead to referents being interpreted to serve explanations that support preconceptions [58] and confirm biases [59].

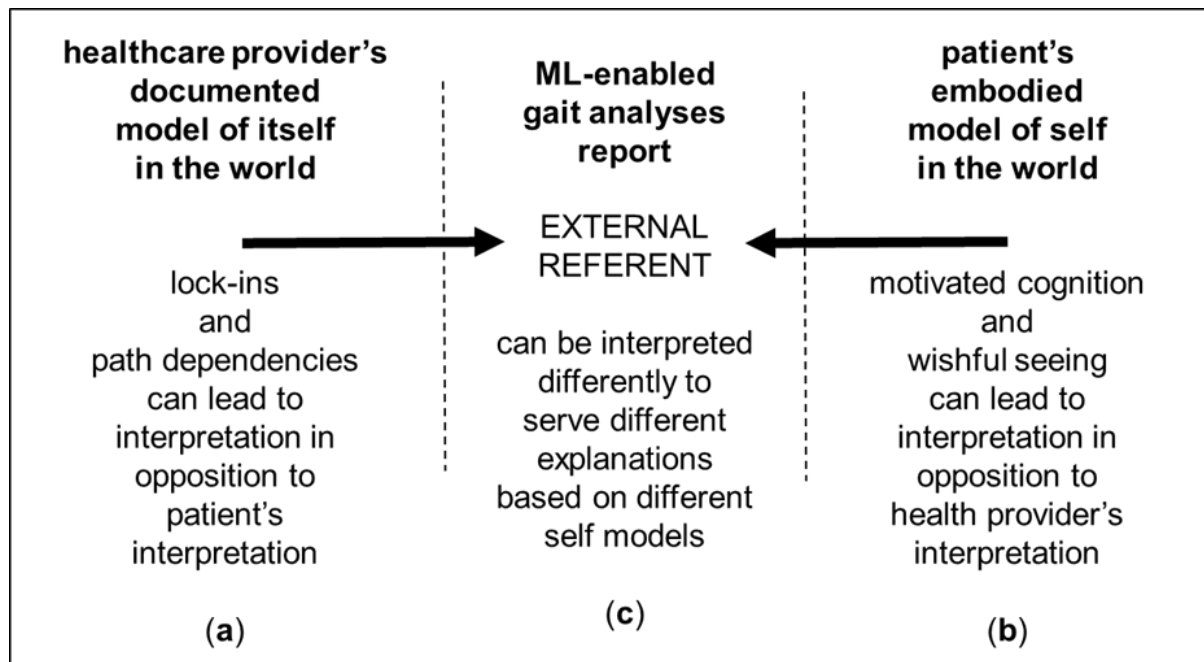


Figure 2: Opposing interpretations of same referent. (a) In accordance with the principle of least collaborative effort, representatives of organization can make more reference to its internal model than to the external referent. (b) In accordance with the principle of least collaborative effort, individual person can make more reference to own internal model than the external referent. (c) External referent is not effective as a boundary object due to participants making more reference to their own internal models than the content of the referent.

2.4 Compressed (Simple) versus Expressive (Meaningful) Referents

Referent facilitated exchanges that are compressed/simple while being sufficiently expressive/meaningful are preferable for least effort. However, such exchanges may not be possible at the beginning of exchanges between participants that have different internal models of self in the environment. Rather, individuals can initially have divergent rich mental representations. These can eventually converge to shared succinct representations that are accessed automatically [29], but this requires collaborative effort to overcome risks and ambiguities.

In terms of survival risk arising from internal generative model complexity, and environmental ambiguity affecting internal generative model prediction accuracy, under-fitting of internal generative models could happen if a few compressed/simple exchanges are inferred erroneously to be sufficiently expressive/meaningful due to their ambiguity (i.e. Type 1 error). Conversely, over-fitting of internal generative models could happen if excessive exchanges are made because many compressed/simple exchanges are inferred erroneously not to be sufficiently expressive/meaningful due to their ambiguity (i.e. Type 2 error). Thus, efforts to achieve minimum description length and minimum message length for compression/simplicity in exchanges between participants need to take into account the internal generative models of participants in exchanges [50].

In addition to Type 1 errors, i.e. false positive inference of relevance, and Type 2 errors, i.e. false negative inference of irrelevance [60], there can also be Type 3 errors. Such errors arise from discrepancy between the intention of exchanges and the information in exchanges. For example, in medical research, discrepancy between research focus and research question [61]. Type 3 errors can be considered to be

inept positives, because the information that is being exchanged between participants may be correct but is not relevant to the intention of the exchanges between participants. Risk arising from internal generative model under-fitting or over-fitting because of Type 3 error can be more likely if the intention of exchanges between participants is ambiguous. Type 3 errors can be related to what has been described as issue distance [62].

Consideration of potential for Type 1, 2, and 3 erroneous inferences highlights the need for deliberative methods to be applied to minimize their occurrence [63]. Not least, because there can be circular causation between the fit of internal generative models to the environment and erroneous inferences, in which badly fitted internal models are more likely to generate erroneous inferences and erroneous inferences are more likely to worsen the fit of models. For shared interpretation of AML-enabled gait analyses report documents, deliberative methods need to be applied to minimize risk and ambiguity related to the intent and the content of exchanges between participants, such as people with gait disorders and organizations providing healthcare.

2.5 Minimizing risk

It has been argued that human capabilities for making shared interpretations evolved when people were hunter-gatherers and securing resources necessary for survival depended on cooperation with others [28], [27] including sharing information with others [64], [65]. These were in times when the evolution of behaviors often took place in environments that changed little over millennia. By contrast, today, rapid changes to environments come from the introduction of artificial intelligence (AI) and the infrastructures that facilitate its application [66]. Accordingly, today, human internal generative models need to synchronize concurrently with unfamiliar environments and with the mental states of unfamiliar people through paying attention to what can be unfamiliar referents, such as AML-enabled gait analysis report documents. Such cognitive challenges exceed those in circumstances during which human adaptive cognition evolved [67], and can go against the human need to minimize neurological energy expenditure [68], [69], [70], [71].

Hence, participants may experience cognitive overload in trying to achieve shared interpretations, which depend upon fitting concurrently their internal generative models with unfamiliar environments and with unfamiliar participants. Cognitive overload can arise from characteristics of information and from the cognitive effort required to create related cognitive schema [72]: for example within an internal generative model. Within cognitive load theory, it is posited that a fundamental bottleneck in creating new cognitive schema is the limited processing capacity of human working memory, which can handle only a very limited number—possibly no more than two or three—of novel interacting elements [73].

Hence, cognition can be overwhelmed by the number of information elements and interactions between them that need to be processed simultaneously [74]. Moreover, the cognitive load of trying to make sense of information and the cognitive load of creating new schema are additive as both can consume working memory resources simultaneously [75]. As cognitive overload is detrimental, cognitive task analysis techniques have been developed in order to assess cognitive loads and how they can be reduced [76].

Cognitive task analysis can be applied to referents such as gait analysis reports in order to minimize the potential for cognitive overload during their interpretation hindering the concurrent synchronization of internal generative models with unfamiliar environments (e.g. gait recording laboratory) and with the mental states of unfamiliar people (e.g. healthcare providers). Application of cognitive task analysis is important to facilitate least effort for participants in exploring the information in gait analysis report documents. Cognitive task analysis can encompass different types of cognitive distance: for example, issue distance and semantic distance. Issue distance represents the cognitive effort required to understand that a shift in goal is needed to achieve completion of the task [62]. Semantic distance can be considered in terms of path length comprising the amount of steps needed to traverse from one word to another word

in the same path. The more steps that are needed, the more less likely that words, which are intended to be related, will be inferred to be related [77].

2.6 Minimizing ambiguities

If information from exchanges with the environment are ambiguous, exchanges may not provide sufficient information to verify the accuracy of past predictions from the internal generative model and inform the formulation of accurate predictions from the internal generative model in the future. Ambiguities need to be minimized to reduce potential for participants to revert to their existing internal generative models. It is important to avoid reverting to existing internal models in a changing world because it can increase potential for erroneous inferences [78]. For example, erroneous preconceptions can override diagrammatic depictions of information, such as tree diagrams [6].

Information in exchanges between people, and between people and technologies, can have linguistic, presentational, and conceptual ambiguities [79]. Conceptual ambiguities can span across disciplines [80]. It has been argued that the term, medically unexplained symptoms is a barrier to improved care because it is conceptual ambiguous [81]. Presentational issues include color and layout [82], [83]. The color of a line can be ambiguous, and the thickness of a line can be ambiguous. Presentational ambiguity can also arise from the way in which written text is presented [84]. Linguistic issues can be lexical, syntactic, semantic, and phonological. Lexical ambiguity can arise when a word, has more than one meaning. Syntactic ambiguities can arise from sentences that can be parsed in more than one way. Semantic ambiguity can arise when the meaning of a sentence could be determined only with the help of greater knowledge sources [85]. Also, semantic ambiguity can arise if the same words elicit different cognitive or emotional states [86]. Phonological ambiguity can arise when a set of sounds can be interpreted in more than one way [87].

Potential ambiguities in AML-enabled gait analysis reports can be addressed through, for example, concept mediation. This involves establishing conceptual equivalence among participants with diverse backgrounds [88], [89]. Also, ambiguities can be reduced through so-called concretization, which increases specificity, for example through direct reference to a solid object [90]. For example, line colors and thicknesses can be less likely to be sources of ambiguity in pictures of physical objects [91].

3 Explainability, transparency, and interpretability

As explained above and summarized in Table 1, risks arise from individuals' internal models of self in the world being over-fitted or under-fitted to the world in which they intend to survive. Ambiguities affect the predictive accuracy of internal models by not providing sufficient information to verify the accuracy of past predictions from the internal generative model and inform the formulation of accurate predictions from the internal generative model in the future. Referents such as gait analyses report documents can contribute to risks by being sources of cognitive overload. Also, referents such as gait analyses report documents can include linguistic, presentational, and/or conceptual ambiguities. Apropos, to facilitate shared interpretation, referents need to support minimizing cognitive distance, and referents need to support concretization and concept mediation. However, there is little consideration of these fundamental issues in literature concerned with explainability, transparency, and interpretability.

Challenge	Internal model	Referent	To share interpretation
Risks	Over- or under- fitted to world	Can cause cognitive overload	Minimize cognitive distance
Ambiguities	Erroneous inferences about world	Can be linguistic, presentational, conceptual	Concretization, concept mediation

Table 1: Challenges in achieving shared interpretations.

3.1 Explainability

The term, explainability in the context of AI, can be associated with explanations of the internal workings of artificial intelligence (AI), including machine learning, during training and when making decisions. Such explainability can be summarized with the abbreviation xAI (eXplainable artificial intelligence) [92]. However, there is also need for AI to be explainable to humans via the interfaces that they use in, so-called, human-centered AI: i.e. HCAI. This is because AI is deployed within tools that are used by people [93]. In other words, to use a colloquial automotive analogy, there should be explainability of what is under the hood and explanation of what is above the hood.

