Anna van Buerenplein 1
2595 DA Den Haag
P.O. Box 96800
2509 JE The Hague
The Netherlands

www.tno.nl

T +31 88 866 00 00

TNO report

TNO 2019 R11941

Hybrid AI White Paper

Date	13 December 2019
Author(s)	André Meyer-Vitali, Roos Bakker, Michael van Bekkum, Maaike de Boer, Gertjan Burghouts, Jurriaan van Diggelen, Judith Dijk, Corrado Grappiolo, Joachim de Greeff, Albert Huizing, Stephan Raaijmakers
Copy no	
No. of copies	
Number of pages	27 (incl. appendices)
Number of appendices	0
Sponsor	
Project name	Early Research Programme on Hybrid Artificial Intelligence
Project number	060.38605

Hybrid Artificial Intelligence combines the best of two worlds: the power of recent advances in deep learning with the possibility to explicitly model human knowledge in connected systems. Together, this new generation of AI becomes more controllable, explainable and fair, in line with European norms and values.

Contents

1. Motivation

Current approaches in Artificial Intelligence (AI), particularly machine learning, have recently reached an unprecedented impact in science and society. There are, however, concerns about interpretability and accountability of such AI, which limit its usefulness and trustworthiness. Current learning AI systems operate almost exclusively in a statistical (model-free) mode, which limits their potential and performance [1]. Many works point out that there is a fundamental need for knowledge representation, reasoning and sharing that is integrated with machine learning systems in order to provide sound and explainable models to alleviate these problems [2, 3, 5, 6, 7, 8].

Hybrid AI (HAI) systems are considered to be the next wave of AI, integrating the fundamental cognitive abilities of intelligent agents: being able to learn from their environment, reason about what has been learned [2, 3, 4] and to share the acquired knowledge. In such integrated systems, machine learning approaches provide robust learning, whereas semantic models and knowledge representation allow for reasoning capabilities [2]. Consequently, they bring a deeper level of understanding to intelligent systems and overcome some of the limitations of current AI systems.

Current AI approaches are highly specialised, in the sense that they are very good at performing well-defined tasks. Transfer of learned abilities and applying them to other contexts (e.g. transferring the driving skills of an autonomous car to a golf cart) is currently very limited [5, 8]. Understanding context is necessary for an agent's ability to cope with unknown situations, for providing feedback to users and being aware of its own functioning [8]. Current AI systems can fail ostentatiously when cases appear in their context that do not fit the training model [8]. Incidents with autonomous vehicles abound. Furthermore, current AI systems depend on huge training data sets: in order to reach acceptable levels of accuracy, algorithms require large quantities of example data [5, 8], which is in stark contrast with how humans are able to integrate new information through, e.g., one-shot learning. Artificial Intelligence is not a new discipline, but one that has been at the forefront of targeting the most ambitious, complex and multi-disciplinary philosophical and technical challenges for 70 years [9]. Recent successes in AI seem to demonstrate its amazing short-term potential. While this is certainly true for a small subset of AI technologies (such as deep convolutional neural networks) and applications, it is a mistake to generalise this potential and to be either over-excited by exaggerated expectations or scared by dramatised potential threats. Either view is founded in unrealistic optimism about AI's potential - certainly within the authors' lifetimes. Instead, we aim for practical solutions to support humankind in coping with their imminent global challenges [10].

Artificial Intelligence is an umbrella term used for a large collection of diverse techniques (cf. fig. 1). The recently popular subfield of Deep Learning (DL) achieves great success in winning at games like chess and Go. However, Deep Learning is shallow. DL does statistical pattern matching in huge amounts of data (Big Data) without knowing what the data is actually about. Even complex games are still easy to win with the current supercomputers, because they provide complete information with absolute certainty. The real world is nothing like this: it's full of uncertainties,

inconsistencies and lack of information. AI based on DL alone cannot make sense of the world. Therefore, Hybrid AI attempts to enrich learning from data with reasoning using knowledge and social interaction. The goal of Hybrid AI is to bridge the semantic gap between data and symbolic knowledge (concepts, relationships and rules) in order to solve problems that are currently out of the reach of even the most advanced AI systems.

