machine learning & knowledge extraction

Article

Representing Human Ethical Requirements in Hybrid Machine Learning Models: Technical Opportunities and Fundamental Challenges

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

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

Abstract: Hybrid machine learning encompasses predefinition of rules and ongoing learning from data. Human organizations can implement hybrid machine learning (HML) to automate some of their operations. Human organizations need to ensure that their HML implementations are aligned with human ethical requirements as defined in laws, regulations, standards, etc. The purpose of the study reported here was to investigate technical opportunities for representing human ethical requirements in HML. The study sought to represent two types of human ethical requirements in HML: locally simple and locally complex. The locally simple case is road traffic regulations. This can be considered to be a relatively simple case because human ethical requirements for road safety, such as stopping at red traffic lights, are defined clearly and have limited scope for personal interpretation. The locally complex case is diagnosis procedures for functional disorders, which can include medically unexplained symptoms. This case can be considered to be locally complex because human ethical requirements for functional disorder healthcare are less well defined and are more subject to personal interpretation. Representations were made in a type of HML called Algebraic Machine Learning. Our findings indicate that there are technical opportunities to represent human ethical requirements in HML because of its combination of human-defined top down rules and bottom up data-driven learning. However, our findings also indicate that there are limitations to representing human ethical requirements: irrespective of what type of machine learning is used. These limitations arise from fundamental challenges in defining complex ethical requirements, and from potential for opposing interpretations of their implementation. Furthermore, locally simple ethical requirements can contribute to wider ethical complexity.

Keywords: algebraic machine learning; artificial intelligence; functional disorders; human ethical requirements; hybrid machine learning; psychomotor; road traffic regulations; world models

check for updates

Citation: Fox, S.; Rey, V.F. Representing Human Ethical Requirements in Hybrid Machine Learning Models: Technical Opportunities and Fundamental Challenges. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 580–592. https://doi.org/10.3390/make6010027

Received: 11 December 2023
Revised: 16 February 2024
Accepted: 28 February 2024
Published: 8 March 2024

1. Introduction

Hybrid machine learning can be considered as an example of a third wave of artificial intelligence (AI). In the first wave, deterministic "hand-crafted" rule-based expert systems were introduced. Such systems may be referred to as Good Old Fashioned AI (GOFAI). They are of limited usefulness amidst the complexity of real-world dynamics, because of the difficulty of identifying each possible case in advance for an expert system and programming each case as hard-coded behavior in rule-based implementations. The machine learning of the second AI wave is a probabilistic data-driven approach, which has brought some advances compared to GOFAI. Yet, ML can have limited potential to transfer what it has learned from training data to situations that differ from the training data. Also, ML implementations can rely on goals needing to be expressed as a single numerical value. Consequently, there is some interest in what can be described as a third wave of AI, which involves combining deterministic and probabilistic approaches in hybrid machine learning [1,2]. One hybrid approach is Algebraic Machine Learning (AML), which allows for the explicit embedding of rules directly and uses multiple data inputs [3,4]. It is not the purpose of the paper to compare AML and other types of hybrid ML. Rather, AML is an example of hybrid machine learning, which is used here to illustrate technical opportunities and fundamental challenges.

The purpose of the study reported here was to investigate technical opportunities for representing human ethical requirements in AML. The study sought to represent two types of human ethical requirements in AML: locally simple and locally complex. Here, local refers to the place and time of interaction between people and human ethical requirements as defined in regulations and implemented by technology. The locally simple case is road traffic regulations. This can be considered to be a relatively simple case because human ethical requirements for road safety, such as stopping at red traffic lights, are defined clearly and have limited scope for personal interpretation. The locally complex case is diagnosis procedures for functional disorders, which can include medically unexplained symptoms. This can be considered to be locally complex because human ethical requirements for functional disorder healthcare are less well defined and are subject to personal interpretation. Findings are reported in the remaining three sections of the paper. Next, in Section 2, the representation in AML of human ethical requirements for road safety is reported. In addition, it is explained how locally simple ethical requirements can contribute to wider ethical complexity. Then, in Section 3, the representation in AML of human ethical requirements in diagnosis procedures for functional disorders is reported. In conclusion, in Section 4, practical implications are discussed and directions for further research are proposed.

Overall, the paper makes four contributions to debate about how human ethical requirements can be represented in machine learning. First, technical opportunities for representation of human ethical requirements in a hybrid machine learning are described in detail. Second, fundamental limitations to representation of human ethical requirements in hybrid machine learning are explained. Third, the implications of fundamental limitations of hybrid machine learning are related to other types of machine learning. These contributions are relevant to AI system engineering [5] and AI-informed decision making [6]. Fourth, a contribution is made to machine ethics [7]. The paper goes beyond previous studies that have considered different forms of normative ethics [8], i.e., normative statements about what should be done, by focusing on the behavioral ethics of what people actually do when under pressure [9–12], such as driving faster than speed limits and/or across road junctions as traffic lights turn from amber to red [13]. Behavioral ethics is fundamentally an ecological analytical framework within which people might not adhere to regulations, etc., if doing so would be unfair because doing so would undermine their survival in their preferred states. An example of a preferred state is having a job that pays enough to support oneself and a young family. Behavioral ethics can be summarized by the title of the paper: "Why good people sometimes do bad things" [9]. Human organizations that need their HML implementations to operate in alignment with human ethical requirements as defined in regulations, etc., need to take behavioral ethics into account [13,14].

2. Representing Human Ethical Requirements in AML-Enabled Traffic Predictions

2.1. Ethical Requirements

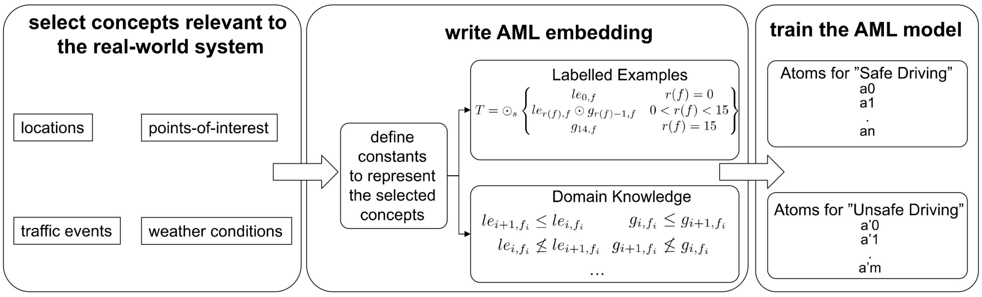
The sanctity of human life is an important ethical construct in many cultures [15]. Road traffic regulations are an example of everyday normative ethics related to the protection of human life. Road traffic regulations are implemented through, for example, traffic lights that indicate where and when drivers must stop their vehicles to reduce the risk of road accidents that could cause grave injuries to people. Road traffic regulations can entail locally simple cases. This is because human ethical requirements for road safety, such as where and when to stop road vehicles, are defined clearly by traffic lights and have limited scope for personal interpretation. For example, although the color red can have different associations generally in different cultures, red is used in traffic lights to indicate stop throughout the world [16,17]. Also, the consistent ordering of traffic lights can provide visual clarity for people who have color-vision deficiency: for example, red for stop always being positioned at the top. In addition, technological innovations are being developed to support drivers who have color-vision deficiencies in cities [18], where there can be a many traffic lights. This example illustrates how big ethical ideas can be expressed as vague general statements, such as the sanctity of human life [15], which can become increasingly specific as they come closer to particular implementations.