It has been argued that, so called, causability, is an antecedent of explainability [1]. In this context, causability has been defined as the extent to which an explanation of a statement to a user achieves a specified level of causal understanding [94]. Some recent opinions related to AI do not recognize that Aristotle opined more than two thousand years ago that causation and explanation cannot be understood separately [95]. Furthermore, recent opinions related to AI involve composition of new descriptions of interactions between causation and explanation rather than making reference to the more than two thousand years of debate concerning this topic. For example, reference can be made to the relevant entry in The Oxford Handbook of Causation in which it is stated that a causal model of explanation maintains that to explain some phenomenon is to give some information about its causes. Yet, it is also stated that not all explanations are causal and not all causes are explanatory [96].

As well as making little reference to the more than two thousand years of relevant debate about explanation, xAI literature and HCAI literature includes little consideration of the potential for explanations to entail risks and ambiguities that can undermine shared interpretations. However, there is some recognition that different people can have different requirements for explanations [97], and that even experts may not agree about explainability [98].

3.2 Transparency

The term, transparency in the context of AI, is associated with the extent to which the internal workings of artificial intelligence can be observed [99]. This is summarized in well-established comparisons of metaphorical opaque “black box” computing and transparent “glass box” computing [100], which have recently been applied to AI [101]. It has been argued that as the complexity of the AI systems and algorithms grow, people increasingly consider them as “black-boxes” that defy understanding [102]. Accordingly, there are widespread efforts to improve transparency.

Often, these efforts are not all encompassing. For example, there is focus on improving the transparency of AI model or on improving the transparency of data outputs from AI models. These can be considered in terms of global characteristics or local characteristics when global transparency encompasses an entire AI model and local transparency encompasses individual data samples. It can be possible to bridge the global and local through feature engineering that involves designing a model around features that are not opaque to any potential end-users. For example, a gait analysis model can be designed around features that both healthcare providers and patients can understand, such as head posture; left/right stride length; left-arm/right-arm swing; walking speed, and body swaying [103].

However, the display of important features is not inherent to some AI such as deep neural networks. Furthermore, even when model interpretation via feature importance is available at a global level, estimating how the model behaves for an individual example is non-trivial. For example, knowing the feature importance for a random forest model is not sufficient for users to grasp what-if scenarios at a particular example level, such as increasing/decreasing feature values by a certain amount [98]. Accordingly, some types of AI models can only ever be black-boxes, and attempts to make them explainable are not likely to be successful [2]. Yet, advocates of transparent models give little consideration of the potential for transparent models to provide referents, which entail risks and ambiguities that can undermine shared interpretations. For example, transparent models may entail cognitive overload.

3.3 Interpretability

Within research concerned with AI and ML, there is no one universally accepted definition of interpretability [104]. However, distinction is drawn between post hoc explanation of black-box models and directly interpretable/self-explanatory models. It has been argued that, for example, deep neural networks require post hoc explanation, but classic approaches in machine learning and AI, such as decision tree and rule-based models, can directly develop interpretable models. That is provided there is sufficient domain knowledge and feature engineering involved in the formulation of decision tree and rule based models [98].

A recent comprehensive survey of interpretability has yielded a taxonomy [105] that distinguishes between interpretability methods to *explain* black-box models (interpretability methods to explain deep learning models and any black-box model); interpretability methods to *create* transparent models; interpretability methods to *restrict* discrimination and enhance fairness in machine learning models; and interpretability methods to *analyse* the sensitivity of machine learning model predictions (traditional sensitivity analysis methods and adversarial example-based sensitivity analysis).

There is no mention in this very thorough and comprehensive survey of interpretability methods [105] of consideration of shared interpretation or application of hermeneutics in the field. Furthermore, there is no mention of there being consideration of fundamental interrelationships between risks and internal models complexity. Nor is there mention of fundamental interrelationships between environmental ambiguity and the accuracy of predictions from internal generative models. The survey does reveal that there is much striving for simplicity in the field, but the survey also reveals that there is a lack of consideration of fundamental trade-offs between compression/simplicity versus expressive/meaningful, which can contribute to type 1, type 2, and/or type 3 errors in inference. Moreover, the survey reveals that there is lack of consideration of underlying cognitive issues in interpretation. Notably, neither the word, cognition, nor the word, cognitive, are included in the survey. Accordingly, the survey reveals a lack of application of methods to reduce potential to address cognitive overload or to reduce potential sources of ambiguities. Hence, it is apparent that while there is a wide range of excellent work being done in the field of AI interpretability and ML interpretability, there has been little previous consideration of facilitating shared interpretation through consideration of fundamental factors.

4 Evaluation of AML-enabled analyses report

AML in its current form, has two important components: atoms and constants (Figure 3). Constants define the nature of the machine learning classification problem, while atoms are learned by the algorithm. Problems are described as embeddings including data (classification instances made of constants) and formulas about constants (the knowledge about the problem at hand). Once the atoms for a specific class are learned, decisions are taken based on the number of atoms missing for that class. The more atoms are present (not missing), the more evidence there is for that class. An atom is considered to be not missing when at least one of its constants is present. In our gait analysis examples, we use human-

interpretable gait features to predict if depression was detected or not. The data set is from a previous study by others, which was focused on detecting depression through gait data [103]. Interpretable features from that previous study are represented using constants that indicate their range ($G_v(f)$ constants for feature v if bigger than v and $LE_v(x)$ for feature f is smaller or equal to value v). Each atom of a class is connected to one or more of these features, generating the structure shown in Figure 3.

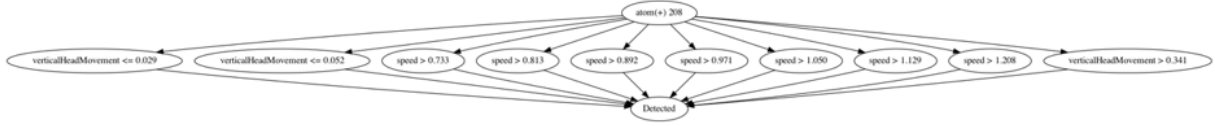


Figure 3: One AML atom connected to constants

In terms of explainability, some everyday words are used such as speed and head movement. Thus, there is potential for human reading of the diagram. However, the overall structure, its components, and interrelationships between them are not necessarily self-explanatory. In order to interpret this diagram correctly, one needs to interpret atom presence as an “or” between the constants that represent feature intervals in the middle layer. For example, the top atom is present if and only if at least of one of the constants is valid. In this example, one constant can be vertical head movement being too high or vertical head movement being too low or the walking speed being too high. Moreover, interval values are not succinct, with the same one variable, for example speed, appearing many times instead of being summarized in one entry. This illustrates that showing the overall raw structure of the atoms, even for one example, can be ambiguous unless further pre-processing is carried out.

Regarding the knowledge for the problem at hand, Figure 4 shows an AML embedding description for gait analysis using AML Description Language. The embedding itself represents the notion of ordered intensity. That is, specifying that interpretable features are represented as discretised numbers. Regarding interpretability, it requires not only knowledge of programming language words, but also of the AML Description Language. Thus, it is not easily interpretable for non-experts. Regarding transparency, the definitions are clear for experts and the final trained model can be a “glass-box” model: at least for experts. In terms of shared interpretability, we have observed confusion on the possible solutions the model can generate: even among people who have backgrounds in computer science. Hence, it is questionable whether AML Description Language can be used to formulate boundary objects even for the description of everyday lived experience such as gait, which involve specifying interpretable features as discretised numbers, for example the speed of walking.

```

1 import numpy as np
2 from amldl import (
3     ADD,
4     APP,
5     C,
6     CV,
7     CMP,
8     Descriptor,
9     EXC,
10    F,
11    HEADER,
12    INC,
13    M,
14    R,
15    T,
16    V,
17    S,
18 )
19
20 def embedding(inconsistent):
21     _i = 1
22     # We used 10 gait features, each with values from 0 to 15.
23     NUMBER_OF_GAIT_FEATURES = 10
24     GAIT_VARIABLE_RANGE = 16
25
26     with Descriptor() as theEmbedding:
27         # Create the constants representing
28         if HEADER("feature constants created"):
29             # Create the constants related to less or equal.
30             for v in range(GAIT_VARIABLE_RANGE - 1):
31                 CV("LE" + str(v), NUMBER_OF_GAIT_FEATURES)
32             # Create the constants related to greater.
33             for v in range(GAIT_VARIABLE_RANGE - 1):
34                 CV("G" + str(v), NUMBER_OF_GAIT_FEATURES)
35
36             # Create constants related to the two classes: not detected (0)
37             # and detected (1).
38             if (HEADER("Outputs")):
39                 CV("O", 2)
40
41             # # Create ordering between the constants, creating intensity ordering.
42             if HEADER("Inside feature separation"):
43                 theEmbedding.REGION = 1
44                 # Ordering for the greater than chain.
45                 for x in range(GAIT_VARIABLE_RANGE - 2):
46                     print("G" + str(x), _i, x, x+1)
47                     ADD(INC(T("G" + str(x), _i), T("G" + str(x+1), _i)))
48                     ADD(EXC(T("G" + str(x+1), _i), T("G" + str(x), _i)))
49
50                 # Ordering for the less or equal than chain.
51                 for x in range(GAIT_VARIABLE_RANGE - 2):
52                     ADD(INC(T("LE" + str(x+1), _i), T("LE" + str(x), _i)))
53                     ADD(EXC(T("LE" + str(x), _i), T("LE" + str(x+1), _i)))

```

Figure 4: Part of AML embedding for gait analysis using AML Description Language

Another challenge for shared interpretability is cognitive overload. If one has to analyse many atoms, it soon becomes not feasible to interpret the model, even if each atom alone may be transparent. This problem is illustrated in Figure 5, which comprises less than 30 atoms.



Figure 5: Several AML chains of input – atom – outputs

Considering that final AML models can have thousand of atoms, there is low practical potential for direct interpretability. Hence, there is need for generation of a secondary explanatory representation such as a tree. Given the amount of information present, it is challenging for participants to achieve a shared interpretation: instead of picking and choosing interpretations that support their existing internal models.

As it not feasible to look at all atoms, one may simplify the model by looking at only the most relevant ones. One way to do this is to use atom presence as a feature and then build a decision tree. Figure 6 shows tree generated by using CART from the AML-enabled gait analysis. This illustrates that a tree generated using standard technique such CART has potential for increasing explainability. In particular, everyday English language words are used within the tree, such as head posture; walking speed; left-arm/right-arm swing; left/right stride length; and body swaying. Also, the tree has a consistent rationale in its descriptive progression from there being a sample of 222 participants in gait analysis.