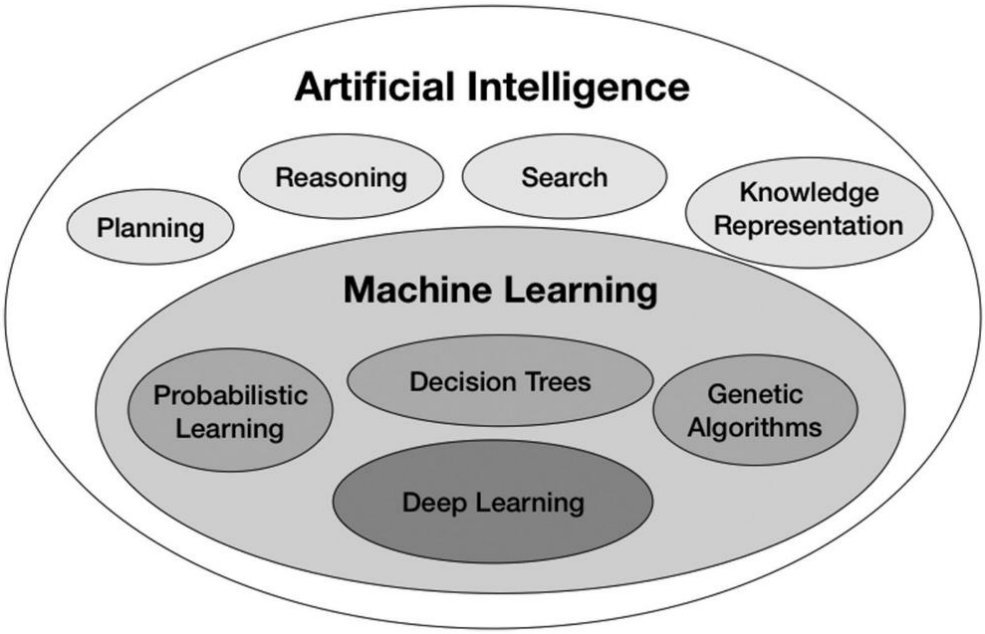

Figure 1: Artificial Intelligence and some of its subfields [11].

Due to their integrated nature, Hybrid AI systems will feature improvements on both accounts: they will be able to understand the context they operate in from their reasoning and abstraction abilities and adapt accordingly [8]. Hybrid AI should thus be able to overcome the limitations of current AI when observing new phenomena in open world situations by being able to combine data and contextual, descriptive models as input to the learning process [2, 5]. For example, providing pictures of a thus far unseen bird combined with a model that describes its features (two legs, beak, wings, etc.) will allow for a more efficient learning process [2, 8]. These systems will also be far better at explaining how they arrived at a certain conclusion: instead of just detailing input features, such a system will not only be able to recognise the bird, but also explain why it came to this conclusion based on symbolic descriptive models [2, 6, 7]. Such symbolic explanations are required for sharing knowledge and for coordinating tasks with other agents and humans. Rarely, a single system engineering approach can solve all possible problems. For example, human brains operate using two different concurrent systems [12]. Accordingly, it only makes sense that AI systems would benefit from a similar integration of different mechanisms. It appears advantageous to reuse knowledge when learning, as well as to adapt knowledge by experience. Scientists have always strived to learn from experimentation in order to discover new laws that, themselves, enabled new experiments. This approach of continuous improvement has led to scientific progress, which can be used for AI systems to become increasingly smart and transparent. Furthermore, a lot of time and energy can be saved by avoiding to learn what is already known (‘reinventing the wheel’).

2. Definition

The heart and raison d'être of Hybrid AI is the goal of bridging the semantic gap.

