Hybrid AI: An Enabler for Trustworthy AI Systems by Enhancements along Four Dimensions

Albert Huizing[1], **Cor Veenman**[1,2], **Mark Neerincx**[1,3], **André Meyer-Vitali**[1], **Joris Sijs**[1], **Gertjan Burghouts**[1], **Maarten Kruithof**[1], **Judith Dijk**[1]

Abstract. In recent years, AI based on deep learning has achieved tremendous success in specialized tasks such as speech recognition, machine translation, and the detection of tumours in medical images. Despite these successes there are also some clear signs of the limitations of the current state-of-the-art in AI. For example, biases in AI-enabled face recognition and predictive policing have shown that prejudice in AI systems is a real problem that must be solved. In this position paper, we argue that current AI needs to be enhanced along four dimensions to become more trustworthy: *environment*, *purpose*, *collaboration*, and *governance*. Hybrid AI offers the potential for advancements along these four dimensions by combining two different paradigms in AI: knowledge-based reasoning and optimization, and data-driven machine learning. Some hybrid AI design patterns show how these paradigms can be combined to harness the advantages of both approaches while at the same time overcoming their limitations. We introduce two classes of systems that are enabled by hybrid AI: autonomous systems and human-machine teams. Several examples show how hybrid AI can be employed to make these system classes more trustworthy.

1 INTRODUCTION

Recent breakthroughs in Artificial Intelligence (AI) based on deep learning have allowed machines to perform at the same level as (or even surpass) humans in specialized tasks such as image classification, speech recognition, and machine translation. These breakthroughs are enabled by the tremendous growth in computational power, the availability of large annotated datasets, and new efficient machine learning algorithms. Most of the recent successes in AI can be attributed to supervised deep learning which is a machine learning approach that uses large labelled training sets and a gradient-based backpropagation algorithm to adapt millions of parameters in a layered deep neural network. The availability of big data and enormous computing power provided by modern graphical processing units are the main contributors to this success.

Despite these successes there are also some disturbing signs of undesirable behaviour of AI. For example, unwanted biases in algorithms for face recognition and fraud detection have shown that prejudice and bias in AI systems is a real problem that needs to be solved [1][2]. Furthermore, accidents with self-driving cars indicate that AI cannot yet be trusted to operate autonomously in safety-critical applications [3].

[1] TNO, Netherlands, email: albert.huizing@tno.nl
[2] LIACS, Leiden University, Netherlands
[3] Delft University of Technology, Netherlands

In contrast with the currently successful deep learning approach, knowledge-based AI uses symbolic knowledge representation, logic reasoning and optimization which offers the benefit of explainability and predictability but lacks the adaptability and effective handling of uncertainty that modern machine learning offers. Hybrid AI is a recent trend in AI that addresses the current limitations in AI by combining the best of knowledge-based methods and data-driven machine learning.

The purpose of this position paper is to argue that AI needs to be enhanced along four dimensions to become more trustworthy and that hybrid AI can enable these enhancements. The paper is organised as follows. Section 2 describes the four dimensions *environment*, *purpose*, *collaboration*, and *governance*. Section 3 introduces hybrid AI design patterns that describe different ways of combining knowledge-based methods and machine learning. Section 4 describes how two distinct AI system classes can be made more trustworthy by applying hybrid AI. Finally, section 5 gives conclusions.

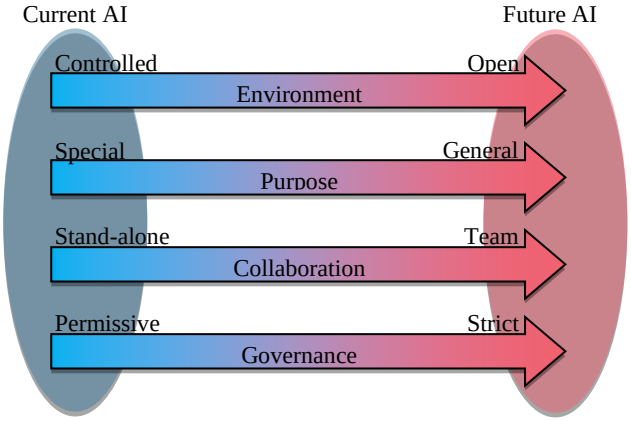

Figure 1. Improvements of AI along four dimensions.