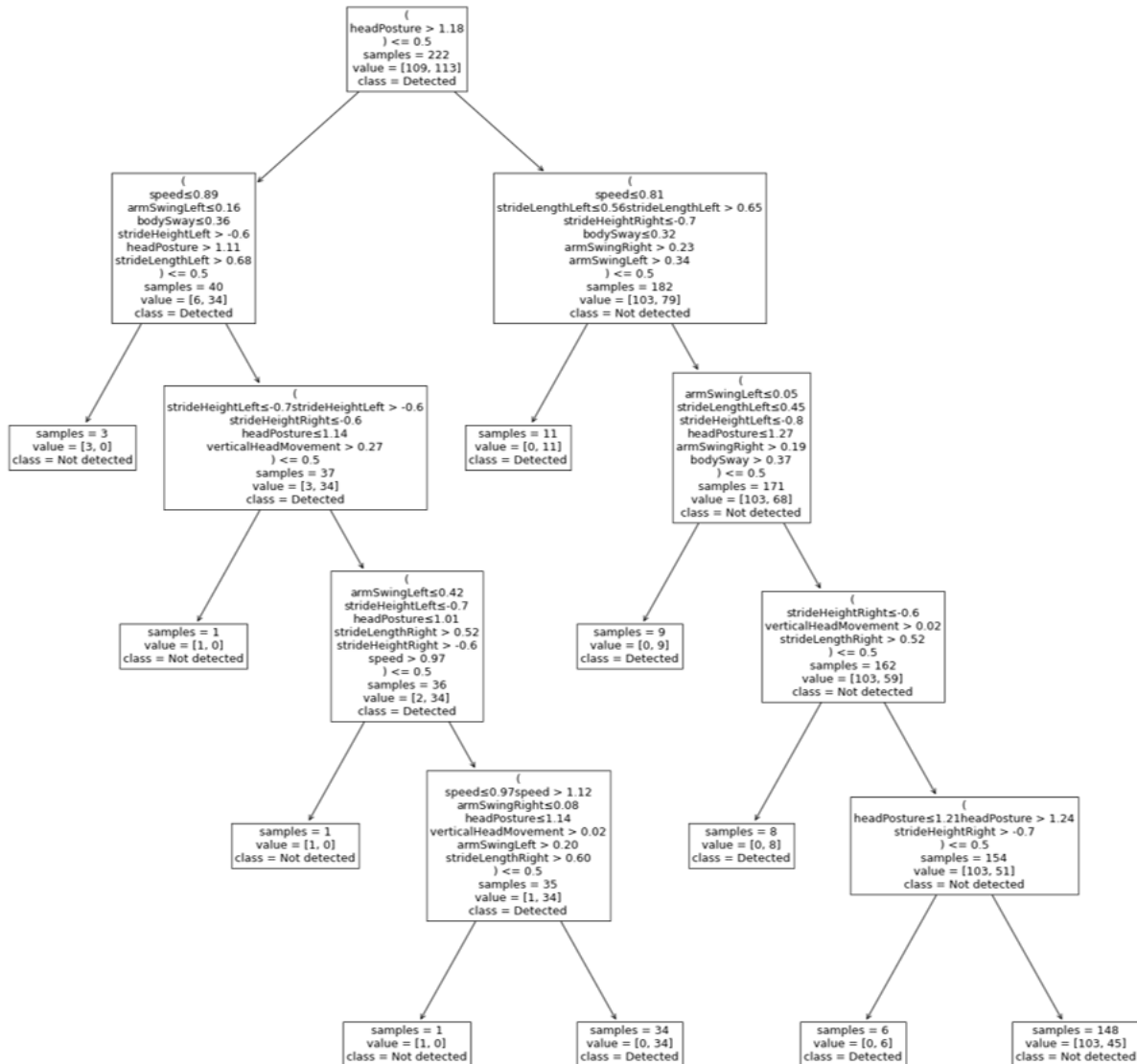


Figure 6: Tree representation of results of AML-enabled gait analysis

At the top of the tree, which is the beginning of the tree, gait indicators of depression were not detected in 109, but were detected in 113. Hence, the classification stated in the first tree node is Detected. In this first node, the analysed feature is head posture. The results of subsequent feature analyses are shown in lower nodes. In particular, the next level of the tree in the left node shows that analyses of walking speed, arm swing left, body sway, stride height, head posture, and stride length indicate that 34 of 40 participants have indicators of depression. By contrast, in the right node shows that indicators of depression are not detected in the majority of the other 182 participants (i.e. 222 participants minus 40 participants). The majority being 103 participants. Presentation of the analysis results proceeds with this rationale from the top to the bottom of the tree.

The tree is transparent. However, the extent to which the tree is interpretability is questionable. This is because the tree provides post hoc explanation of a model rather than direct interpretation of a model.

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

With regard to shared interpretability, the tree shown in Figure 6 presents risks and ambiguities. In particular, there are many information elements and interactions that need to be processed in order to understand why categorizations are made. Furthermore, there can be both types of cognitive distance to overcome. For example, there can be issue distance between healthcare providers and people with gait disorders because the goal of the tree presentation is not stated clearly. Also, there can be semantic distance because spatio-temporal terms are presented in different sequences in different nodes. In addition, there are ambiguities. For example, the top node of the tree contains the information: $(\text{headPosture} > 1.18) \leq 0.5$, and the information that depression is detected in the majority. By contrast, in the fifth level, there is $(\text{headPosture} \leq 1.21 \text{ headPosture} > 1.22 \text{ strideHightRight} > -0.7) \leq 0.5$, and the information that depression is not detected in majority. Thus, there is a lack of obvious consistency about the measurements and how they relate to categorization.

5 Development work to improve the shared interpretability of AML

One of the major hurdles for shared interpretability is in the nature of interpreting atoms misses, where an atom is said to be missing if all its constants are not present. In order to decrease potential for cognitive overload, it can be best to analyse only one feature at a time. Furthermore, one can focus on the most important features that the AML model is using to distinguish between the classes.

As our AML model makes decisions based on the number of misses, we built a new decision tree where each decision is on a single interval (that means one feature constant) and that constant is selected using the information gain of the AML prediction based on the number of misses for the both class models. An example can be seen in Figure 7, where we have capped the number of decisions to 5.

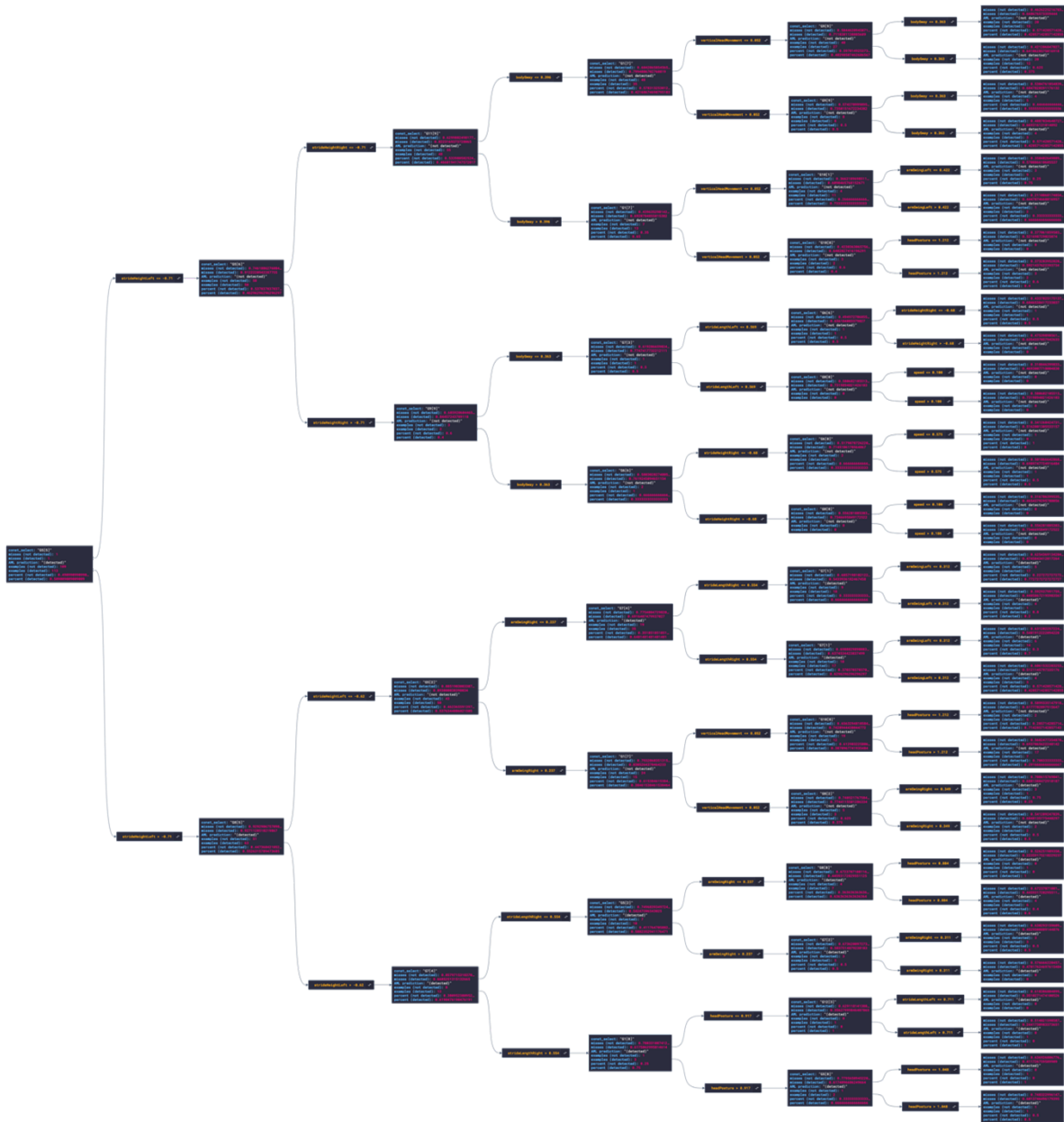


Figure 7: Example of developed tree

As shown in Figure 8, potential for cognitive overload can be decreased if we zoom to the first decision. This is because one only needs to analyse both sides of a single feature interval, as is common in decision trees. On the other hand, the diagram is still using terms related to AML, such as misses, which may be ambiguous. For example, there can be a type 1 error if it is assumed that the words not detected are restating for emphasis the word misses. However, this can be followed by a type 2 error when reading the next line in the tree box: misses (detected), if it is assumed that sometimes misses means not detected but sometimes misses means something else. Overall, there can be a type 3 error as the goal of not detecting

and detecting misses is not related to a clearly stated purpose for the gait analysis. Furthermore, words such as “predicted”, have a specific meaning in machine learning, in particular representing the decision provided by a given model, while having a different meaning in common parlance, such as “predict the future”.

By changing words, it is possible future development work could reduce potential for ambiguity within a fixed group of participants, such as one computer scientist, one medical doctor, and one patient. For example, the word prediction could be changed to the word inference or finding, and the meaning of those words in this context could be agreed among those participants. However, these words would not necessarily be intuitively understandable to another group of participants.