A Hybrid AI system of autonomous intelligent agents would ideally be adaptive and flexible to learn new concepts, interpret and relate concepts in the context of common and acquired knowledge, reason about hypotheses and consequences, take actions and communicate with other agents to collaboratively solve complex problems (distribute tasks among specialists). We focus on Hybrid AI systems that aim to assist humans in daily life by providing smart support tools (collaborative decision support) and by operating autonomously on their behalf (autonomous decision making), while considering given goals and norms. Crucial for such AI tools is the ability to interpret the (social) context in which they operate.
In general, a hybrid system is one that incorporates and integrates more than one (computing) paradigm. There are many ways to characterise the opposing features of a hybrid system of AI components, as shown in table 1.

Probabilistic	Deterministic
Statistical	Logical
Sub-symbolic	Symbolic
Neuro-	Symbolic
Intuitive	Conceptual
Empirical	Rational
Implicit	Explicit
Data(-driven)	Knowledge(-driven)
Solipsistic	Collaborative
Detached	Contextual
Reactive	Intentional
System 1 (Fast Thinking)	System 2 (Slow Thinking)

Table 1: Contrasting characteristics of AI approaches.

In comparison with a current AI system a Hybrid AI system would be (cf. fig. 2):

- suitable for general purpose tasks, instead of being limited to special purposes;
- able to operate in dynamic and unpredictable situations;
- collaborate with humans and other systems (agents); and
- suitable for situations where privacy, safety, security and ethical constraints are critical.

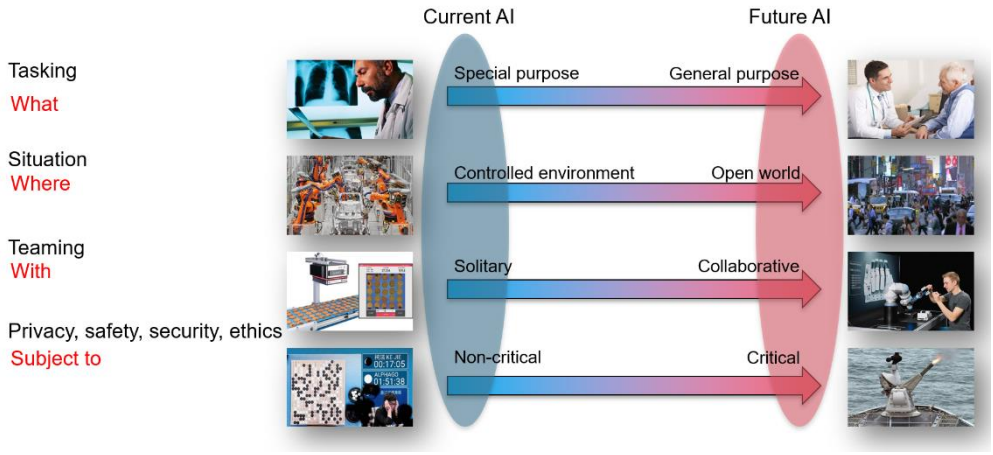

Figure 2: Characteristics of future AI systems.

None of the existing technologies is sufficient to build such systems. By bringing diverse relevant technologies together, Hybrid AI appears particularly promising to enable and to reap the potential impact. Therefore, intensive research should be conducted for making significant progress in this technology and its applications in many domains.

2.1 Semantic Gap

Hybrid AI unites the traditionally disparate camps of rational reasoners and empirical learners. Throughout the history of AI these camps followed their separate research agendas. However, more recently it appears that a combined approach would be more successful in pushing the limits of AI. Important issues, such as building meaningful, transparent and trustworthy AI systems, force the camps to unite. Machine learning (ML) is a generic approach that relies on detecting patterns in data to build statistical models of experience. Symbolic reasoning approaches allow for precision and transparency in defining knowledge in descriptive structural models. As collaborative systems of autonomous intelligent agents (robots, software agents, drones, smart assistants and home devices, cloud services or mobile apps, as well as human beings) the competing approaches can bridge the semantic gap and will enhance each other's strengths to create a new kind of adaptive, learning, collaborative and semantically rich AI.