However, despite more than one hundred years of increasing traffic regulations, there continue to be road traffic accidents. Contributing to this can be individual people's changes in balance between their moral motivation and their ethical temptation [9–11]. Within behavioral ethics, being able to resist ethical temptation can depend upon a person having sufficient internal control to be able to resist external pressure. In particular, having sufficient self-regulatory control [19]. Combinations of psychological fatigue and physical fatigue can undermine self-regulatory control [20]. For example, a driver could have reduced self-regulatory control due to psychological and physical fatigue caused by many hours of driving that have involved having to stop at many traffic lights. If the driver is already very late (time pressure [12]) for an important meeting with the biggest customer of the driver's employer (organizational pressure [21]), the driver may be more likely to not slow down and stop when traffic lights are changing from amber to red.

Hence, a human driver may agree generally with the need for traffic regulations as described in the road traffic authority's official documentation and as implemented with traffic lights. Nonetheless, in particular situations that have led to psychomotor stress, the same human driver may accelerate above the speed limit to drive across road junctions as traffic lights turn to red [13]. Predicting the violation of traffic regulations, such as driving through red traffic lights, is an established topic in machine learning [22].

2.2. Representing Ethical Requirements in AML

As summarized in Figure 1, selected concepts for an application are first described through the definition of constants to represent those concepts. In this case, the concepts relevant to the real-world system are time-insensitive data, such as locations (e.g., latitude and longitude coordinates) and points of interest (e.g., road crossings, traffic lights), and time-sensitive data, such as traffic events (e.g., road closures, road works), and weather conditions (e.g., fog, snow) [22].



select concepts relevant to the real-world system
locations
points-of-interest
traffic events
weather conditions
write AML embedding
define constants to represent the selected concepts
Labelled Examples
$T = \odot_s \left\{ \begin{array}{cc} le_{0,f} & r(f) = 0 \\ le_{r(f),f} \odot g_{r(f)-1,f} & 0 < r(f) < 15 \\ g_{14,f} & r(f) = 15 \end{array} \right\}$
Domain Knowledge
$le_{i+1,f_i} \leq le_{i,f_i}$ $g_{i,f_i} \leq g_{i+1,f_i}$
$le_{i,f_i} \not\leq le_{i+1,f_i}$ $g_{i+1,f_i} \not\leq g_{i,f_i}$
...
train the AML model
Atoms for "Safe Driving"
a0
a1
.
an
Atoms for "Unsafe Driving"
a'0
a'1
.
a'm

Figure 1. Steps involved in the formulation of an AML model: road traffic example.

AML models are mathematical models comprising algebraic representations. AML is founded upon several core algorithms that can learn from human-defined constraints and from data [3,4]. The concepts and relationships between them, which are selected to represent those aspects of the real-world system that the AML model interacts with, are application specific.

Defining concepts for an application of AML is akin to other hybrid machine learning approaches. In AML, constants are the primitives used by the algebra. As with any machine learning application, it is necessary to specify the input that the machine learning model will receive. For example, for neural networks, it is necessary to define the size of the image before training an image classifier. Defining constants for AML is as flexible as defining inputs for sub-symbolic approaches, as images and time series data can also be represented by constants. Moreover, for hybrid machine learning approaches, it is also necessary to specify the high-level concepts and their relationships with the underlying data.

As shown in the second box of Figure 1, constants will in turn be used to specify labeled training examples and to specify domain knowledge. This specification is called an embedding. Models for that embedding are learned using the algorithms described in [3,4]. As summarized in the third box of Figure 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, and outputs. Relationships between inputs and outputs defined in atoms are binary. In this case, the binary definition is safe driving/unsafe driving. Safe driving refers to adherence to road traffic regulations. Unsafe driving refers to violation of road traffic regulations. 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 learned as new atoms are generated during training. As illustrated in Figure 2, this is an example of the representation of ethics in machine learning world models in general, and AML world models in particular, being a computer engineering challenge.

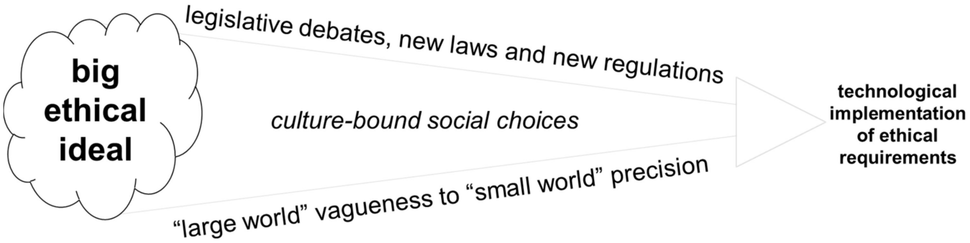

big ethical ideal
legislative debates, new laws and new regulations
culture-bound social choices
"large world" vagueness to "small world" precision
technological implementation of ethical requirements

Figure 2. From big ethical ideal to technological implementation of ethical requirements.

This is because implementations take place at the end of a typical engineering progression from vagueness to precision, as follows: big ethical ideal—debate about culture-bound social choices—legislation of new laws—definition of regulations—data from observations—definition of implementation requirements—technological implementations. Within the framing summarized in Figure 2, important debates about complex interactions between ethical ideals and social choices take place before the definition of requirements for technological implementations, which are the focus of this paper.