2 FOUR DIMENSIONS OF AI

For AI to become more effective and accepted in society, there is a need for future AI to improve from the current state-of-the art. Figure 1 shows these needs along four dimensions: (i) *environment* (ii) *purpose*, (iii) *collaboration*, and (iv) *governance*. For the near future we foresee a need for AI to evolve from operations in a controlled environment to operations in an open world, from special purpose tasks to more general purpose problem solving, from a stand-alone system to a team of humans and AI, and from applications where the governance of the AI can be permissive to

applications where governance needs to be strict with respect to compliance with laws, ethical norms, and societal values. The AI challenges that emerge from these needs are discussed in more detail in the next sections.

2.1 Environment

When AI was introduced in the second half of the 20th century, it demonstrated considerable success in solving problems that were previously unattainable by computers. This first generation of AI used knowledge representations such as heuristics, rules and ontologies, and deductive reasoning to solve problems such as search and planning. However, it soon became clear that knowledge-based AI could only solve well-defined problems in carefully controlled environments where uncertainty is minimal and explicit knowledge instead of intuition mostly defines the solution to the problem.

Machine learning techniques such as support vector machines and deep neural networks use large labelled data sets to solve problems. Machine learning does not depend on explicit knowledge representations thereby reducing the need for scarce domain experts and broadening the range of environments in which it can operate. Uncertainty in the environment can also be handled better by machine learning than knowledge-based reasoning methods because it exploits the diversity and fluctuations in the training data to achieve statistically impressive results even when numerous potentially correlated parameters are involved. There are however still significant problems when using machine learning in open environments: e.g. environments with rare but important events, adverse conditions, and unforeseen situations. In these environments there is often few data or no data at all available for training. This scarcity of labelled training data is a big challenge for machine learning which restricts the applications in which AI can be deployed effectively and safely. For example, the danger of relying on machine learning in safety-critical applications such as an Advanced Driver Assist System (ADAS) is vividly illustrated by a recent experiment where a digit on a traffic sign was slightly modified [4]. The interpretation of the image of the traffic sign by the ADAS was 85 mph instead of 35 mph. Clearly, such an error could cause major problems and lethal accidents.

In summary, AI faces challenges in an open environment in the identification and characterization of unforeseen situations, a lack of training data and acting effectively and safely in unknown situations. Potential solutions for these challenges include compositional reasoning, simulation models to generate training data for rare events, and context awareness to preclude undesirable behaviour of AI-enabled systems [5].

2.2 Purpose

AI using deep neural networks can currently outperform humans on specialized tasks such as the detection of tumours in medical images after being extensively trained on large labelled image sets [6]. However, if the purpose of the task changes slightly, for example from localisation of a tumour to segmentation of a medical image, the loss function that encodes the purpose of the task and the neural network architecture must be redesigned, and the network must be retrained again. Approaches such as transfer learning can address this problem by reusing parts of a network that has been trained on large publicly available databases such as ImageNet. However, there are often significant differences between images used in different domains. For example, in medical images variations in local textures are used to detect tumours while in natural image datasets there is generally a global subject present [7]. Significant progress has been achieved with transfer learning using homogeneous data sets in decision support systems. Considerably more challenging are applications that involve heterogeneous data sources and planning and control of effectors such as in mobile robots [8].

Instead of designing a specific loss function for each task and tuning task-specific parameters until the machine learning algorithm performs satisfactorily, it would be useful to have a more general-purpose approach in which the problem to be solved can be described at a higher abstraction level. The challenge for general-purpose AI is to offer a user the flexibility to conduct a variety of tasks according to user preferences, ethics, and societal values, while avoiding a detailed specification by the user of how the AI should carry out these tasks [9]. To decide which course of action is best in the current situation, the AI needs a world model and domain knowledge to assess the impact of different actions [10].