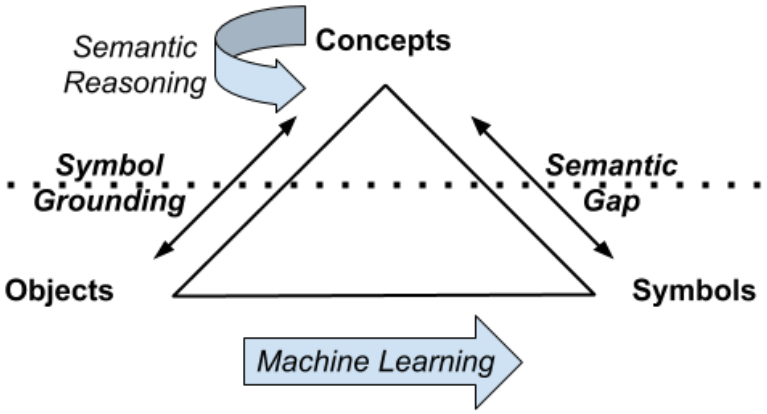

Figure 3: The Semantic Gap in the Semantic Triangle (Hybrid Agent).

In reference to Peirce's semiotic triangle[1] (cf. fig. 3), the ***semantic gap*** is the difference between a symbol and a concept. The gap between an object and a concept can be defined as ***symbol grounding***. Machine learning provides an object's data with a symbol (or label; usually, a classification or prediction), but lacks the connection to the concept that the symbol refers to. The concepts are part of semantic models (knowledge representation, ontologies), using which the symbol can be connected to a corresponding concept and, hence, ***bridge the semantic gap***. This bridge is further spun by agents interacting with their context (including other agents, humans and the physical environment), thereby sharing and updating their individual knowledge, intentions and points of view.
Hybrid AI addresses knowledge management in all phases of use: from knowledge acquisition and exploitation to sharing. Knowledge acquisition requires establishing links between learned structures and human-understandable concepts (as formulated using knowledge graphs, for example). As the knowledge of the world constantly changes, it is necessary to add, replace, update and remove (instances of) concepts and links, based on a temporal model of knowledge. Knowledge models include concepts, properties, relationships and rules. They can provide causal relationships between actions and effects [13].