In this case, debate about culture-bound social choices [23,24] can lead to consensus that there need to be actions to minimize road traffic dangers. This can be followed by the legislation of broad road traffic laws, which can provide a basis for the formulation of detailed road traffic regulations. These regulations can then provide a basis for local governments to prepare location-specific road traffic management plans. Subsequent monitoring of road traffic can provide data, which can be used to train machine learning applications to predict risks of road traffic violations. For example, for each city block, machine learning can be used to predict, given its general features and recent (two hours) history, whether or not an accident will occur in the next 15 min [22]. More generally, the engineering progression from a big ethical ideal to technology implementations can be seen as a progression from the intractable unpredictability of vaguely defined "large worlds" to the more tractable, more predictable, precisely defined "small worlds" [25,26] in which specific rules can be implemented to try to reduce unethical behavior [27].

2.3. *Opposing Interpretations of Ethical Requirement Representations*

Locally, road traffic regulations can involve simple cases of representing human ethical requirements. This is because road safety requirements, such as where and when to stop road vehicles, are defined clearly and have limited scope for personal interpretation.

However, traffic regulations can be involved in wider ethical complexity. Firstly, traffic regulations are culturally bound social choices, which can be changed by new culturally bound social choices [23,24]. For example, in recent years, some lights and signs have been removed in order to improve the efficiency and safety of road traffic [28,29]. Thus, location-specific ethical requirements, as indicated by traffic lights, speed signs, etc., can change from one day to the next. This can lead to situations where people can have to pay fines and accept other penalties for driving at a speed above a new speed limit, despite having driven legally at that same speed at that same place every day for many previous years. This can happen if, for example, satellite navigation systems are not immediately up to date, and hence drivers do not have comprehensive information about where new speed limits apply [30]. Under such circumstances, some people who are penalized may not consider that road traffic regulations are applied ethically. More widespread complexity arises from the practice of delivery organizations arranging deliveries by self-employed gig economy delivery drivers to have short time durations and payment reductions if they deliver late. Such arrangements can entail organizational pressure and time pressure for drivers, which can be on top of chronic resource depletion due to poor employment conditions that can reduce moral motivation [13]. Such situations can lead to delivery drivers exceeding speed limits. However, it is only the delivery drivers who are penalized for violating road traffic regulations. This is because they are categorized as being self-employed: rather than being employees of the delivery company [31]. Accordingly, gig economy drivers may not interpret fines for violating traffic regulations as being ethical, even though they may not dispute that they have driven above speed limits. Eventually, there may be new culturally bound social choices that lead to new laws, regulations, and procedures, which could end the categorization of gig economy drivers as being self-employed [32]. This example illustrates the diverse scope of what needs to be updated in an AML model as ethical requirements change at specific locations, such as speed limits, and at many locations as employment regulations related to road traffic are changed. The updating of ethical requirements at specific locations could be carried out through automatic data collection via, for example, MapQuest. By contrast, the updating of ethical requirements due to changes in driver employment regulations would need to be human-defined.

3. Representing Human Ethical Requirements in AML-Enabled Diagnoses

3.1. Ethical Requirements

Some societies seek to provide universal healthcare [33]. In such societies, it would be considered unethical to deny healthcare to people who are not in good health. However, demand for healthcare services can exceed the supply of healthcare services, and triage may be needed in their allocation. Triage involves deciding which people have most need for healthcare and are most likely to respond positively to healthcare. However, deciding which people are most in need of healthcare and are most likely to respond positively to healthcare is particularly difficult for functional disorders. This is because these are disorders that impair normal bodily functioning but cannot be explained fully. Studies indicate that functional disorder patients “have often been misdiagnosed, correctly diagnosed after lengthy delays, and/or subjected to poorly delivered diagnoses that prevent diagnostic understanding and lead to inappropriate treatments, iatrogenic harm, unnecessary and costly evaluations, and poor outcomes” [34]. The term iatrogenic harm refers to inappropriate healthcare treatment that aggravates rather than improves a health problem [35]. Overall, high economic costs can arise from delayed diagnosis of functional disorders [36], which can reduce total healthcare budgets for all who are in need of healthcare. Furthermore, there is an ethical requirement to avoid conflicting interpretations of functional disorder diagnoses. This is because functional disorder patients who do not agree with a diagnosis may be less likely to respond to treatment [37]. By contrast, successful explanation of a functional disorder can contribute positively to treatment [38].

This example illustrates how big ethical ideals can be expressed as vague general statements, such as universal healthcare, which can become increasingly specific as they come closer to particular implementations. In this case, the aspiration to provide healthcare to all those in need of healthcare is related to the need for improved diagnoses of functional disorders. One option for applying machine learning in functional motor disorder diagnoses is to apply machine learning in gait analyses used to detect affective disorders, such as depression. The term gait refers to walking, running, and other means of natural locomotion combined with posture. It has been argued that gait analyses can provide a readily quantifiable objective approach to monitoring depression and related affective disorders [39]. This is important because, regardless of what type of event may initiate functional disorders, they can be amplified or perpetuated by affective disorders [34]. Apropos, machine learning has been applied to Kinect-recorded gait data in order to facilitate recognition of anxiety and depression [40]. In the next Section 3.2, a description is provided of representations in Algebraic Machine Learning, which can be used to examine relationships between gait and disorders. This example illustrates how representation of ethical requirements in machine learning can be concerned with specific dimensions of human psychomotor functioning, which are related to big ethical ideals such as universal healthcare, through causal chains that begin with vague general statements and end with precise technical representations.

3.2. Representing Ethical Requirements in AML

All AML models encompass selected concepts and relationships between them to represent those aspects of the real-world systems that they interact with. The first box in Figure 3 shows concepts relevant to this case. As shown in the second box of Figure 3, those constants will in turn be used to specify labeled training examples and to specify domain knowledge. As in the first case described in Section 2 above, and in all applications of AML, this specification is called an embedding. Models for that embedding are learned using the algorithms described in [3,4].

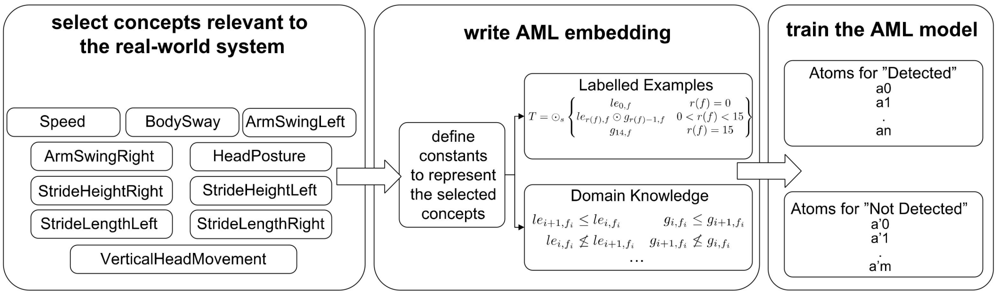

select concepts relevant to the real-world system
Speed
BodySway
ArmSwingLeft
ArmSwingRight
HeadPosture
StrideHeightRight
StrideHeightLeft
StrideLengthLeft
StrideLengthRight
VerticalHeadMovement
write AML embedding
define constants to represent the selected concepts
Labelled Examples
$T = \odot_s \left\{ \begin{matrix} le_{0,f} & r(f)=0 \\ le_{r(f),f} \odot g_{r(f)-1,f} & 0 < r(f) < 15 \\ g_{14,f} & r(f)=15 \end{matrix} \right\}$
Domain Knowledge
$le_{i+1,f_i} \leq le_{i,f_i}$ $g_{i,f_i} \leq g_{i+1,f_i}$
$le_{i,f_i} \not\leq le_{i+1,f_i}$ $g_{i+1,f_i} \not\leq g_{i,f_i}$
...
train the AML model
Atoms for "Detected"
a0
a1
.
an
Atoms for "Not Detected"
a'0
a'1
.
a'm