2.3 Collaboration

Current AI mainly interacts in a pre-determined way with humans (and other systems) in their environment and acts like a smart stand-alone tool that is employed to solve specific problems. This predetermined interaction and fixed task allocation between humans and a smart AI tool helps to manage expectations and assure safety, but it also limits the effectiveness of combined human intelligence and artificial intelligence in complex and dynamic environments. Effective collaboration between humans and AI demands mutual understanding of each other's abilities and shortcomings. Currently, a proper level of mutual understanding and anticipation is lacking. Consequently, there is a need for AI that learns (1) to understand and interpret human abilities [11], and (2) to self-improving forms of collaboration [12].

Another aspect that limits the use of machine learning in collaborative systems is the black box nature of deep neural networks. Even for machine learning experts it is hard to understand how a deep neural network arrives at its conclusions. An explanation to an expert or user of how and why the AI arrived at certain conclusions is a challenge that has been widely recognized [13]. Such an explanation capability is supported by symbolic communication in terms of domain knowledge that humans understand [14].

2.4 Governance

Machine learning software behaves differently from conventional software in the sense that decisions are based on training data and not on rules and control flows engineered by humans. This has the advantage that less effort is needed to develop the software. Furthermore, novel solutions may be found to problems that have eluded scientists and software engineers. However, the disadvantage is that there is less (or no) awareness of unwanted biases in the data set. For some applications such as machine translation and recommender systems, the impact of biases is limited, and governance is permissive because humans can correct or compensate for mistakes. However, in many AI applications a stricter governance is needed because the tolerance for errors and biased decisions is low. A well-known example of biased decision

making is AI-based photo-categorization software that labels images of people with a dark skin as gorillas [16].

Fairness, i.e. decisions that are free of unwanted biases, is one of the pillars of the responsible use of AI. Fairness is sometimes in conflict with the accuracy of decisions that is also desired for a responsible use of AI. Most deep learning algorithms can achieve a high accuracy only by having access to large data sets. Individuals and organisations sometimes contribute willingly to large data sets to reap the benefits of useful machine learning applications. In case of sensitive data, however, most individuals and organisations are reluctant to disclose these data. This implies that they also do not benefit from the use of AI. A possible solution for this dilemma is multiparty computation that enable machine learning methods to learn from confidential data without disclosure of the data [17]. The final element of responsible use of AI addressed in this paper is that decision making also needs to be transparent. A lack of transparency leads to distrust and potentially to rejection of AI in society. However, transparency can also be at odds with the confidential treatment of data on which decisions are based. Fairness, Accuracy, Confidentiality and Transparency (FACT) are four objectives for the responsible use of AI that need to be addressed [14]. The challenge for AI is to conduct a trade-off between these conflicting objectives that depends on the context.

3 HYBRID AI DESIGN PATTERNS

The limitations of current data-driven machine learning methods have been identified by leading AI researchers and a combination with knowledge-based reasoning and optimization has been proposed as a potential solution to these limitations [18][19] [20][21][22]. There are, however, many ways to combine these approaches. We have adopted the boxology proposed by van Harmelen and ten Teije to categorise different hybrid AI methods [23]. The hybrid AI boxology use design patterns with two different elements: ovals for algorithms and boxes for their input and output. The oval algorithms represent knowledge-based methods (KR) or data-driven machine learning methods (ML). The input and output represented by rectangles concern knowledge or data. Some examples of hybrid AI design patterns that enable responsible human-machine teaming and safe autonomous systems are presented in this section. A more detailed discussion on hybrid AI use cases is provided in section 4.

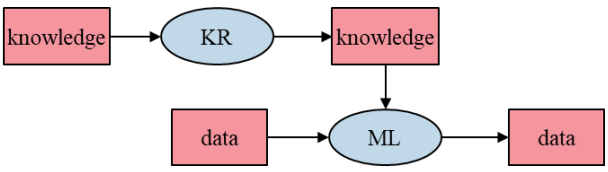

Figure 2. Hybrid AI design pattern using knowledge as a prior for machine learning.

Figure 2 illustrates a first example of a hybrid AI design pattern in which knowledge is used as prior information for machine learning. An example of a method that uses this pattern is a Logic Tensor Network which integrates fuzzy logic with a neural network to allow efficient learning from noisy data in the presence of logical constraints [24]. Sections 4.1.2 and 4.2.1 discuss the potential of this design pattern for novelty detection and bias mitigation.