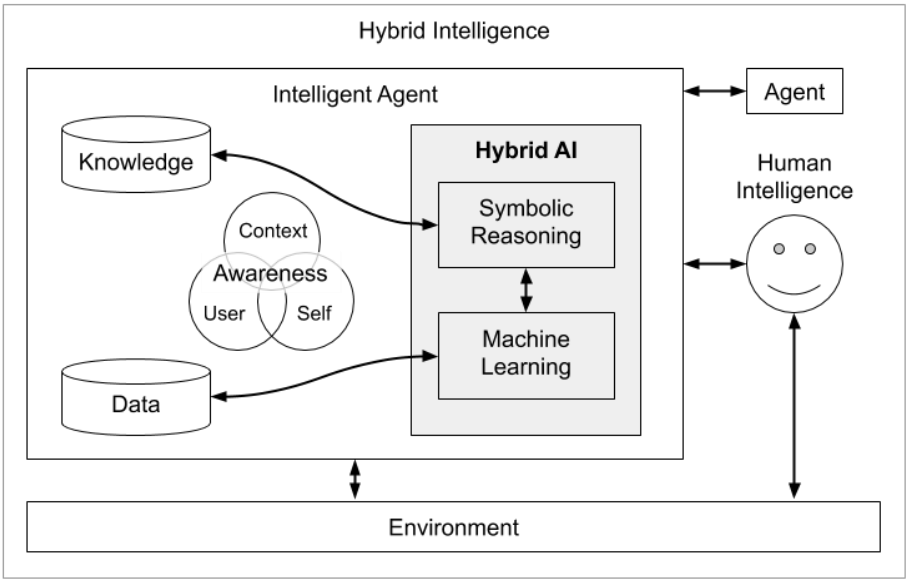

Figure 4: Hybrid AI vs. Hybrid Intelligence.
In figure 4, Hybrid AI is defined as the combination of reasoning on symbolic knowledge with machine learning from data (objects) embodied in an intelligent agent. Hybrid AI is part of systems of Hybrid Intelligence (HI), where intelligent agents learn and reason and interact with humans and other agents in relation to the environment that they are situated in. Intelligent agents need to acquire awareness about themselves, their context and users (both human and other agents).

2.2 Machine Learning

There are a multitude of techniques and methods for machine learning that have been amply described elsewhere [9, 14, 15]. Deep convolutional neural networks

[1] For example, object: a cat - symbol: "Tom" - concept: cats.

are very successful at many tasks, such as playing games, image classification, speech recognition and translation. Typically, big data is considered as the solution to overcome the bottleneck of formal knowledge. Big data proved successful in various applications, indeed. However, in many cases, vast amounts of data are not available for many reasons, such as restrictions of privacy, confidentiality, (computational) cost, time-consumption or because data does not exist, is unethical or too dangerous to collect. Humans are able to learn and generalise from little data and are able to achieve Deep Understanding, because we can conceptualise examples in ways that Deep Learning cannot.

A possible work-around for gathering data is to produce synthetic data from gaming and simulation models, which is known as Digital Twins. However, the question may be raised why the modelling knowledge required to develop those gaming and simulation environments could not be used explicitly in the learning component and to avoid the use of generated fake data.

Another possible work-around is to use transfer learning. In this type of learning, a model is trained on a general or common domain with a lot of data and re-trained on the specific domain, which requires less data. Although this is a viable method, it often does not use symbolic methods, yet.

In the context of Hybrid AI, combining learning with knowledge graphs and of learning concepts from data are of interest, for example.

2.3 Symbolic Reasoning

Reasoning can be defined as the capacity for consciously making sense of things, establishing and verifying facts, applying logic, and adapting or justifying beliefs, i.e. the ability to make inferences based on new or existing information. It is a general process of thinking rationally in order to find ‘valid’ conclusions when trying to understand sensory observations from the environment, to think about cause and effect, etc.

A reasoning process can be based on pure logic (propositional, first-order, higher-order, etc.), where the information used for the reasoning process (predicates) is represented with certainty. There are, however, many situations where we have imperfect or unknown information about the real world, e.g. due to ignorance, errors, unreliable sources or lack of observations. In those cases, we need to be able to reason with a measure of uncertainty, typically represented by probabilities, which leads to methods of probabilistic inference.

A knowledge representation in symbols serves to transfer learning from one situation to another. That is, a symbol 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 we learned in one place to a problem that we may encounter somewhere else. Combinations of symbols that express their interrelations could be called *reasoning*, and when humans string a number of signs together to express thought, it can be called *symbolic manipulation*. A plausible definition of "reasoning" [16] that covers both inference based on logic and probabilistic inference, could then be: "algebraically manipulating previously acquired knowledge (symbols) in order to answer a new question".

Reasoning may be subdivided into forms of reasoning based on forms of logic associated with the strict sense, such as deductive reasoning or inductive reasoning and more informal reasoning activities, such as intuitive reasoning or common sense reasoning. The latter employs reasoning based on experiences or common-sense knowledge, akin to the human ability to make presumptions about everyday events.

3. Benefits

The potential benefits of Hybrid AI systems lie in their increased trustworthiness, transparency, explainability, controllability and responsibility. They may facilitate the implementation of legal and ethical norms in future distributed collaborative systems in order to contribute to more human-aware systems. We elaborate on the benefits with respect to controllability, explainability and responsibility of Hybrid AI systems after explaining some common issues.