Figure 3. Steps involved in the formulation of an AML model: diagnosis example.

The gait analysis embedding comprises relevant features, as represented by AML. In this case, the domain knowledge consists of the notion of ordered intensity (i.e., telling the algebra that each feature is a number), while the training examples come from an existing real-world dataset [40]. The relevant 10 features are summarized in the first box of Figure 3. These features are based on analysis of previous studies by others that are reported in [41–48] and combined in [40]. As summarized in the third box of Figure 3, once the embedding is defined, AML learns by creating atoms. The core AML algorithms check that the human-defined constraints are maintained in the subsequent binary relationships between the inputs and outputs, which are learned as new atoms are generated during training. In the case of gait analyses, the real-world system that AML interacts with at the beginning consists of numbers describing gait features. These numbers come from conversion by specialist software of gait recordings are made by video cameras and/or wearable sensors. At the end, the real-world system that AML interacts with is the communication media of healthcare providers, such as their computer monitors.

As summarized in Figure 4, the performance of the AML can be considered in terms of perceptual, instrumental, and epistemic inference: i.e., cycles of active 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 internal representations in world models [14,49].

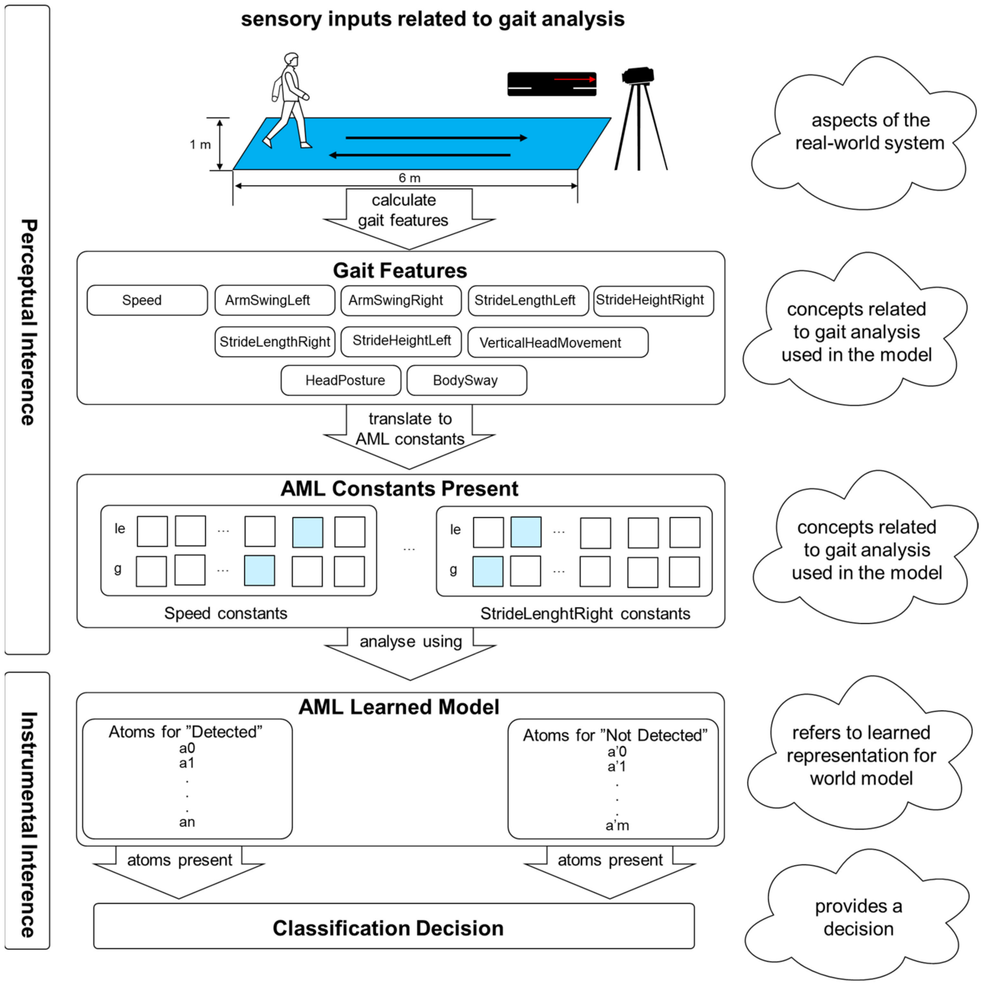


Figure 4. Operation of a trained AML model.

A machine learning world model, like all other world models, encompasses selected concepts and relationships between them to represent those aspects of the real-world system that it interacts with. Figure 3 above provides a summary of the selected concepts and relationships between them in this case of gait analysis. When evaluating a gait example, AML perceptual inference is in terms of constants present for that example (as indicated by the blue squares in Figure 4). 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. The accuracy of AML models can be evaluated using standard machine learning metrics such as the Macro F1 Score, which is calculated from the precision and recall of the test. Precision is the number of true positive predictions divided by the number of all positive predictions, including those not identified correctly. The recall is the number of true positive predictions divided by the number of all samples that should have been identified as positive. The Macro F1 Score is the mean of the F1 of each individual class.

As summarized in Figure 5, 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. Figure 5 summarizes that AML is a type of hybrid machine learning, as internal representations and relationships between them involve both constraints being defined by people and data-driven learning from new labeled examples.