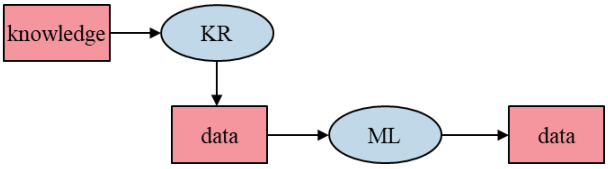

Figure 3. Hybrid AI design pattern using a knowledge-based model to generate training data for a machine learning method.

Figure 3 shows a hybrid AI design pattern where a knowledge-based simulation model is used to generate training data for a machine learning method. This method is useful in applications such as defence and security where representative and balanced training data sets are expensive or difficult to acquire [25].

A third example of a hybrid AI design pattern is shown in Figure 4 where a knowledge-based model is used to predict data that is measured by a system and the prediction errors are used to update the model. This pattern is useful for anomaly detection and competence assessment by autonomous systems, see section 4.1.1.

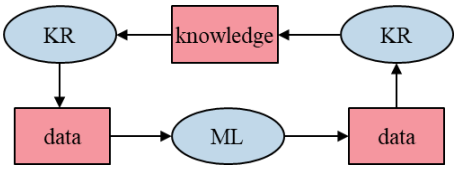

Figure 4. Hybrid AI design pattern using a model to predict measured data and prediction errors are used to update the model.

4 AI SYSTEM CLASSES

To illustrate the potential of hybrid AI for trustworthy intelligent systems, we propose to differentiate two AI system classes: autonomous systems and human-machine teams. Autonomous systems are being employed to replace humans in hazardous environments, in applications where reaction time is critical, or in jobs where skilled human workers are scarce. The challenge of AI for autonomous systems is to conduct tasks effectively and safely in an open environment without direct human intervention for an extended period of time. This primarily requires AI to be improved along the first two dimensions in Figure 1, i.e. *environment* and *purpose*. The second system class concerns a human-machine team which exploits the complementary capabilities of humans and AI to become more effective at conducting tasks while at the same time assuring compliance with laws, ethics, and societal values. To achieve this, AI needs to collaborate with other team members (humans and machines) and this requires enhancements along the final two dimensions in Figure 1, i.e. *collaboration* and *governance*. The difference in focus for AI in these two system classes is also illustrated by Figure 5 and Figure 8 with AI for autonomous systems focusing on the interaction loop with the environment and AI for human-machine teams concentrating on the collaboration loops within the team.

4.1 Autonomous system

Figure 5 illustrates an autonomous system that employs a combination of reasoning and optimization, a knowledge base, machine learning, and a data base to support the Observe, Orient,

Decide, (OODA) loop that is needed to operate in an open world [26]. Deductive reasoning and domain knowledge in the form of a world model enables the autonomous system to interpret the objective of the task that is specified by the user at a high abstraction level. Knowledge of the external world and the functional capabilities of the autonomous system is valuable for planning an effective course of action and adaptation of the system configuration in a complex dynamic environment. This use case of hybrid AI is elaborated in section 4.1.1. Furthermore, compositional reasoning helps to characterize unforeseen situations in terms of symbols that a human operator understands [5]. In addition, domain knowledge helps to partly solve the lack of training data that is characteristic for an open world. For example, hierarchical novelty detection using a taxonomy of objects could be used to alleviate the problem of characterizing novel objects. This approach, which uses the hybrid AI design pattern shown in Figure 3, is discussed in more detail in section 4.1.2. Another application of this hybrid AI pattern is to restrict the output of machine learning algorithms to labels that are valid in the current context. For example, misinterpretation of traffic signs could be avoided by restricting the output to official traffic signs.