3.1 Common Issues

3.1.1 Open World

In an open world, situations and objects may emerge that have not been foreseen when the system was initially designed. To enable safe and meaningful operation in an open world, an AI agent must be able to [17]:

1. Detect that the world has changed with respect to the current world model.
2. Identify how and why the world has changed.
3. Adapt the response according to the new world model.
4. Update the model of the world.

Table 2 shows an initial concept for an open world hierarchy that was introduced by DARPA [17]. This hierarchy may not be accurate in multiple aspects, such as the order of levels, missing or redundant levels, the lack of inclusion of additional relevant dimensions (e.g., whether the novelty is local or global, the frequency with which novel situations occur, etc.), or others.

Open World Novelty Hierarchy		
Entities and Attributes	0	**Instances**: previously unseen objects or entities.
	1	**Classes**: previously unseen classes or objects or entities.
	2	**Attributes**: change in a feature of an object or entity not previously relevant to classification or action.
	3	**Representations**: change in how entities and features are specified, corresponding to transformation of dimensions or coordinate system, not necessarily spatial or temporal.
Interactive	4	**Relations** (static): change in allowed relationships between entities, such as in object-class hierarchies or adjacency (if in physical space).
	5	**Interactions** (dynamic): change in allowed relationships between entities, i.e., “rules of the game”, resulting in state changes.

	6	**Capabilities**: change in allowable actions of entities with respect to pre-conditions, post-conditions, or side-effects.
External	7	**Environment**: change in the world affecting all entities similarly, depending on factors independent of the entity.
	8	**Goals**: change in objectives of actions, especially in a multi-entity or adversarial environment.
	9	**Context**: change in meaning of actions that provides a different interpretation or narrative frame on a series of interactions.

Table 2: Open World Novelty Hierarchy.

3.1.2 *User Interaction*

Communication of humans with Hybrid AI systems is preferably bidirectional and takes the form of a dialogue. Felicity conditions for fruitful, dialogue-based communication include rich, semantic representations of subject matter. This becomes especially important when humans are facilitated to insert domain or world knowledge in an AI system. On the other hand, extracting semantic information from the eventual sub-symbolic contents (e.g. hidden layers) of an AI component is crucial for successful human-AI interfacing.

3.2 Controllability

Future AI agents will have to conduct a variety of tasks in dynamic complex environments with uncertainties and unforeseen situations. At the same time, the AI agent must adhere to safety standards and satisfy legal and ethical constraints. The implications of an unpredictable environment are that an AI agent must be able to adapt to the actual situation and learn from experience to optimise its utility for the current task. While adaptivity and learning can lead to an effective use of the capabilities and resources of an AI agent in well-known situations, it could also lead to unsafe or unethical behaviour in unforeseen situations. Controllable AI agents should therefore include appropriate safeguards to enable human intervention and control when necessary. This chapter will describe some topics in which hybrid AI methods can contribute to controllable AI agents. First, the key requirements for controllable AI agents will be defined.

3.2.1 Key Requirements

Some of the key requirements for a controllable AI agent that can be derived from the problem description are:

- **Operations in an Open World**
  A controllable AI agent must be able to operate in an open world that differs from the world for which it was initially designed.
- **User Control**
  A controllable AI agent must be able to accept and learn from the user.
- **Self-Assessment**
  A controllable AI agent must be able to assess its own performance so that the user, or the AI agent itself, can determine if it can meet the task objectives.

- **Self-Management**
  A controllable AI agent must be able to adapt its own configuration to the actual situation and learn from previous experiences to optimise its performance.
- **Explainability**
  A controllable AI agent must be able to explain to the user what it is going to do and why it is doing this. This topic is addressed below.