Figure 8: First decision in the developed tree

Moreover, as shown in Figure 9, if the view of the tree is expanded, it is apparent that there is increased potential for cognitive overload. Apropos, it is important to note that after 3 decisions most nodes still have more than half their number of misses for both classes. This means that there is still a lot of the model’s structure that is not incorporated in this diagram. Even at the five decisions shown in Figure 7, there is no tree node with fewer than a fifth of its atoms present. Hence, there are fundamental barriers to the effectiveness of common representations like trees as boundaries objects: even when they are intended to describe analyses of everyday lived experience, such as walking, and include everyday words such as stride. Thus, AML in its current form, offers limited potential for overcoming fundamental challenges in shared interpretability.



Figure 9: Expansion of the developed tree

This has wider implications for the shared interpretability of machine learning, because AML has a comparatively simple shallow structure of inputs-atoms-outputs from which the relatively simple structure of trees can be generated. Accordingly, it cannot be expected that breakthroughs in shared interpretability will inevitably only come from innovations in machine learning. Rather, shared interpretability will continue to depend upon willingness among participants to make sufficient effort to synchronize interpretations through multiple iterative exchanges. In particular, multiple iterative exchanges focused on referents that can initially entail much more risk and ambiguity than participants' internal models.

6 Shared Interpretability of World Models

There can be many outputs from applications of machine learning. The example in the preceding sections is AML-enabled gait analyses, which can be presented as on-screen displays or as on-paper reports. If it is not possible to reach shared interpretation of machine learning outputs, reference to world models could reveal sources of opposing interpretations. However, there would need to be shared interpretation of world models to enable agreement about sources of opposing interpretations of machine learning outputs. In this section, an overview is provided of different types of world models in interactive human-centric machine learning systems. These are the world models of people, organizations, and machine learning. Then, with AML as an example, the risks and ambiguities explained above are related to the embodied world models of individual people, to the documented world models of human organizations, and to machine learning world models. Subsequently, the need for shared interpretability across world models is related to the overall focus of ALMA Work Package 5: Learning world models, ethics and culture with AML.

6.1 World models in interactive human-centric machine learning systems

Human-artificial intelligence systems involve interactions between individual people and the machine learning implementations of human organizations. Although there are fundamental differences in the forms of individual people's, human organizations' and machine learning world models, all can provide the basis for perceptual, instrumental and epistemic inference. Perceptual inference refers to inferring

sensory stimuli from predictions based on internal representations in world models. Instrumental inference involves inferring action options and consequences in the environment. Epistemic inference refers to updating beliefs in world models. For brevity, cycles of perceptual, instrumental, and epistemic inference can be referred to as active inference [50], [106].

The embodied world models of individual people are psychomotor world models, which can be extremely difficult to define because they can involve dynamic non-conscious interactions between variables such as personality type and body memory. For example, gait is related to personality in ways that are not fully understood. This problem is exacerbated by the difficulty of defining where one personality type ends and another begins. Furthermore, gait is related to memory in ways that are not fully understood. This problem is exacerbated by the difficulty of defining what aspects of memory are in the mind and what aspects are in the body [107]. Furthermore, embodied world models can comprise a broad meta-model and many activity-specify models. Accordingly, definitions of embodied world models should encompass meta-model variables such as personality type, and activity-specific models variables such as past experience, which can combine positively or combine negatively [108].

The world models of human organizations can also be considered in terms of broad meta-models and activity-specific models. Strategic plan documents are an example of organizations' broad meta-models, and documented work procedures are examples of activity-specific models. Broad high level plans can be combined with activity-specific procedures within quality management systems (QMS), which can provide the basis for organizational active inference [109]. QMS can incorporate legal regulations that can provide the basis for machine learning implementations [110]. In order to achieve objectives set-out in strategic plan documents, human organizations' implementations of machine learning are narrow implementations that are intended to improve the performance of specific activities within particular situations [106]. Accordingly, machine learning implementations can be considered to be examples of extended cognition, which involves human organizations extending out some cognitive efforts to technologies [111].

World models for machine intelligence were described in the 1960s as follows: "A computer program capable of acting intelligently in the world must have a general representation of the world in terms of which its inputs are interpreted" [112]. More recently, computer scientists have considered that the scope of world models is selected concepts and relationships between them to represent the real system [113]. Since 2020, the world models construct has been related to deep learning and reinforcement learning [114]. Moreover, there has been increasing interest in a framing of the world models construct that is applicable to natural and artificial agents [50], [115]. However, there is no universally agreed computational basis for world models, and there are limitations with all current implementation efforts [116]. While debate has increased among neuroscientists and computer scientists about world models, there have been increasing examples of practical implementations that are separate from this scientific debate. Consider, for example, small autonomous delivery robots that are already being operated in urban environments in several countries [117][117]. These have a model of the world comprising selected concepts and relationships between them (such as sidewalks and pedestrians), which represent those aspects of the real-world systems that the robots interact with [118]. Furthermore, these autonomous delivery robots carry out perceptual, instrumental, and epistemic inference. In other words, they infer sensory stimuli from predictions based on the internal representations in their world models (e.g. the intermittent white lines in front of them indicate road crossings). They infer action options and consequences in the environment (e.g. the intermittent white lines can be followed to cross the road safely). They update beliefs in their world models (e.g. there are new intermittent white lines, which indicates that there is a new road crossing). This is just one example of what is referred to in science as perceptual, instrumental, and epistemic inference being carried out in practice in a wide variety of applications with a wide variety of computational methods [109].

6.2 Shared Interpretability of World Models: An AML example

Based on the preceding explanation, a machine learning *world model* can be described as any machine learning model that has two characteristics. First, it encompasses selected concepts and relationships between them to represent those aspects of the real-world system that it interacts with. Figure 10 below shows a summary of the steps involved in the formulation of a AML world model.

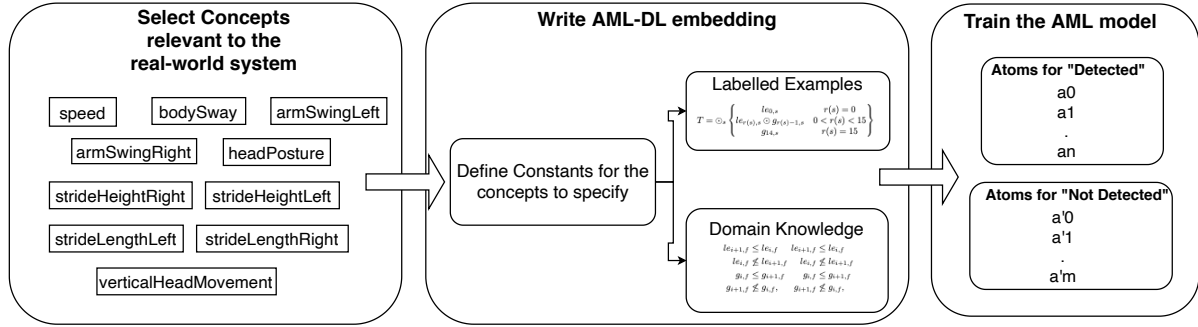


Figure 10: Steps involved in the formulation of a AML world model.

AML-models are mathematical models comprising algebraic representations. AML is founded upon several core algorithms that can learn from both human-defined constraints and data. As we can see in Figure 10, the selected concepts for the desired application are first described using the AML description language (AML-DL). This involves defining the relevant constants for the problem at hand. Constants define features such as different walking speeds. As shown in the second box of Figure 10, those constants will in turn be used to specify labelled training examples and to specify domain knowledge. This specification is called an embedding. In the previous deliverable D5.5, we defined our gait analysis embedding. This means we selected relevant features such as the person's speed and head posture, and represented those in AML-DL. In our case, the domain knowledge consists of the notion of ordered intensity (telling the algebra each feature is a number) while the training examples come from an existing real world dataset. For completeness, we will explain here again the embedding. The relevant 10 features selected are those from [103], which can be seen in the first box of Figure 10.

Walking speed: Participants' walking speed was measured according to their spine movement along the Z-axis.

Body swaying: Body swaying is measured by the maximum difference in position of the left-shoulder and the right-shoulder on the X-axis during a gait cycle.

Left-arm/Right-arm swing: It was defined as the maximum difference of the left/right wrist moving along the Z-axis during a gait cycle.

Head posture: Quantifies the head posture during a gait cycle by averaging the angle between the vertical direction and the connection line between the neck and the clavicle in the plane which consisted of the Y-axis and the Z-axis.

Left/Right stride length: It measured the maximum change in the horizontal direction of the left/right foot during a gait cycle.

Left/Right stride height: It measured the maximum change in the vertical direction of the left/right foot during a gait cycle.

Vertical head movement: maximum vertical amplitude of head along the Y-axis during a gait cycle.

For each subject, each feature is calculated using its mean during two gait cycles [103]. Each feature is a number, which is discretised using a simple approach where each feature will be uniformly divided into 16 bins. We define for each feature $30 = 2 \times (16 - 1)$ constants separated into two sets of 15 constants each, one set for “lower or equal” constants and another set for “greater than” constants. A total of $10 \times 30 = 300$ constants are used to describe the input and another 2 constants are used for the output, which represents the two groups of subjects (detected and not detected). For each feature f , the 30 constants associated to the feature are $le_{0f}, le_{1f}, \dots, le_{14f}$ and $g_{0f}, g_{1f}, \dots, g_{14f}$. The formal knowledge we want to express is the notion of intensity, which we will represent as ordered sets

$$le_{i+1f} \leq le_{if} \quad (1)$$

$$le_{if} \not\leq le_{i+1f} \quad (2)$$

$$g_{if} \leq g_{i+1f} \quad (3)$$

$$g_{i+1f} \not\leq g_{if}, \quad (4)$$

forming one ascending chain and one descending chain for each feature. This is the domain knowledge that can be seen in the the second box of Figure 10 and represents that each feature corresponds to an intensity, i.e a numerical value and, therefore, represented as an ordered set in a discrete system. The second part of the embedding is related to the training examples. Each example can be represented as a vector r corresponding to the $f = 0, \dots, 10$ discretized feature readings for a subject and one output value k representing the group the subject belongs to. For each training example, a term T was created representing the problem input:

$$T = \odot_s \left\{ \begin{array}{ll} le_{0f} & r(f) = 0 \\ le_{rf} \odot g_{r-1f} & 0 < r(f) < 15 \\ g_{14f} & r(f) = 15 \end{array} \right\} \quad (5)$$

and the training example was embedded with the positive relation

$$O_k \leq T \quad (6)$$

and 1 negative relation expressing that when some data belongs to one group he does not belong to the other,

$$O_k \not\leq T \quad (7)$$

This can also be seen in the second box of Figure 10, with its three components representing the complete theory we have used to embed the gait analysis problem. In other words, our embedding consists on how training data is labelled, **Eqs. 6** and **Eqs. 7**, and the formal knowledge relations **Eqs. 1**.