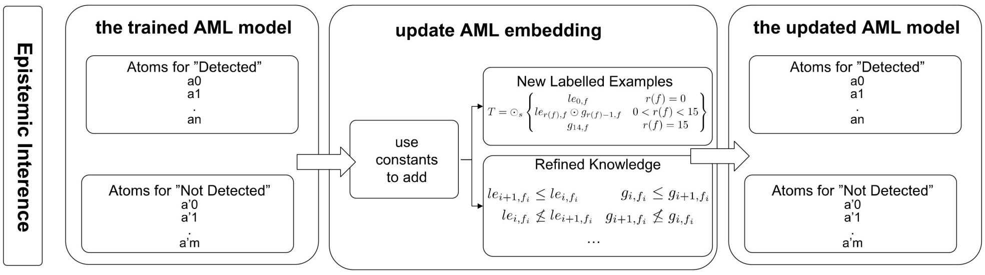

Epistemic Inference
the trained AML model
Atoms for "Detected"
a0
a1
.
an
Atoms for "Not Detected"
a'0
a'1
.
a'm
update AML embedding
use constants to add
New Labelled Examples
$T = \odot_s \left\{ \begin{matrix} le_{0,f} & r(f) = 0 \\ le_{r(f),f} \odot g_{r(f)-1,f} & 0 < r(f) < 15 \\ g_{14,f} & r(f) = 15 \end{matrix} \right\}$
Refined Knowledge
$le_{i+1,f_i} \leq le_{i,f_i}$ $g_{i,f_i} \leq g_{i+1,f_i}$
$le_{i,f_i} \not\leq le_{i+1,f_i}$ $g_{i+1,f_i} \not\leq g_{i,f_i}$
...
the updated AML model
Atoms for "Detected"
a0
a1
.
an
Atoms for "Not Detected"
a'0
a'1
.
a'm

Figure 5. AML model updating.

3.3. Opposing Interpretations of Ethical Requirement Representations

This example further illustrates that the representation of ethics in machine learning world models in general, and AML world models in particular, is a computer engineering challenge. For example, there can be culturally bound social choices to accept that functional disorders are an important and increasingly widespread medical phenomenon. Then, related normative ethics can be defined 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 (QMSs), which are developed and audited in accordance with international standards. AML, like other computational methods, can be implemented at the end of such progressions from vagueness to specificity [50].

However, even though the core AML algorithms check that the human-defined constraints are maintained, no amount of diligence in computer engineering can ensure that individual people who interact with a healthcare organization's implementation of machine learning will consider that their diagnoses are ethical. Rather, 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, which can resist definition due to dynamic, non-conscious interactions between variables such as personality type and body memory [51]. 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 [52]. In this case, there is the profound challenge that functional disorders have medically unexplained physical symptoms [53]. Thus, as summarized in Figure 6, even the most advanced computer engineering of machine learning world models for healthcare diagnoses cannot easily enable shared understanding of something that cannot yet be explained fully 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 high-quality visual content [54]. Hence, as summarized in Figure 7, the representation of ethical requirements in machine learning world models needs to be carried out with consideration of human organizations' documented world models in their QMS, individual people's embodied psychomotor world models, and the potential for deeply rooted opposition between them [14,55].

As summarized in Figure 7, HML-enabled gait analyses and other related information such as image study findings and self-reporting pain scales can be considered to be boundary objects between a healthcare provider's internal model of itself in the world (i.e., a documented world model) and an individual's model of self in the world (i.e., an embodied world model). Boundary objects have different meanings in different social worlds but nonetheless can be meaningful in 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 [56,57]. However, reference to existing world models can lead to organizations' lock-ins [58] and path dependencies [59] coming into opposition with individuals' motivated cognition [60] and wishful seeing [61]. All of which can lead to new information being interpreted to serve explanations that support preconceptions [62] and confirm biases [63]. Hence, even seeking to support results from analyses carried out with HML with information from other sources may be of limited usefulness in facilitating agreement between healthcare providers and patients. For example, results from gait analyses could be combined with other information from imaging studies, such as scans and self-reported pain scales. However, these can be of limited usefulness because patients' prior expectations can determine their interpretation of imaging studies [64] and can determine the extent to which they experience pain [65].

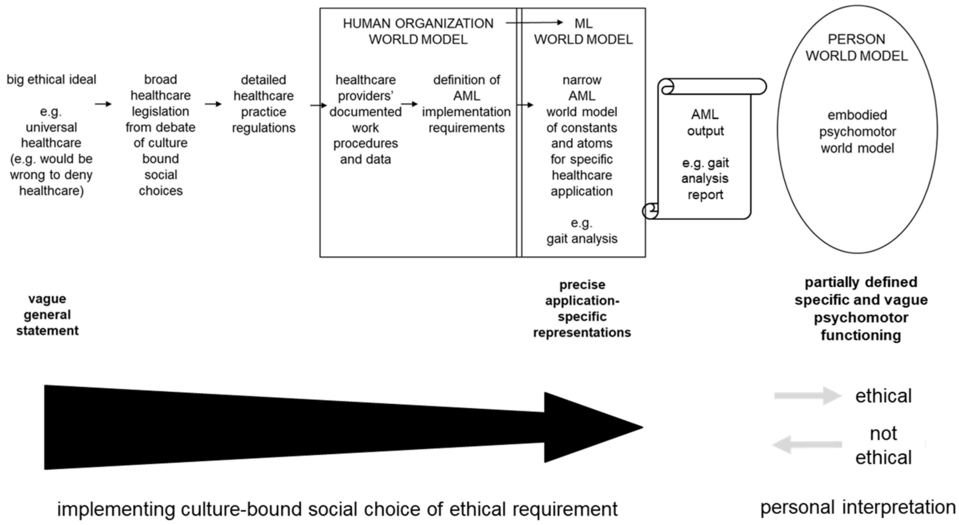

big ethical ideal
e.g. universal healthcare (e.g. would be wrong to deny healthcare)
broad healthcare legislation from debate of culture bound social choices
detailed healthcare practice regulations
HUMAN ORGANIZATION WORLD MODEL
healthcare providers' documented work procedures and data
definition of AML implementation requirements
ML WORLD MODEL
narrow AML world model of constants and atoms for specific healthcare application
e.g. gait analysis
AML output
e.g. gait analysis report
PERSON WORLD MODEL
embodied psychomotor world model
vague general statement
precise application-specific representations
partially defined specific and vague psychomotor functioning
ethical
not ethical
implementing culture-bound social choice of ethical requirement
personal interpretation

Figure 6. Implementation of ethical requirements may not be interpreted as ethical by all.