3.2.2 Interaction with User for Never-ending Learning

In a changing world, an AI agent will encounter new and unknown situations. TNO's clients operate in complex and changing environments. For instance, the army and police both aim for a safe society, while opponents challenge them with changing threats. Research at TNO is focused on helping them to use state-of-the-art AI tools which are capable of operating in changing conditions. A practical example is aggression in nightlife situations, where we develop video AI which detects aggression in an early stage. Every few weeks a new form of aggression is encountered that has previously not been seen. The video AI may not detect it at first; it needs human supervision to improve its model. Moreover, the normal behavioural patterns vary as well, which also lead to new, unseen situations. For instance, a garbage truck passing by during midnight, where personnel is throwing the garbage in the back of the truck, which looks similar to aggression. Interaction with the user is needed for never-ending learning. The user feedback can be considered as top-down symbolic knowledge about the world that gives meaning to bottom-up, sub-symbolic sensor data.

3.2.3 Exploiting Domain Knowledge

It is essential to exploit the available knowledge about the environment that the AI agent is operating in. In the above example of aggression detection: the aggression almost always happens in nightly hours and in one of the main nightlife areas. These facts are available in police databases. Often, a lot of contextual facts about the purpose of the AI agent are known beforehand. Using such facts to deploy the AI agent in the right circumstances increases its performance (in the case of aggression detection it leads to less false positives), and to limit its computational resources (running it during two nights on 3 cameras rather than 24/7 on all 70 cameras). The coupling of such contextual world knowledge on the one hand, and sensory interpretation on the other hand, is an important research topic for Hybrid AI systems.

3.2.4 Context Awareness

In many real-life applications, an AI model is applied within different environments. For instance, a surveillance system that detects truck cargo theft is deployed in various truck parking lots, each with its own layout and infrastructure. Ideally, it is the same AI model but deployed using a different configuration. In case of the cargo monitoring, the AI agent should know about the local layout, i.e., where the parking zone, shop and walking routes are. Similarly, contextual information about how the data is acquired is helpful, e.g., knowing the calibration of a camera enables the AI agent to make hypotheses about the size and speed of objects. Awareness about context (top-down symbolic knowledge) is essential for interpretation.

3.2.5 Context Adaptivity

Moving AI agents may enter new, unseen environments. For instance, a drone that explores a large area may enter a rural environment, while its model was optimised for an urban environment. The AI agent should signal that the environment has changed, so it can take proper actions, such as loading a model with nature classes instead of urban classes, loading another model that is known to work less accurately but more robustly in different conditions, or asking the operator for guidance.

3.2.6 Learning with Few Examples

Machine learning requires a large amount of training samples. However, in many real-life applications, few training samples are available (or even none). The two main reasons are: (a) there are few or no samples of the phenomenon (e.g., a military threat such as the placement of an improvised explosive device (IED)), or (b) it is hard or expensive to acquire many labelled training samples (e.g., observed human stress levels under a wide range of real-life stressors). There are approaches to solve these challenges, which involve the use of domain knowledge to break down the problem into smaller sub-problems, which are easier to model and to acquire labelled samples for. For instance, the IED placement is a long-term scenario which almost never is caught on camera. Another difficulty is that it can happen in many different ways, so training samples will never be completely representative. Yet, it involves short-term activities such as digging and placing the IED, for which abundant video footage is available. Classifiers for such activities can be learned, on top of which interpretation can be carried out by a sequential model such as a finite-state-machine in which the rule set is shaped by world knowledge about the phenomenon. High-level rules model the phenomenon for which few or no samples are available. Similar to such a high-level representation, this approach can also be taken for mid-level representations, such as so-called semantic features that capture parts of the phenomenon. For the example of stress assessment: the mid-level representation may be designed to consist of observable stress indicators for which sufficient labelled samples can be acquired. Again, top-down knowledge is combined with bottom-up sensory observations.