Once the embedding is defined, AML learns by creating atoms. AML-models have a fundamentally simple structure. They comprise only three layers: inputs – atoms – outputs. Relationships between inputs and outputs defined in atoms are binary. An atom can be linked to one or more inputs and is said to be present in an example if at least one of those inputs is present. AML-models expand as they learn patterns from training data. The number of atoms can increase from tens to thousands during training. Nonetheless, the simple three-layer structure persists. The core AML algorithms check that the human-defined constraints are maintained in the subsequent binary relationships between inputs and outputs that are learnt as new atoms are generated during training. It is clear from this description that AML fits the first criteria, as the embedding description and learned atoms represent selected concepts and relationships between them in the context of the real-world system in this gait analysis scenario.

Based on the description in Section 6.1, the second fundamental characteristic of a machine learning world model is that it can carry out perceptual, instrumental, and epistemic inference related to the real-world system that it interacts with. We can see how a trained AML model performs perceptual and instrumental inference by looking at Figure 11.

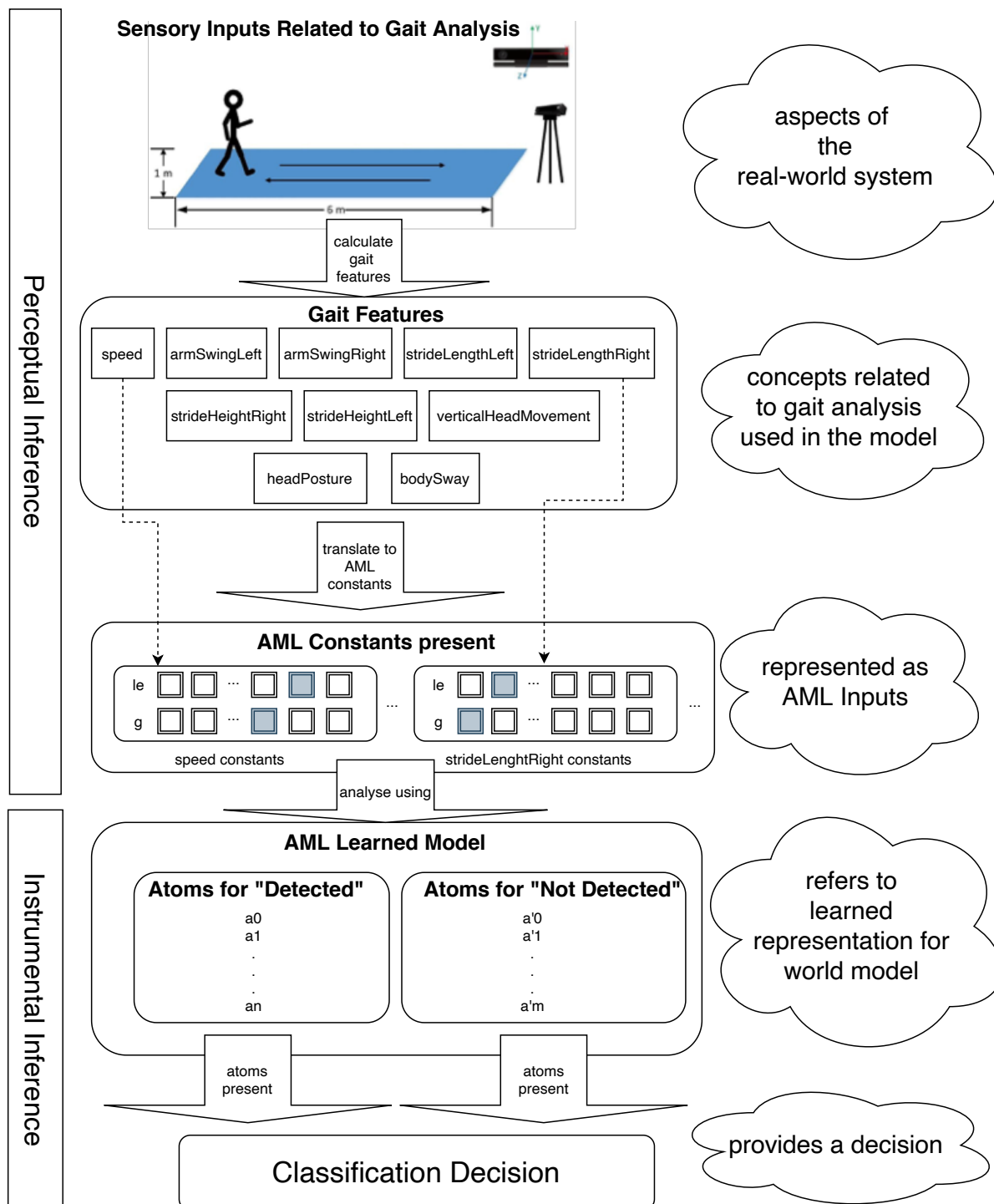


Figure 11: A trained AML model performs perceptual and instrumental inference.

In the case of gait analysis, the real-world system that AML interacts with at the input stage is numbers describing gait features. These numbers come from the conversion by specialist software of gait record-

ings that are made by video cameras and/or wearable sensors. When evaluating a gait example AML perceptual inference is in terms of constants present for that example while the instrumental inference of AML is in consulting its learned representation for its world model (atoms) and based on its analysis (the number of atoms present, i.e. not missing) making classification decisions.

Regarding epistemic inference, in this scenario it is present when updating the AML model to include new gait examples or even new knowledge about the gait analysis problem. The process is summarized in Figure 12.

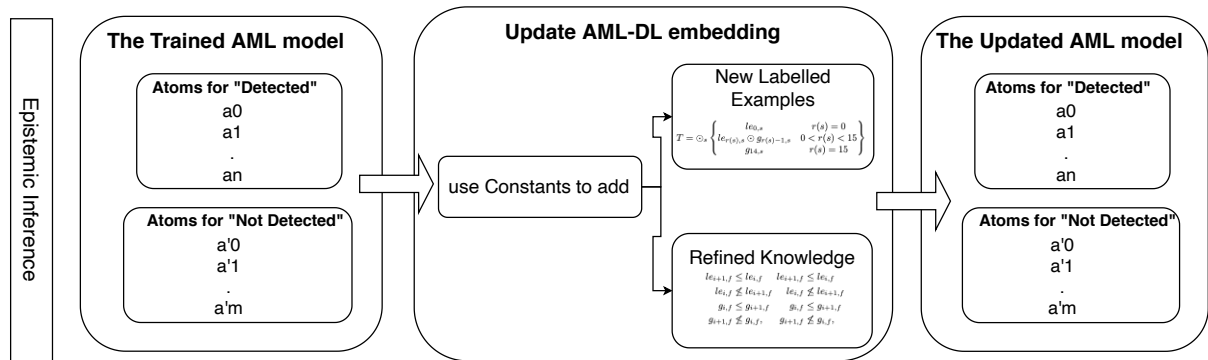


Figure 12: In this scenario AML can perform epistemic inference when updating the model through retraining. This can be done to include new gait examples or even new knowledge regarding the gait analysis problem.

Thus, we can see that AML can fulfill both criteria of a world model. First, as summarized in Figure 10, it encompasses selected concepts and relationships between them to represent those aspects of the real-world system that it interacts with. Second, as summarized in Figures 11 and 12, AML can perform perceptual inference, inferential inference and epistemic inference.

At the output stage, the real-world system that AML interacts with is the communication media of healthcare providers, such as their computer monitors. As explained in sections 4 and 5 above, AML-enabled analyses need to be converted into the clearest possible human-readable formats before their output to communication media. If it is not possible to reach shared interpretation of machine learning outputs, reference to world models could reveal sources of opposing interpretations. However, there would need to be shared interpretation of world models to enable agreement about sources of opposing interpretations of machine learning outputs. In this case, there can be opposing interpretations of AML-enabled gait analyses, which could be presented as on-screen displays or as on-paper reports. In particular, a healthcare provider's interpretation of an AML-enabled gait analyses report could be that the person whose gait has been analysed does not have a functional disorder. By contrast, the person whose gait has been analysed can interpret the same report as providing indications of functional disorder. As summarized in Figure 2 in section 2.3 of this deliverable, in accordance with the principle of least collaborative effort, representatives of a healthcare provider organization could be making more reference to its internal model than to the AML-enabled gait analysis report. At the same time, also in accordance with the principle of least collaborative effort, the person whose gait has been analysed could be making more reference to own internal model than to the AML-enabled gait analysis report.

In this case, it would be unlikely that reference to world models would reveal sources of opposing interpretations of the AML-enabled gait analyses report. There could be most potential for shared interpretation of the healthcare provider's documented world model as defined in its QMS documentation because it is written in natural language. However, as can be seen from Figures 3 and 5, which show atoms and constants in the AML model, the risks and ambiguities involved in efforts to achieve shared interpretation of

the AML output, the gait analyses report, are present in the AML model itself. In particular, as explained in detail in section 2 of this deliverable, there is the risk of cognitive overload for people who are not AML experts that try to understand the AML model, and there are conceptual and presentational ambiguities for non-AML experts. However, in this case, the fundamental challenge to shared interpretation of world models is not the AML model, or any other type of machine learning that could have been used. Rather, it is the fundamental challenge of reaching agreement about the embodied psychomotor world model [108] of the person whose gait has been analysed. For example, research indicates that patients often perceive functional disorders to be primarily physical. By contrast, healthcare providers can perceive patients' functional disorders to be primarily psychological [119].

6.3 Shared interpretability for culture and ethics in AML world models

As stated in the project summary, the ALMA project is concerned with interactive human-centric machine learning systems. In this deliverable, research and development work is reported that has been carried out in ALMA Work Package 5: Learning world models, ethics and culture with AML. In particular, Task 5.5 Representing basic ethical ideas in AML, and Task 5.6 Defining world models in terms of ecosystems' cultures and ethics. There are three types of world models in interactive human-centric machine learning systems: the world models of individual people, of machine learning, and of human organizations. The focus of this deliverable is shared interpretation across world models because shared interpretation is a fundamental characteristic of culture and ethics. For example, a recent study reports that cultural integration in the European Union continues to be based on shared values, but draws attention to the potential for opposing interpretations of values such as democracy [120]. In particular, "even though democracy is almost universally accepted as a preferred form of government, people in autocratic states exhibit a so-called 'authoritarian notion of democracy,' which conflicts with the principles of liberal democracy". This is an example of potential for opposing interpretations to increase when going from abstract information about values to specific information about applications of values in practice. Accordingly, in the work reported in this deliverable is focused on specific information about practical application of an important EU shared value: universal healthcare [121]. The work reported here focused on the challenge to universal healthcare brought by the economic costs arising from the increasing prevalence of functional disorders. Notably, high economic costs can arise from delayed diagnosis of functional motor disorders [122].