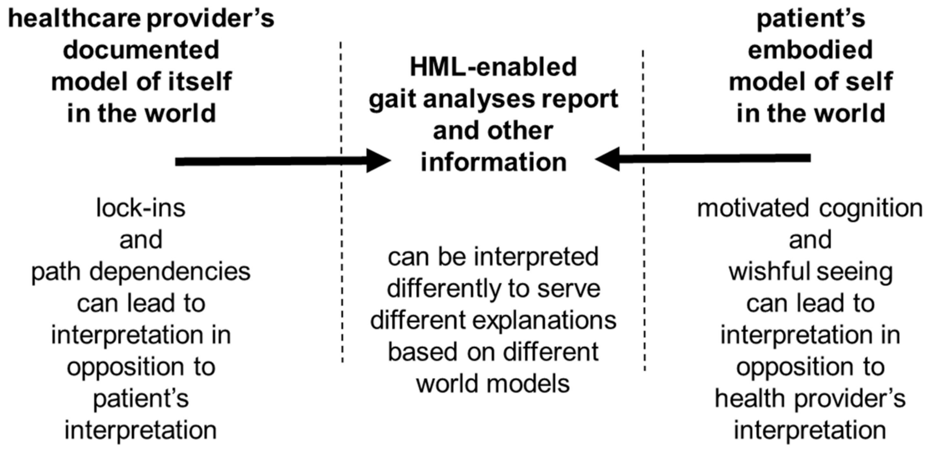

healthcare provider's documented model of itself in the world
HML-enabled gait analyses report and other information
patient's embodied model of self in the world
lock-ins and path dependencies can lead to interpretation in opposition to patient's interpretation
can be interpreted differently to serve different explanations based on different world models
motivated cognition and wishful seeing can lead to interpretation in opposition to health provider's interpretation

Figure 7. Opposing world models can lead to opposing interpretations of the same information.

Thus, in this case, there can be individual people who believe that they are suffering from a functional disorder, but machine learning enabled analysis and other related information indicate a low probability of functional disorder. Yet, they do not believe that they receive the healthcare that they believe they are entitled to receive within the big ethical ideal of universal healthcare. Accordingly, while it may be possible for computer engineering to enable implementation of a culture-bound social choice of an ethical requirement [23,24], that implementation may not be interpreted as ethical by individual people. By contrast, red traffic signals at road junctions are boundary objects that have one pre-agreed meaning irrespective of organizational, machine learning or personal world models. Red traffic signals at road junctions indicate that vehicles must stop, and any driver who does not stop their vehicle at a red traffic light violates traffic regulations that have been agreed as a necessary intervention to contribute to protecting human lives on the roads. In doing so, traffic regulations contribute to an ethical ideal that is prominent in many cultures: the sanctity of human life [15].

4. Conclusions

This paper is concerned with the representation of human ethical requirements in hybrid machine learning models, which, as summarized in Figure 2, can take place at the end of a progression from vagueness to precision. This can involve the following steps: big ethical ideal—debate about culture-bound social choices—legislation of new laws—definition of regulations—data from observations—definition of implementation requirements—technological implementations. Within this framing summarized in Figure 2, important debates about complex interactions between ethical ideals and social choices take place before the definition of requirements for technological implementations, which are the focus of this paper. In this paper, findings have been reported concerning technical opportunities and fundamental limitations for representing human ethical requirements in hybrid machine learning, as illustrated by Algebraic Machine Learning. These contributions are relevant to AI system engineering [5] and AI-informed decision making [6].

With regard to AI system engineering, the findings reported here illustrate that representation of human ethical requirements in hybrid machine learning is computer engineering work that can include both human-defined constraints and training from datasets (Figures 1 and 3–5). Furthermore, this is work that takes place at the end of typical processes that start with very broad general statements and narrow to precise application-specific representations. Previous work has drawn attention to the need to take into account psychological and cognitive aspects of human trust in AI system engineering [5]. This study draws attention to the importance of also taking human psychomotor functioning into account because it can determine the extent to which representations can describe human phenomena that involve dynamic non-conscious interactions between variables such as personality type and body memory. Also, human psychomotor functioning can influence the extent to which decisions informed by machine learning analyses will be perceived as ethical (Figures 6 and 7). Accordingly, this study introduces psychomotor functioning as an important consideration for future research concerned with human interpretations of the fairness of AI-informed decision making [6].

An important direction for further research would be to consider how to incorporate consideration of human psychomotor functioning into efforts to define generic rules for ethical AI implementations. For example, one generic rule that has been proposed is "AI decisions, actions, and communicative processes must be transparent and explainable" [66]. However, there is the fundamental challenge for explainability that some psychomotor phenomena cannot yet be explained, for example, phenomena characterized by medically unexplained symptoms. Furthermore, as summarized in Figure 7 above, while transparency and explainability are necessary, they are not sufficient to bring about agreement about outputs from machine learning. This is because different people who have different embodied psychomotor world models can have opposing interpretations of the same information.

Also, the study reported here goes beyond previous studies concerned with AI system engineering and AI-informed decision making that have not considered the importance of behavioral ethics. Future research into AI systems engineering for transportation systems could encompass predictions of the internal control levels of drivers. On-going improvements in wearable devices could enable psychophysiological fatigue [67], which can be an indicator of potential for self-regulatory control, to be measured. If drivers are willing to wear devices, such as earplugs, for this purpose, AI-informed decision making systems for improving road safety could include such measurements.

Author Contributions: Conceptualization, investigation, and writing, S.F.; software and validation, V.F.R. All authors have read and agreed to the published version of the manuscript.

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

Data Availability Statement: No new data were created.

Acknowledgments: The authors thank Paul Lukowicz for technical advice.