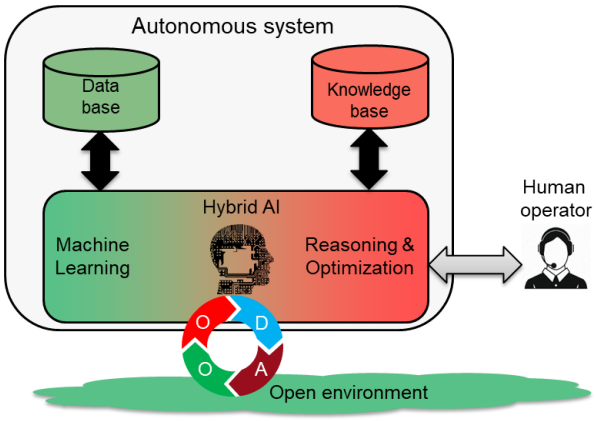

Figure 5. General purpose autonomous system operating in an open environment enabled by hybrid AI.

4.1.1 *Competence assessment*

AI applied in an autonomous system must be able to conduct a variety of tasks in an environment that may differ significantly from the environment for which it was originally designed or trained. In these situations, current AI methods are often unreliable and make mistakes while the confidence level as estimated by the AI itself is high. This happens not only for adversarial examples that have been crafted to deliberately mislead the AI, but also for naturally occurring situations [27]. To be able to operate safely in complex dynamic environments and gain the trust of users, AI must be able to reliably assess its own competencies with or without assistance of a human operator [28].

Here we propose a hybrid AI method that uses an ontology as a knowledge representation of its internal processes and configuration to assess its competencies in the current environment. Figure 6 shows the basic, domain independent modelling element of the ontology that describes not only processes in any system or environment but also the relation between a process, its inputs and its outputs. Each process has a specific performance and a health state. The performance indicates how well the process is carrying out the current task in the current situation. This is based on the quality of the input data and how well the input data distribution matches the expected data distribution. For processes that contain learning elements, the world model is updated and keeps track of the conditions in which the world model is valid according to the hybrid AI pattern in figure 4. The health state indicates how 'healthy' a component is; i.e. if the process is functioning perfectly the health state will be 1, if it is malfunctioning the health state will be 0. The specification of the inputs and outputs rely on the process. A model of the competence of an entire AI-based system configuration that contains multiple processes is composed from a hierarchy of the basic modelling elements. The overall competence assessment for the entire system configuration is based on an aggregation of the performance of the individual processes.

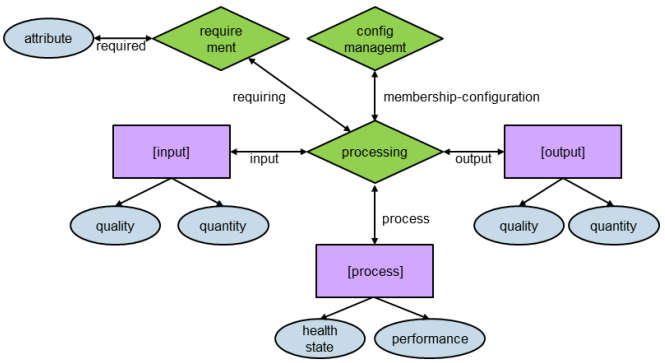

Figure 6. Main modelling element of a competence assessment ontology. Purple squares are entities, blue ovals are attributes and green diamonds are relations.

4.1.2 *Hierarchical novelty detection*

In an open world, novel objects will be encountered by an autonomous system that have not been or rarely seen before. Current machine learning techniques for object classification that learn from lots of examples do not perform well in this case [29]. One way to address this problem is to use the fact that novel objects almost invariably share some characteristics with objects that have been seen before. These common characteristics can be exploited by a hierarchical novelty detection method that uses a task-specific taxonomy of object classes to avoid errors that can lead to undesirable consequences [30]. In an object class taxonomy, each known object class is a superclass of its children classes and subclass of its parent class. The leaves in a taxonomy represent the most specific object classes while the root represents the most generic object class. The use of an object class taxonomy as prior knowledge in a machine learning algorithm is an example of the hybrid AI design pattern shown in Figure 2.