As described in section 3 of this deliverable, the focus on shared interpretability goes beyond the state-of-the-art in research concerned with explainability, transparency, and interpretability, which although necessary are not sufficient for shared interpretation of information related to culture and ethics. In this deliverable, complex linkages between culture and ethics are illustrated with the specific case of AML-enabled gait analyses. In particular, there is an ethical requirement to avoid conflicting interpretations of analyses related to functional disorders, including gait disorders. This is because different interpretations can lead to high treatment costs with poor treatment outcomes [11]. For example, functional disorder patients who do not agree with a diagnosis may be less likely to respond to treatment. This ethical requirement is aligned with the shared EU cultural value of universal healthcare in two ways. First, people in the EU should receive timely diagnoses in order to maximize good health in the EU population. Second, as the proportion of retirement age people in the EU population increases, providing universal healthcare depends upon careful management of the medical costs of maximizing good health.

In addition, the research and development work reported in this deliverable illustrates that the incorporation of culture and ethics into machine learning world models in general, and AML world models in particular, is a computer engineering challenge. This is because implementations take place at the end of the typical engineering progression from vagueness to precision as follows: debate about social choices – legislation of new laws – definition of regulations in practice standards – organizational documentation of work procedures – data from work procedures – machine learning implementations. For example, there can be culturally-bound social choices to accept that functional disorders are an important and increas-

ingly widespread medical phenomenon. Then, related normative ethics can be set-out in laws related to healthcare, which provide regulatory frameworks for human organizations' operating procedures: in this case healthcare providers' operating procedures for carrying out diagnoses of functional disorders. For many organizations, operating procedures are defined within their quality management systems, which are developed and audited in accordance with international standards such as ISO 9000. AML, like other computational methods, will be implemented at the end of such progressions from vagueness to specificity.

However, the research and development work reported in this deliverable also illustrates that there can be profound challenges to address before opposing participants in interactive human-centric machine learning systems can agree shared interpretation of machine learning models and their outputs. These profound challenges can arise from human embodied psychomotor functioning. In this case, there is the profound challenge that functional disorders have medically unexplained physical symptoms [123]. Thus, even the most advanced computer engineering of machine learning world models for medical diagnoses cannot easily enable shared understanding of something that cannot yet be explained by medical science. Furthermore, it is important to note that there are decades of evidence that organizations and individuals can continue to have opposing interpretations of exactly the same information even if it is explained in detail with presentation of high quality visual content [124]. Hence, as illustrated by the work reported in this deliverable, the incorporation of culture and ethics into machine learning world models needs to be carried out with consideration of human organizations' world models, individual people's world models, and the potential for deeply-rooted opposition between them [125]. Thus far, there is no evidence that AML will be better or worse than other types of machine learning for incorporating culture and ethics into world models. Rather, this is an open challenge for machine learning, which provides the the biggest challenge for successful application of AML.

7 Conclusions

As stated in the project summary, the ALMA project is concerned with interactive human-centric machine learning systems. There are three types of world models in interactive human-centric machine learning systems: the world models of individual people, of machine learning, and of human organizations. In this document, Deliverable (D) 5.8 - AML based world models - Final, we have reported findings from our continued research into the potential of Algebraic Machine Learning (AML) enable world models that encompass culture and ethics. This is the third and concluding deliverable in the AML based world models series: Initial (D5.4), Revised (D5.5), and Final (D5.8). In D5.4, we related AML-enabled analyses to interfaces between individuals' embodied models of themselves in the world and organizations' documented models of themselves in the world (i.e. world model interfaces). In D5.5, we considered the interpretability in world model interfaces of outputs from AML-enabled analyses. Our research reported in D5.5 indicated that AML has some potential to improve the interpretability of machine learning, but that AML would need to be developed further to bring about realization of that potential.

Accordingly, here we have reported our research and development work to improve the interpretability of AML. In particular, we have evaluated AML-enabled gait analyses in terms of important constructs related to explainable artificial intelligence (xAI): explainability, transparency, and interpretability. Furthermore, we have evaluated in terms of shared interpretability of AML-enabled analyses outputs.

Shared interpretation is important because different models of self in the world can lead to different people making different, and even opposing, interpretations of the same information. Apropos, there is an ethical requirement to avoid conflicting interpretations of analyses related to functional disorders, including gait disorders. This is because different interpretations can lead to high treatment costs with poor treatment outcomes. For example, functional disorder patients who do not agree with a diagnosis are less likely to respond to treatment.

In addition to evaluating AML in terms of important xAI constructs, we have developed AML application in order to improve its potential to facilitate shared interpretation. This development work has focused on maintaining the accuracy of outputs from AML-enabled analyses while reducing the complexity of their presentation. Development work has been informed by reference to hermeneutics, which is the study of interpretation processes.

Results of development work have also been evaluated in terms of explainability, transparency, interpretability, and shared interpretability. Evaluation indicates improvement, but highlights that barriers to shared interpretability are deeply-rooted in human embodied psychomotor functioning and will remain challenging.

Futhermore, we have related the challenges of shared interpretability to individuals' embodied psychomotor models of themselves in the world, organizations' documented models of themselves in the world, and machine learning world models with an AML example. Seeking to achieve shared interpretability of world models is important when it is not possible to reach shared interpretation of machine learning outputs, and reference to world models could reveal sources of opposing interpretations. Thus far, there is no evidence that AML will be better or worse than other types of machine learning for incorporating culture and ethics into world models. Rather, this is an open challenge for machine learning, which provides the the biggest challenge for successful application of AML in interactive human-centric machine learning systems.

The work reported in the AML based world models series of deliverables, Initial (D5.4), Revised (D5.5), and Final (D5.8), was carried out in Task 5.5 Representing basic ethical ideas in AML, and Task 5.6 Defining world models in terms of ecosystems' cultures and ethics. The titles of these two Tasks can be considered as research questions. First, how can basic ethical ideas be represented in AML. Figure 13, below, Representing basic ethical ideas in AML, provides a summary of our findings. In particular, basic ethical ideas can be represented in AML at the end of a progression from vague general statements to precise application-specific representations. This is because implementations take place at the end of the typical engineering progression from vagueness to precision as follows: debate about social choices – legislation of new laws – definition of regulations in practice standards – organizational documentation of work procedures – data from work procedures – machine learning implementations.

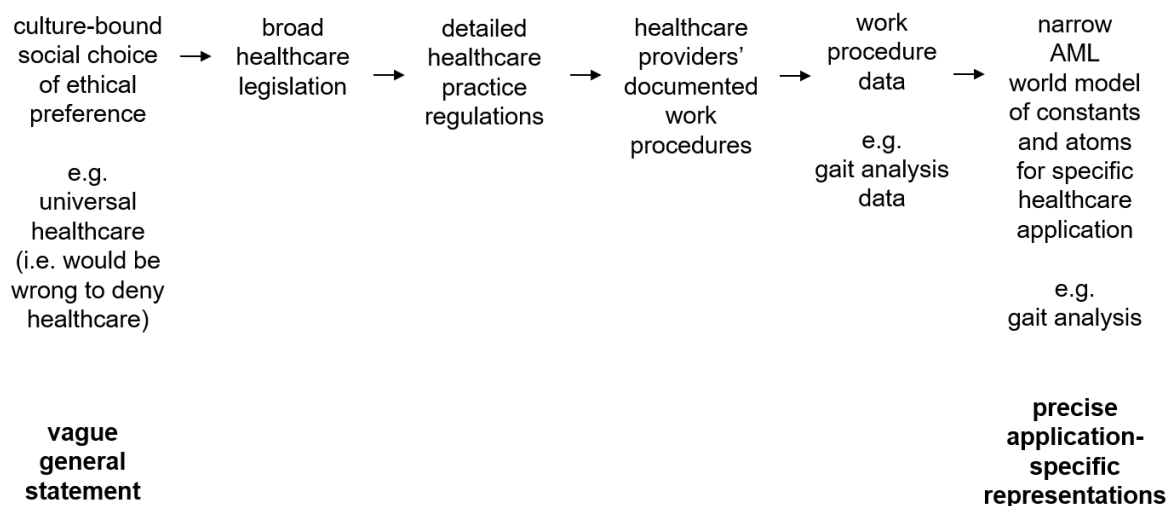


Figure 13: Representing basic ethical ideas in AML

Also, the title of Task 5.6 can be considered as a research question: how can world models be defined in terms of ecosystems' cultures and ethics. Figure 14 below, Definition of world models in terms of ecosystems cultures and ethics, provides a summary of our findings. In particular, Figure 14 shows that interactive human-centric machine learning systems involve interactions between the world models of human organizations, their machine learning implementations, and individual people.

Figure 14 shows that in an ecosystem, such as an healthcare ecosystem, some people can interpret decisions made with the assistance of machine learning to be ethical decisions but some people can interpret them to be unethical decisions. For example, in this case, there can be individual people who believe that they are suffering from a functional disorder, but machine learning enabled analysis indicates a low probability of functional disorder. Hence, they do not receive the healthcare that they believe they are entitled to receive within universal healthcare.

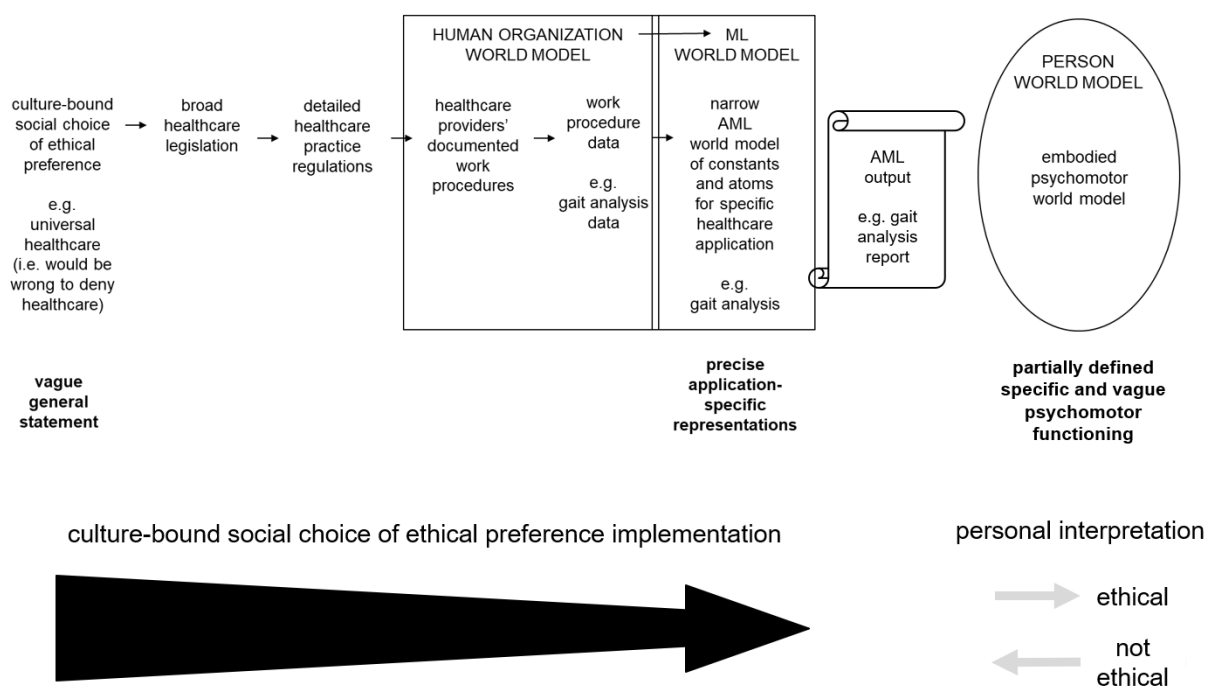


Figure 14: Definition of world models in terms of ecosystems cultures and ethics

Following this deliverable, research in Work Package 5 will deliver D5.6 Embedding of ethics and cultural concerns into AML based world models - Initial, and D5.7 Embedding of ethics and cultural concerns into AML based world models - Final. In those future deliverables, we will report our further research into the potential of Algebraic Machine Learning (AML) to improve interactive human-centric machine learning systems in complex settings.