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

References

1. Button, A.; Merk, D.; Hiss, J.A.; Schneider, G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nat. Mach. Intell.* **2019**, *1*, 307–315. [CrossRef]
2. Wang, J.; Zhang, Q.; Zhao, D.; Chen, Y. Lane change decision-making through deep reinforcement learning with rule-based constraints. In Proceedings of the International Joint Conference on Neural Networks, Budapest, Hungary, 14–19 July 2019; p. N-2051.
3. Martin-Maroto, F.; de Polavieja, G.G. Semantic Embeddings in Semilattices. *arXiv* **2022**, arXiv:2205.12618.
4. Martin-Maroto, F.; de Polavieja, G.G. Algebraic Machine Learning. *arXiv* **2018**, arXiv:1803.05252.
5. Fischer, L.; Ehrlinger, L.; Geist, V.; Ramler, R.; Sobiezky, F.; Zellinger, W.; Brunner, D.; Kumar, M.; Moser, B. AI System Engineering—Key Challenges and Lessons Learned. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 56–83. [CrossRef]
6. Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and Explanation in AI-Informed Decision Making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579. [CrossRef]
7. Anderson, M.; Anderson, S.L.; Armen, C. Towards machine ethics. In *AAAI-04 Workshop on Agent Organizations: Theory and Practice*; American Association for Artificial Intelligence: San Jose, CA, USA, 2004; pp. 2–7.
8. Tolmeijer, S.; Kneer, M.; Sarasua, C.; Christen, M.; Bernstein, A. Implementations in machine ethics: A survey. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 132. [CrossRef]
9. Bersoff, D.M. Why good people sometimes do bad things: Motivated reasoning and unethical behavior. *Pers. Soc. Psychol. Bull.* **1999**, *25*, 28–39. [CrossRef]
10. De Cremer, D.; Van Dick, R.; Tenbrunsel, A.; Pillutla, M.; Murnighan, J.K. Understanding ethical behavior and decision making in management: A behavioural business ethics approach. *Br. J. Manag.* **2011**, *22*, S1–S4. [CrossRef]
11. De Cremer, D.; Vandekerckhove, W. Managing unethical behavior in organizations: The need for a behavioral business ethics approach. *J. Manag. Org.* **2017**, *23*, 437–455. [CrossRef]
12. Lee, E.-J.; Yun, J.H. Moral incompetency under time constraint. *J. Bus. Res.* **2019**, *99*, 438–445. [CrossRef]
13. Fox, S. Behavioral ethics ecologies of human-artificial intelligence systems. *Behav. Sci.* **2022**, *12*, 103. [CrossRef] [PubMed]
14. Fox, S. Human-artificial intelligence systems: How human survival first principles influence machine learning world models. *Systems* **2022**, *10*, 260. [CrossRef]
15. Baranzke, H. "Sanctity-of-Life"—A Bioethical Principle for a Right to Life? *Ethic Theory Moral Pract.* **2012**, *15*, 295–308. [CrossRef]
16. Kawai, C.; Zhang, Y.; Lukács, G.; Chu, W.; Zheng, C.; Gao, C.; Gozli, D.; Wang, Y.; Ansorge, U. The good, the bad, and the red: Implicit color-valence associations across cultures. *Psychol. Res.* **2023**, *87*, 704–724. [CrossRef] [PubMed]
17. Yan, F.; Li, B.; Zhang, W.; Hu, G. Red-light running rates at five intersections by road user in Changsha, China: An observational study. *Accid. Anal. Prev.* **2016**, *95*, 381–386. [CrossRef]
18. Ochoa, A.; Oliva, D. Smart traffic management to support people with color blindness in a Smart City. In Proceedings of the IEEE Latin American Conference on Computational Intelligence (LA-CCI), Gudalajara, Mexico, 7–9 November 2018; pp. 1–8.
19. Wang, Y.; Wang, G.; Chen, Q.; Li, L. Depletion, moral identity, and unethical behavior: Why people behave unethically after self-control exertion. *Conscious. Cogn.* **2017**, *56*, 188–198. [CrossRef] [PubMed]
20. Evans, D.R.; Boggero, I.A.; Segerstrom, S.C. The nature of self-regulatory fatigue and "ego depletion" lessons from physical fatigue. *Pers. Soc. Psychol. Rev.* **2016**, *20*, 291–310. [CrossRef] [PubMed]
21. Umphress, E.E.; Bingham, J.B.; Mitchell, M.S. Unethical behavior in the name of the company: The moderating effect of organizational identification and positive reciprocity beliefs on unethical pro-organizational behavior. *J. Appl. Psychol.* **2010**, *95*, 769–780. [CrossRef]
22. Moosavi, S. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, IL, USA, 5–8 November 2019; pp. 33–42.
23. Coleman, J.; Ferejohn, J. Democracy and social choice. *Ethics* **1986**, *97*, 6–25. [CrossRef]
24. Plott, C.R. Ethics, social choice theory and the theory of economic policy. *J. Math. Sociol.* **1972**, *2*, 181–208. [CrossRef]
25. Savage, L. *The Foundations of Statistics*; Wiley: New York, NY, USA, 1954.