To achieve hierarchical novelty detection, a machine learning algorithm is first trained in a supervised way by presenting it with a data set of known objects and the associated hierarchy of labels in the taxonomy from leaf to root. After being trained, the algorithm assigns the most specific object class for known objects while for novel objects the nearest superclass is assigned. To illustrate the potential benefits of this approach for dilemmas that might be faced by autonomous vehicles, Figure 7 shows an example of an object class taxonomy. At the highest level in the taxonomy below the root there are four object super classes (vehicle, pedestrian,

animal, and small obstacle). Objects that are detected by the vehicle and assigned to different super classes may lead to very different decisions and actions by the autonomous vehicle. For example, the detection of a plastic bag in front of the car would normally not cause the autonomous vehicle to brake, while the detection of a pedestrian or animal should lead to an emergency stop. A novel object such as a tree branch should be assigned to the small obstacle superclass and lead to similar behaviour as for a plastic bag. On the other hand, a novel object such as a cat should not be assigned to the superclass small object but to the superclass animal leading to an emergency stop if no evasive manoeuvre is possible [31].

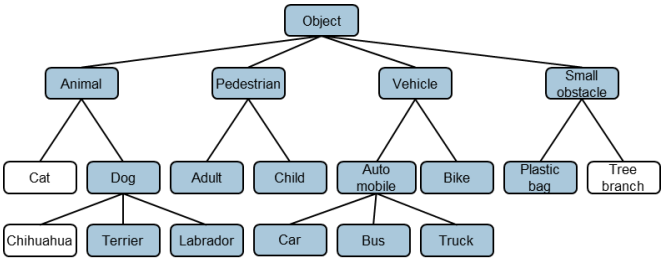

Figure 7. Example of an object class taxonomy for an autonomous vehicle. Known object classes are indicated as boxes with solid blue shading and unknown classes as white shaded boxes.

4.2 Human-machine team

In human-machine teams, AI and human team members need to interact on a regular basis to exploit the complementary skills and capabilities of human and artificial intelligence in the execution of a task. A prerequisite for effective and responsible team operations is that the team members have a shared view of the objectives that should be achieved and the capabilities and limitations of the team members. In addition, the team members should be able to use each other's data and knowledge to learn from each other [32]. This shared view and team learning ability is enabled by the exchange of knowledge and data between the team members, see Figure 8.

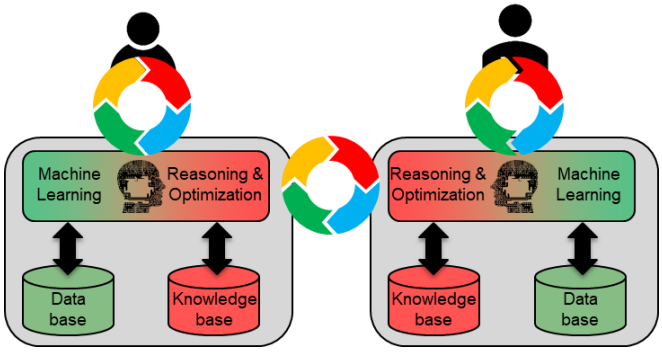

Figure 8. Human-machine team enabled by hybrid AI.

Domain knowledge represented by heuristics and physical, biological or behavioural models can express how entities in the world relate to each other and can predict how the environment changes because of certain events and actions. A reasoning process serves to transfer domain knowledge, which is a compressed form of data and experience collected during many years of experience, from one task to another. That is, knowledge offers a level of abstraction above the concrete and granular details of a sensory experience or observation, an abstraction that allows humans to transfer what they learned in one place to a problem that they may encounter somewhere else.

Another important aspect is that although there is a benefit in cooperation by pooling skills and resources, laws and intellectual property concerns may preclude sharing data that could be used to conduct the task. For example, privacy laws impose restrictions on insurance companies and hospitals to share patient data while the combined data could be used to more effectively treat patients [33]. This requires a reasoning and optimisation process supervised by humans that balances the need for data and knowledge sharing and the need to know.

4.2.1 Hierarchical bias mitigation

The goal of fair decision making in a human-machine team is to avoid unlawful biases and decisions in sensitive societal applications [34]. For instance, a fair decision support system using AI should not suggest lower wages to women or should not favour specific ethnic groups in suggesting loans and insurance policy schemes. One of the main challenges for fair decision support is that if a historical data set contains biases and a machine learning algorithm is trained to make accurate predictions on such data, then the decisions made by humans based on these predictions will be biased as well. An example of biased decision making is the System for Risk Indication (SyRI) employed by the Dutch Ministry of Social Affairs to predict the likelihood of an individual committing benefits or tax fraud [2]. In February 2020, the Dutch high court ruled against the use of this system because it violates human rights. While details of SyRI have not been disclosed by the Dutch government, it has become apparent that SyRI was used to detect fraud only in disadvantaged neighbourhoods. This way of working reinforces existing biases because fraud is only detected in poor neighbourhoods.