References

- [1] Donghee Shin. “The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI”. In: *International Journal of Human-Computer Studies* 146 (2021), p. 102551.
- [2] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.
- [3] Adán Cabello. “Interpretations of quantum theory: A map of madness”. In: *What is quantum information* (2017), pp. 138–143.
- [4] Roscoe Pound. *Interpretations of legal history*. Cambridge University Press, 1967.
- [5] Judith Balmaña et al. “Conflicting interpretation of genetic variants and cancer risk by commercial laboratories as assessed by the prospective registry of multiplex testing”. In: *Journal of Clinical Oncology* 34.34 (2016), p. 4071.
- [6] Laura R Novick and Kefyn M Catley. “When relationships depicted diagrammatically conflict with prior knowledge: An investigation of students’ interpretations of evolutionary trees”. In: *Science Education* 98.2 (2014), pp. 269–304.
- [7] Karen O’Brien et al. “What’s in a word? Conflicting interpretations of vulnerability in climate change research”. In: *CICERO Working Paper* (2004).
- [8] Arul Mishra and Himanshu Mishra. “Border bias: The belief that state borders can protect against disasters”. In: *Psychological science* 21.11 (2010), pp. 1582–1586.
- [9] Thomas Metzinger. “Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples”. In: *Progress in brain research* 168 (2007), pp. 215–278.
- [10] Jaap Van den Heuvel et al. “An ISO 9001 quality management system in a hospital: bureaucracy or just benefits?” In: *International Journal of Health Care Quality Assurance* (2005).
- [11] Katherine E MacDuffie et al. “Stigma and functional neurological disorder: a research agenda targeting the clinical encounter”. In: *CNS spectrums* 26.6 (2021), pp. 587–592.
- [12] Sarah C Lidstone, Lindsey MacGillivray, and Anthony E Lang. “Integrated therapy for functional movement disorders: time for a change”. In: *Movement disorders clinical practice* 7.2 (2020), p. 169.
- [13] Josef Bleicher. *Contemporary hermeneutics: Hermeneutics as method, philosophy and critique*. Vol. 2. Routledge, 2017.
- [14] István M Fehér. “Hermeneutics and the Sciences”. In: *Hermeneutics and Science*. Springer, 1999, pp. 1–12.
- [15] Michael D Myers. “Dialectical hermeneutics: a theoretical framework for the implementation of information systems”. In: *Information systems journal* 5.1 (1995), pp. 51–70.
- [16] Karl J Friston and Christopher D Frith. “Active inference, communication and hermeneutics”. In: *cortex* 68 (2015), pp. 129–143.
- [17] Karl Friston and Christopher Frith. “A duet for one”. In: *Consciousness and cognition* 36 (2015), pp. 390–405.
- [18] Shaun Gallagher and Micah Allen. “Active inference, enactivism and the hermeneutics of social cognition”. In: *Synthese* 195.6 (2018), pp. 2627–2648.
- [19] Hendry F Chame, Ahmadreza Ahmadi, and Jun Tani. “A hybrid human-neurorobotics approach to primary intersubjectivity via active inference”. In: *Frontiers in psychology* 11 (2020), p. 584869.
- [20] Francesco Donnarumma, Haris Dindo, and Giovanni Pezzulo. “Sensorimotor communication for humans and robots: improving interactive skills by sending coordination signals”. In: *IEEE Transactions on Cognitive and Developmental Systems* 10.4 (2017), pp. 903–917.
- [21] Jiaxuan Wang et al. “Learning credible models”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 2417–2426.
- [22] Murat Dikmen and Catherine Burns. “The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending”. In: *International Journal of Human-Computer Studies* 162 (2022), p. 102792.

- [23] Jinglu Jiang, Surinder Kahai, and Ming Yang. “Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty”. In: *International Journal of Human-Computer Studies* 165 (2022), p. 102839.
- [24] Jobst Landgrebe. “Certifiable ai”. In: *Applied Sciences* 12.3 (2022), p. 1050.
- [25] Jens Zimmermann. *Hermeneutics: A very short introduction*. OUP Oxford, 2015.
- [26] Klaus Bruhn Jensen. “The double hermeneutics of communication research”. In: *Javnost-The Public* 25.1-2 (2018), pp. 177–183.
- [27] Jared Vasil et al. “A world unto itself: human communication as active inference”. In: *Frontiers in psychology* 11 (2020), p. 417.
- [28] Brian Skyrms. “The stag hunt”. In: *Proceedings and Addresses of the American Philosophical Association*. Vol. 75. 2. JSTOR. 2001, pp. 31–41.
- [29] Axel Constant, Andy Clark, and Karl J Friston. “Representation wars: Enacting an armistice through active inference”. In: *Frontiers in Psychology* 11 (2021), p. 598733.
- [30] Bart Nooteboom. “Learning by interaction: absorptive capacity, cognitive distance and governance”. In: *Journal of management and governance* 4.1 (2000), pp. 69–92.
- [31] Roger C Conant and W Ross Ashby. “Every good regulator of a system must be a model of that system”. In: *International journal of systems science* 1.2 (1970), pp. 89–97.
- [32] Jelle Bruineberg et al. “Free-energy minimization in joint agent-environment systems: A niche construction perspective”. In: *Journal of theoretical biology* 455 (2018), pp. 161–178.
- [33] Ryan Smith, Karl J Friston, and Christopher J Whyte. “A step-by-step tutorial on active inference and its application to empirical data”. In: *Journal of mathematical psychology* 107 (2022), p. 102632.
- [34] Gaia Tavoni, Vijay Balasubramanian, and Joshua I Gold. “What is optimal in optimal inference?” In: *Current opinion in behavioral sciences* 29 (2019), pp. 117–126.
- [35] Ryan Smith et al. “An active inference approach to modeling structure learning: Concept learning as an example case”. In: *Frontiers in computational neuroscience* 14 (2020), p. 41.
- [36] Stephen Fox. “Synchronous generative development amidst situated entropy”. In: *Entropy* 24.1 (2022), p. 89.
- [37] Arjen Stolk et al. “Cerebral coherence between communicators marks the emergence of meaning”. In: *Proceedings of the National Academy of Sciences* 111.51 (2014), pp. 18183–18188.
- [38] Arjen Stolk, Lennart Verhagen, and Ivan Toni. “Conceptual alignment: How brains achieve mutual understanding”. In: *Trends in cognitive sciences* 20.3 (2016), pp. 180–191.
- [39] Herbert H Clark and Deanna Wilkes-Gibbs. “Referring as a collaborative process”. In: *Cognition* 22.1 (1986), pp. 1–39.
- [40] Michael Tomasello et al. “Joint attention as social cognition”. In: *Joint attention: Its origins and role in development* 103130 (1995), pp. 103–130.
- [41] Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*. Vol. 142. Citeseer, 1986.
- [42] Joanna Szufnarowska et al. “Is ostension any more than attention?” In: *Scientific Reports* 4.1 (2014), pp. 1–4.
- [43] Herbert H Clark and Susan E Brennan. “Grounding in communication.” In: (1991).
- [44] Michael Tomasello. *A natural history of human thinking*. Harvard University Press, 2014.
- [45] Michael Tomasello. *Origins of Human Communication*. Harvard University Press, 2008.
- [46] Jasmeen Kanwal et al. “Zipf’s law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication”. In: *Cognition* 165 (2017), pp. 45–52.
- [47] Noah D Goodman and Michael C Frank. “Pragmatic language interpretation as probabilistic inference”. In: *Trends in cognitive sciences* 20.11 (2016), pp. 818–829.
- [48] Bethan L Davies. “Least Collaborative Effort or Least Individual Effort: Examining the Evidence”. In: *Leeds Working Papers in Linguistics and Phonetics* 12 (2007), pp. 1–20.
- [49] Christopher J Whyte and Ryan Smith. “The predictive global neuronal workspace: A formal active inference model of visual consciousness”. In: *Progress in neurobiology* 199 (2021), p. 101918.