26. Binmore, K. On the foundations of decision theory. *Homo Oecon.* **2017**, *34*, 259–273. [CrossRef]
27. Mulder, L.B.; Rink, F.; Jordan, J. Constraining temptation: How specific and general rules mitigate the effect of personal gain on unethical behavior. *J. Econ. Psychol.* **2020**, *76*, 102242. [CrossRef]
28. Cassini, M. Traffic lights: Weapons of mass distraction, danger and delay. *Econ. Aff.* **2010**, *30*, 79–80. [CrossRef]
29. Baker, L. Removing roads and traffic lights speeds urban travel. *Scientific American*, 1 February 2009. Available online: https://www.scientificamerican.com/article/removing-roads-and-traffic-lights(accessed on 1 March 2024).
30. Morris, S. Wales Is Bringing in a 20mph Speed Limit. *The Guardian*, 15 September 2023. Available online: https://www.theguardian.com/politics/2023/sep/15/wales-is-bringing-in-a-20mph-speed-limit-why-and-what-will-happen(accessed on 7 February 2024).
31. Christie, N.; Ward, H. The health and safety risks for people who drive for work in the gig economy. *J. Transp. Health* **2019**, *13*, 115–127. [CrossRef]
32. Russon, M.-A. Uber Drivers Are Workers Not Self-Employed, Supreme Court Rules. *BBC News*, 19 February 2021. Available online: https://www.bbc.com/news/business-56123668(accessed on 7 February 2024).
33. Webb, E.; Offe, J.; van Ginneken, E. Universal Health Coverage in the EU: What do we know (and not know) about gaps in access? *Eurohealth* **2022**, *28*, 13–18.
34. Espay, A.J.; Aybek, S.; Carson, A.; Edwards, M.J.; Goldstein, L.H.; Hallett, M.; LaFaver, K.; LaFrance, W.C.; Lang, A.E.; Nicholson, T.; et al. Current concepts in diagnosis and treatment of functional neurological disorders. *JAMA Neurol.* **2018**, *75*, 1132–1141. [CrossRef] [PubMed]
35. Nielsen, G.; Buszewicz, M.; Edwards, M.J.; Stevenson, F. A qualitative study of the experiences and perceptions of patients with functional motor disorder. *Disabil. Rehabil.* **2020**, *42*, 2043–2048. [CrossRef] [PubMed]
36. Tinazzi, M.; Gandolfi, M.; Landi, S.; Leardini, C. Economic costs of delayed diagnosis of functional motor disorders: Preliminary results from a cohort of patients of a specialized clinic. *Front. Neurol.* **2021**, *12*, 786126. [CrossRef]
37. Lidstone, S.C.; MacGillivray, L.; Lang, A.E. Integrated therapy for functional movement disorders: Time for a change. *Mov. Disord. Clin. Pract.* **2020**, *7*, 169. [CrossRef]
38. Stone, J.; Carson, A.; Hallett, M. Explanation as treatment for functional neurologic disorders. *Handb. Clin. Neurol.* **2016**, *139*, 543–553.
39. Hausdorff, J.M.; Peng, C.K.; Goldberger, A.L.; Stoll, A.L. Gait unsteadiness and fall risk in two affective disorders: A preliminary study. *BMC Psychiatry* **2004**, *4*, 39. [CrossRef] [PubMed]
40. Zhao, N.; Zhang, Z.; Wang, Y.; Wang, J.; Li, B.; Zhu, T.; Xiang, Y. See your mental state from your walk: Recognizing anxiety and depression through Kinect-recorded gait data. *PLoS ONE* **2019**, *14*, e0216591. [CrossRef] [PubMed]
41. Roether, C.L.; Omlor, L.; Christensen, A.; Giese, M.A. Critical features for the perception of emotion from gait. *J. Vision.* **2009**, *9*, 15. [CrossRef] [PubMed]
42. Coulson, M. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *J. Nonverbal Behav.* **2004**, *28*, 117–139. [CrossRef]
43. Gross, M.M.; Crane, E.A.; Fredrickson, B.L. Methodology for assessing bodily expression of emotion. *J. Nonverbal Behav.* **2010**, *34*, 223–248. [CrossRef]
44. Hazlett, R.L.; McLeod, D.R.; Hoehn-Saric, R. Muscle tension in generalized anxiety disorder: Elevated muscle tonus or agitated movement? *Psychophysiology* **1994**, *31*, 189–195. [CrossRef] [PubMed]
45. Michalak, J.; Troje, N.F.; Fischer, J.; Vollmar, P.; Heidenreich, T.; Schulte, D. Embodiment of sadness and depression—Gait patterns associated with dysphoric mood. *Psychosom. Med.* **2009**, *71*, 580–587. [CrossRef]
46. Brandler, T.C.; Wang, C.; Oh-Park, M.; Holtzer, R.; Verghese, J. Depressive symptoms and gait dysfunction in the elderly. *Am. J. Geriatr. Psychiatry* **2012**, *20*, 425–432. [CrossRef]
47. Natale, M.; Bolan, R. The effect of Velten's mood-induction procedure for depression on hand movement and head-down posture. *Motiv. Emot.* **1980**, *4*, 323–333. [CrossRef]
48. Lemke, M.R.; Wendorff, T.; Mieth, B.; Buhl, K.; Linnemann, M. Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. *J. Psychiatr. Res.* **2000**, *34*, 277–283. [CrossRef]
49. Friston, K.; Moran, R.J.; Nagai, Y.; Taniguchi, T.; Gomi, H.; Tenenbaum, J. World model learning and inference. *Neural Netw.* **2021**, *144*, 573–590. [CrossRef] [PubMed]
50. Winfield, A. Ethical standards in robotics and AI. *Nat. Electron.* **2019**, *2*, 46–48. [CrossRef]
51. Fox, S. Psychomotor predictive processing. *Entropy* **2021**, *23*, 806. [CrossRef] [PubMed]
52. Fox, S. Practical implications from distinguishing between Pearl blankets and Friston blankets. *Behav. Brain Sci.* **2022**, *45*, E194. [CrossRef] [PubMed]
53. Price, J.R.; Okai, D. Functional disorders and 'medically unexplained physical symptoms. *Medicine* **2016**, *44*, 706–710. [CrossRef]
54. Perloff, R.M. A three-decade retrospective on the hostile media effect. *Mass Commun. Soc.* **2015**, *18*, 701–729. [CrossRef]
55. Clark, C.J.; Winegard, B.M. Tribalism in war and peace: The nature and evolution of ideological epistemology and its significance for modern social science. *Psychol. Inq.* **2020**, *31*, 1–22. [CrossRef]
56. Ayobi, A.; Stawarz, K.; Katz, D.; Marshall, P.; Yamagata, T.; Santos Rodriguez, R.; Flach, P.A. Machine learning explanations as boundary objects: How ai researchers explain and non-experts perceive machine learning. In *Proceedings of the Joint Proceedings of the ACM IUI 2021 Workshops (Vol. 2903), CEUR Workshop Proceedings*; College Station, TX, USA, 13–17 April 2021.

57. Star, S.L.; Griesemer, J.R. Institutional ecology, translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939. *Soc. Stud. Sci.* **1989**, *19*, 387–420. [CrossRef]
58. Arthur, W.B. Competing technologies, increasing returns, and lock-in by historical events. *Econ. J.* **1989**, *99*, 116–131. [CrossRef]
59. Schreyögg, G.; Sydow, J.; Holtmann, P. How history matters in organisations: The case of path dependence. *Manag. Organ. Hist.* **2011**, *6*, 81–100. [CrossRef]
60. Nurse, M.S.; Grant, W.J. I'll see it when I believe it: Motivated numeracy in perceptions of climate change risk. *Environ. Comm.* **2020**, *14*, 184–201. [CrossRef]
61. Dunning, D.; Balcetis, E. Wishful seeing: How preferences shape visual perception. *Curr. Dir. Psychol. Sci.* **2013**, *22*, 33–37. [CrossRef]
62. Tetlock, P.E. Theory-driven reasoning about plausible pasts and probable futures in world politics: Are we prisoners of our preconceptions? *Am. J. Pol. Sci.* **1999**, *43*, 335–366. [CrossRef]
63. Pohl, R. *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*; Psychology Press: London, UK, 2004.
64. Zogas, A. "We have no magic bullet": Diagnostic ideals in veterans' mild traumatic brain injury evaluations. *Patient Educ. Couns.* **2022**, *105*, 654–659. [CrossRef] [PubMed]
65. Atlas, L.Y.; Wager, T.D. How expectations shape pain. *Neurosci. Lett.* **2012**, *520*, 140–148. [CrossRef] [PubMed]
66. Muller, H.; Mayrhofer, M.T.; Van Veen, E.B.; Holzinger, A. The ten commandments of ethical medical AI. *Computer* **2021**, *54*, 119–123. [CrossRef]
67. Kalanadhabhatta, M.; Min, C.; Montanari, A.; Kawsar, F. FatigueSet: A multi-modal dataset for modeling mental fatigue and fatigability. In *Pervasive Computing Technologies for Healthcare*; Lewy, H., Barkan, R., Eds.; PH 2021. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; Springer: Cham, Switzerland, 2022; Volume 431, pp. 204–217.

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