One of the simplest methods to avoid unwanted biases is to remove protected or sensitive attributes from the input data. However, when applying machine learning to these data, biases may be introduced through proxies of the protected attribute(s). For example, the postal code could be used as a proxy of ethnicity by a machine learning algorithm.

In this section, we introduce the concept of a hierarchical bias mitigation algorithm to reduce bias from a historical dataset by generating a transformed dataset that is still readable by humans. The algorithm uses a knowledge representation in the form of a taxonomy of the protected or sensitive attribute in which higher abstraction levels of the attribute reduce the probability of biases with respect to lower abstraction levels. with geographical location, then we may find a bias with data points from Rotterdam West and Rotterdam North. The hierarchical bias mitigation algorithm would then set the geographical attributes of these data points to Rotterdam. On the other hand, if the bias involves Rotterdam and Amsterdam, we can set the geographical property to The Netherlands, see Figure 9.

To be able to reduce the bias in the dataset, a machine learning algorithm is trained to minimize a loss function that comprises three parts. The first part is an inverted loss on the protected attribute such as ethnicity. The better the algorithm can predict this attribute from the unbiased dataset the higher the loss. The second part of the loss function is the prediction loss that measures how

accurate the model can predict the attribute that is going to be used for the unbiased dataset. The third part measures how close the unbiased dataset is to the original dataset. This loss is used to make sure that attributes that do not contribute a lot to the prediction and do not contribute to the bias are still accurately represented in the unbiased dataset.

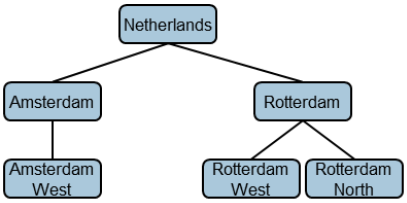

Figure 9. Geographical taxonomy for hierarchical bias mitigation.

5 CONCLUSIONS

In this position paper, we have argued that current AI needs to improve along four dimensions to become more trustworthy: *environment*, *purpose*, *collaboration*, and *governance*. We have also reasoned that hybrid AI which combines knowledge-based methods, and data-driven machine learning can address the challenges to improve AI along those dimensions. Some hybrid AI design patterns in this position paper illustrate different ways to combine knowledge-based and data-driven methods. To clarify the potential benefits of hybrid AI we introduce two distinct system classes that focus on different AI interactions: AI for autonomous systems primarily interacts with the environment and AI for human-machine teams mainly interacts with the team members. We have described several hybrid AI uses cases that we are currently exploring in our endeavour to make AI for autonomous systems and human-machine teams more trustworthy. Detailed descriptions and concrete results of those hybrid AI use cases will be published in forthcoming technical papers.

ACKNOWLEDGEMENTS

This position paper was funded by the TNO Early Research Program Hybrid AI. The authors acknowledge the valuable discussions with Frank van Harmelen and Annette ten Teije about hybrid AI design patterns.

REFERENCES

[1] P. Grother, M. Ngan, K. Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, NIST NISTIR 8280 (2019)

[2] M. Hunt, *Automated welfare fraud detection system contravenes international law, Dutch court rules*, Global Government Forum https://www.globalgovernmentforum.com/ (2020)

[3] S. S. Banerjee et al., *Hands Off the Wheel in Autonomous Vehicles?: A Systems Perspective on over a Million Miles of Field Data*, 48th Conf. Dependable Systems and Networks (DSN), pp. 586-597 (2018)

[4] S. Povolny and S. Trivedi, *Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles*, /www.mcafee.com/blogs (2020)

[5] B. Lake et al., *Building machines that learn and think like people*, Behavioral and Brain Sciences, 40, E253 (2016)

[6] S.M. McKinney et al. *International evaluation of an AI system for breast cancer screening*. Nature 577, 89–94 (2020)

[7] M. Raghu, C. Zhang, G. Brain, J. Kleinberg, S. Bengio, *Transfusion: Understanding Transfer Learning for Medical Imaging*, Advances in Neural Information Processing Systems 32 (2019)

[8] K. Weiss, T.M. Khoshgoftaar, D. Wang, *A survey of transfer learning*, Journal of Big Data 3, 9 (2016).