- [50] Karl Friston et al. “World model learning and inference”. In: *Neural Networks* 144 (2021), pp. 573–590.
- [51] Yukie Nagai. “Predictive learning: its key role in early cognitive development”. In: *Philosophical Transactions of the Royal Society B* 374.1771 (2019), p. 20180030.
- [52] Amid Ayobi et al. “Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning”. In: (2021).
- [53] Susan Leigh Star and James R Griesemer. “Institutional ecology, translations’ and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907–39”. In: *Social studies of science* 19.3 (1989), pp. 387–420.
- [54] W Brian Arthur. “Competing technologies, increasing returns, and lock-in by historical events”. In: *The economic journal* 99.394 (1989), pp. 116–131.
- [55] Georg Schreyögg, Jörg Sydow, and Philip Holtmann. “How history matters in organisations: The case of path dependence”. In: *Management & organizational history* 6.1 (2011), pp. 81–100.
- [56] Matthew S Nurse and Will J Grant. “I’ll see it when I believe it: Motivated numeracy in perceptions of climate change risk”. In: *Environmental Communication* 14.2 (2020), pp. 184–201.
- [57] David Dunning and Emily Balcetis. “Wishful seeing: How preferences shape visual perception”. In: *Current directions in psychological science* 22.1 (2013), pp. 33–37.
- [58] Philip E Tetlock. “Theory-driven reasoning about plausible pasts and probable futures in world politics: are we prisoners of our preconceptions?” In: *American Journal of Political Science* (1999), pp. 335–366.
- [59] Rüdiger Pohl and Rüdiger F Pohl. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, 2004.
- [60] Roxy Peck and Jay L Devore. *Statistics: The exploration & analysis of data*. Cengage Learning, 2011.
- [61] Sharon Schwartz and Kenneth M Carpenter. “The right answer for the wrong question: consequences of type III error for public health research.” In: *American journal of public health* 89.8 (1999), pp. 1175–1180.
- [62] Barbara Sheehan et al. “Cognitive analysis of decision support for antibiotic prescribing at the point of ordering in a neonatal intensive care unit”. In: *AMIA annual symposium proceedings*. Vol. 2009. American Medical Informatics Association. 2009, p. 584.
- [63] Stephen Fox. “Relevance: A framework to address preconceptions that limit perceptions of what is relevant”. In: *International Journal of Managing Projects in Business* (2015).
- [64] John Sweller. “Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture”. In: *Instructional science* 32.1 (2004), pp. 9–31.
- [65] John Sweller and Susan Sweller. “Natural information processing systems”. In: *Evolutionary Psychology* 4.1 (2006), p. 147470490600400135.
- [66] Stephen Fox. “Behavioral Ethics Ecologies of Human-Artificial Intelligence Systems”. In: *Behavioral Sciences* 12.4 (2022), p. 103.
- [67] David A Raichlen and Gene E Alexander. “Adaptive capacity: an evolutionary neuroscience model linking exercise, cognition, and brain health”. In: *Trends in neurosciences* 40.7 (2017), pp. 408–421.
- [68] David Attwell and Simon B Laughlin. “An energy budget for signaling in the grey matter of the brain”. In: *Journal of Cerebral Blood Flow & Metabolism* 21.10 (2001), pp. 1133–1145.
- [69] Ed Bullmore and Olaf Sporns. “The economy of brain network organization”. In: *Nature reviews neuroscience* 13.5 (2012), pp. 336–349.
- [70] Jeremy E Niven and Simon B Laughlin. “Energy limitation as a selective pressure on the evolution of sensory systems”. In: *Journal of Experimental Biology* 211.11 (2008), pp. 1792–1804.
- [71] Biswa Sengupta, Martin B Stemmler, and Karl J Friston. “Information and efficiency in the nervous system—a synthesis”. In: *PLoS computational biology* 9.7 (2013), e1003157.
- [72] Paul Chandler and John Sweller. “Cognitive load theory and the format of instruction”. In: *Cognition and instruction* 8.4 (1991), pp. 293–332.

- [73] Fred Paas, Alexander Renkl, and John Sweller. "Cognitive load theory and instructional design: Recent developments". In: *Educational psychologist* 38.1 (2003), pp. 1–4.
- [74] Vahan Yoghourdjian et al. "Scalability of network visualisation from a cognitive load perspective". In: *IEEE transactions on visualization and computer graphics* 27.2 (2020), pp. 1677–1687.
- [75] Fred Paas, Alexander Renkl, and John Sweller. "Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture". In: *Instructional science* 32.1/2 (2004), pp. 1–8.
- [76] Beth Crandall, Gary A Klein, and Robert R Ho man. *Working minds: A practitioner's guide to cognitive task analysis*. Mit Press, 2006.
- [77] Yoed N Kenett et al. "The semantic distance task: Quantifying semantic distance with semantic network path length." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43.9 (2017), p. 1470.
- [78] Adam Linson, Thomas Parr, and Karl J Friston. "Active inference, stressors, and psychological trauma: A neuroethological model of (mal) adaptive explore-exploit dynamics in ecological context". In: *Behavioural brain research* 380 (2020), p. 112421.
- [79] Stephen Fox. "Getting real about ICT: Applying critical realism to the framing of information and communication technologies". In: *Management Research Review* 36.3 (2013), pp. 296–319.
- [80] Marco Yzer, Susan LoRusso, and Rebekah H Nagler. "On the conceptual ambiguity surrounding perceived message effectiveness". In: *Health communication* 30.2 (2015), pp. 125–134.
- [81] Francis Creed et al. *Is there a better term than "medically unexplained symptoms"?* 2010.
- [82] Narender K Ramarapu et al. "The emergence of hypertext and problem solving: An experimental investigation of accessing and using information from linear versus nonlinear systems". In: *Decision Sciences* 28.4 (1997), pp. 825–849.
- [83] Debi Roberson et al. "The Development of Color Categories in Two languages: a longitudinal study." In: *Journal of Experimental Psychology: General* 133.4 (2004), p. 554.
- [84] Julie Lemarié, Hélène Eyrolle, and Jean-Marie Cellier. "The segmented presentation of visually structured texts: Effects on text comprehension". In: *Computers in human behavior* 24.3 (2008), pp. 888–902.
- [85] Kathryn L Baker, Alexander M Franz, and Pamela W Jordan. "Coping with ambiguity in knowledge-based natural language analysis". In: *Proceedings of FLAIRS-94* (1994).
- [86] Bryan S Scha er and Christine M Riordan. "A review of cross-cultural methodologies for organizational research: A best-practices approach". In: *Organizational research methods* 6.2 (2003), pp. 169–215.
- [87] Ram Frost, Laurie B Feldman, and Leonard Katz. "Phonological ambiguity and lexical ambiguity: Effects on visual and auditory word recognition." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16.4 (1990), p. 569.
- [88] Monika Bullinger et al. "Translating health status questionnaires and evaluating their quality: the IQOLA project approach". In: *Journal of clinical epidemiology* 51.11 (1998), pp. 913–923.
- [89] Diane Wild et al. "Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation". In: *Value in health* 8.2 (2005), pp. 94–104.
- [90] Margarita Filippova and V Allakhverdov. "Concretization of the Chosen Meaning in the Process of Ambiguous Figures Perception". In: *Psychology. Journal of Higher School of Economics* 17.2 (2020), pp. 355–366.
- [91] Forrester Cole et al. "How well do line drawings depict shape?" In: *ACM SIGGRAPH 2009 papers*. 2009, pp. 1–9.
- [92] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.
- [93] Jichen Zhu et al. "Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation". In: *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE. 2018, pp. 1–8.

- [94] Andreas Holzinger et al. "Causability and explainability of artificial intelligence in medicine". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4 (2019), e1312.
- [95] Nathanael Stein. "Causation and explanation in Aristotle". In: *Philosophy Compass* 6.10 (2011), pp. 699–707.
- [96] Peter Lipton. "Causation and explanation". In: (2009).
- [97] Vijay Arya et al. "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques". In: *arXiv preprint arXiv:1909.03012* (2019).
- [98] Sanjoy Dey et al. "Human-centered explainability for life sciences, healthcare, and medical informatics". In: *Patterns* 3.5 (2022), p. 100493.
- [99] Victoria Alonso and Paloma De La Puente. "System transparency in shared autonomy: A mini review". In: *Frontiers in neurorobotics* 12 (2018), p. 83.
- [100] Benedict Du Boulay, Tim O'shea, and John Monk. "The black box inside the glass box: Presenting computing concepts to novices". In: *International Journal of Human-Computer Studies* 51.2 (1999), pp. 265–277.
- [101] Arun Rai. "Explainable AI: From black box to glass box". In: *Journal of the Academy of Marketing Science* 48.1 (2020), pp. 137–141.
- [102] Davide Castelvocchi. "Can we open the black box of AI?" In: *Nature News* 538.7623 (2016), p. 20.
- [103] Yameng Wang et al. "Detecting depression through gait data: examining the contribution of gait features in recognizing depression". In: *Frontiers in psychiatry* 12 (2021), p. 661213.
- [104] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. "Interpretable machine learning—a brief history, state-of-the-art and challenges". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2020, pp. 417–431.
- [105] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods". In: *Entropy* 23.1 (2020), p. 18.
- [106] Stephen Fox. "Human–Artificial Intelligence Systems: How Human Survival First Principles Influence Machine Learning World Models". In: *Systems* 10.6 (2022), p. 260.
- [107] Stephen Fox. "Practical implications from distinguishing between Pearl blankets and Friston blankets." In: *Behavioral & Brain Sciences* 45 (2022).
- [108] Stephen Fox. "Psychomotor predictive processing". In: *Entropy* 23.7 (2021), p. 806.
- [109] Stephen Fox. "Active inference: Applicability to different types of social organization explained through reference to industrial engineering and quality management". In: *Entropy* 23.2 (2021), p. 198.
- [110] Alan Winfield. "Ethical standards in robotics and AI". In: *Nature Electronics* 2.2 (2019), pp. 46–48.
- [111] Stephen M Fiore and Travis J Wiltshire. "Technology as teammate: Examining the role of external cognition in support of team cognitive processes". In: *Frontiers in psychology* 7 (2016), p. 1531.
- [112] John McCarthy and Patrick J Hayes. "Some philosophical problems from the standpoint of artificial intelligence". In: *Readings in artificial intelligence*. Elsevier, 1981, pp. 431–450.
- [113] David Ha and Jürgen Schmidhuber. "World models". In: *arXiv preprint arXiv:1803.10122* (2018).
- [114] Yutaka Matsuo et al. "Deep learning, reinforcement learning, and world models". In: *Neural Networks* 152 (2022), pp. 267–275.
- [115] Pietro Mazzaglia et al. "The free energy principle for perception and action: A deep learning perspective". In: *Entropy* 24.2 (2022), p. 301.
- [116] Tadahiro Taniguchi et al. "World models and predictive coding for cognitive and developmental robotics: frontiers and challenges". In: *Advanced Robotics* (2023), pp. 1–27.
- [117] Dylan Jennings and Miguel Figliozzi. "Study of sidewalk autonomous delivery robots and their potential impacts on freight efficiency and travel". In: *Transportation Research Record* 2673.6 (2019), pp. 317–326.
- [118] Luke Dormehl. *Inside the mind of an autonomous delivery robot*. Feb. 2020. UR : <https://www.digitaltrends.com/cool-tech/how-starship-robots-navigate-world/>.

- [119] Elspeth Carruthers et al. “Content analysis of views on the nature of functional neurological disorders”. In: *Journal of Psychosomatic Research* 140 (2021), p. 110303.
- [120] Plamen Akaliyski, Christian Welzel, and Josef Hien. “A community of shared values? Dimensions and dynamics of cultural integration in the European Union”. In: *Journal of European Integration* 44.4 (2022), pp. 569–590.
- [121] Erin Webb, Johanna O’Leary, Ewout van Ginneken, et al. “Universal Health Coverage in the EU: what do we know (and not know) about gaps in access?” In: *Eurohealth* 28.3 (2022).
- [122] Michele Tinazzi et al. “Economic costs of delayed diagnosis of functional motor disorders: preliminary results from a cohort of patients of a specialized clinic”. In: *Frontiers in Neurology* 12 (2021), p. 786126.
- [123] Jonathan R Price and David Okai. “Functional disorders and ‘medically unexplained physical symptoms’”. In: *Medicine* 44.12 (2016), pp. 706–710.
- [124] Richard M Perloff. “A three-decade retrospective on the hostile media effect”. In: *Advances in Foundational Mass Communication Theories*. Routledge, 2018, pp. 196–224.
- [125] Cory J Clark and Bo M Winegard. “Tribalism in war and peace: The nature and evolution of ideological epistemology and its significance for modern social science”. In: *Psychological Inquiry* 31.1 (2020), pp. 1–22.