[9] P. Werkhoven, L. Kester, M. Neerincx, *Telling autonomous systems what to do*. Proceedings ECCE (2018)

[10] P. Elands et al., *Governing ethical and effective behavior of intelligent systems*, Militaire Spectator, 188(6), pp. 303-313 (2019).

[11] J. van Diggelen et al., *Developing Effective and Resilient Human-Agent Teamwork Using Team Design Patterns* IEEE Intelligent Systems, vol. 34, no. 2, pp. 15-24, (2019)

[12] M.A. Neerincx et al., *Socio-Cognitive Engineering of a Robotic Partner for Child's Diabetes Self-Management*. Frontiers in Robotics and AI, 6:118. (2019)

[13] A. Adadi and M. Berrada, *Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)*, IEEE Access, 6: (2018)

[14] W.M.P. van der Aalst, M. Bichler, & A. Heinzl, *Responsible Data Science*, Bus Inf Syst Eng 59, 311–313 (2017)

[15] S.R. Islam et al., *Domain Knowledge Aided Explainable Artificial Intelligence for Intrusion Detection and Response*, AAAI-MAKE, (2020)

[16] T. Simonite, *When It Comes to Gorillas, Google Photos Remains Blind*, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/ (2018)

[17] A. Sangers et al., *Secure multiparty PageRank algorithm for collaborative fraud detection*, IACR Cryptology (2018).

[18] M. Garnelo and M. Shanahan, *Reconciling deep learning with symbolic artificial intelligence: representing objects and relations*, Current Opinion Behavioral Science, 29, pp. 17-23 (2019)

[19] G. Marcus and E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*, Penguin Random House (2019)

[20] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, Basic Books, New York (2018)

[21] P. Battaglia et al., "Relational inductive biases, deep learning, and graph networks", arXiv:1806.01261v2 (2018)

[22] L. de Raedt et al. *NeuroSymbolic= Neural + Logical + Probabilistic.* NeSy'19@ IJCAI (2019)

[23] F. van Harmelen and A. ten Teije, *A Boxology of Design Patterns for Hybrid Learning and Reasoning Systems*, Journal of Web Engineering 18(1), 97–124 (2019)

[24] I. Donadello, L. Serafini, A. D'Avila Garcez, *Logic tensor networks for semantic image interpretation*. In IJCAI, 1596–1602 (2017)

[25] J. Dijk, K. Schutte, S. Oggero, *A vision on Hybrid AI for military applications*, SPIE Conference AI and ML in Defense (2019).

[26] G. Beckers et al., *Intelligent autonomous vehicles with an extendable knowledge base and meaningful human control*, SPIE (2019).

[27] M. Hein, M. Andriushchenko, and J. Bitterwolf, *Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem*, CVPR (2019)

[28] G. Burghouts, A. Huizing, M. Neerincx, *Robotic self-assessment of competence*, Workshop Assessing, Explaining, and Conveying Robot Proficiency for Human-Robot Teaming HRI Conference (2020)

[29] A. Bendale and T.E. Boult, *Toward Open Set Deep Networks*, CVPR (2016).

[30] K. Lee et al., *Hierarchical novelty detection for visual object recognition*, CVPR (2018)

[31] J-F. Bonnefon, A. Shariff, I. Rahwan, *The social dilemma of autonomous vehicle*, Science 352(6293): 1573–1576 (2016)

[32] M. Johnson et al., *Coactive design: Designing support for interdependence in joint activity*, Journal of Human-Robot Interaction, 3(1), 43–69, (2014)

[33] T. Attema, *A New Approach to Privacy-Preserving Clinical Decision Support Systems*, ArXiv https://arxiv.org/abs/1810.01107 (2018)

[34] N. Mehrabi et al., *A Survey on Bias and Fairness in Machine Learning*, arXiv:1908.09635v2 (2019)