3.3 Explainability

Recent developments in AI (e.g., the recent Capsule Networks [18]) attempt to bring the sub-symbolic world closer to the symbolic world, by forcing neural networks to focus on semantic, interpretable structures in their data. While not being hybrid in itself, this approach delivers crucial ingredients for any form of Hybrid AI: a semantic *interlingua* that contributes to human-AI dialogue and increases the potential for explainability.

Hybrid AI can also address more procedural information exchange between humans and machine learning models. An example is the use of neural attention [19], where a neural network emits probabilities linked to its input layer. These probabilities explain one aspect of its computation: the parts of the data that attract most of the network's attention. Allowing humans to interact directly with this attention (e.g., calibrating the neural attention with human attention by steering it into another direction) is a low-level form of Hybrid AI, that bypasses the semantic gap and allows for a more procedural implementation of human-AI dialogue.

Human feedback loops, supported with dialogues, are important for detecting and handling bias. Bias is a pesky problem in AI, with manifestations such as *selection bias* (unfair data selection) and *inductive bias* (the formal assumptions made by an AI algorithm for predicting outcomes, like 'maximising margins between classes by Support Vector Machines is a good idea'). Exploitation of bias by adversaries occurs in *adversarial machine learning* and AI-generated data, such as *fake news*. Explaining which data underlies an AI-produced outcome, and - vice versa - assessing whether human-provided data is biased, is helpful for bias resolution. This demands a variety of information exchange types, some of which can be on quite a low level. For example, the attention mechanism described above may serve to highlight inductive bias by an AI algorithm. Analogy-based explanations (displaying training data similar to test data or its model abstractions of that data) have strong communicative and explanatory value. But it is clear that, for fine-grained, human-understandable bias assessment, semantic abstraction of this type of anecdotal data is helpful.

Human-robot communication can be characterised as shown in figure 5.

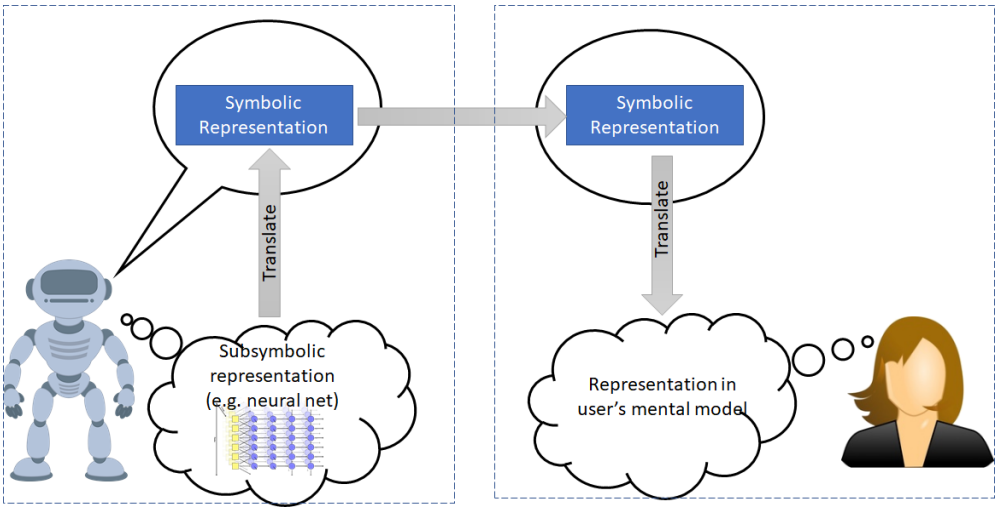

Figure 5: Human-robot communication.

We can identify the following steps in this process:

1. A robot forms an intention regarding which information to convey to a human (represented by the cloud).
2. The robot translates the information (which is represented sub-symbolically) to a symbolic representation which can be conveyed in a message (represented by the grey arrow "translate").
3. The robot pronounces the message (represented by the talk box).
4. The message travels to the receiver and arrives at the human (represented by the ellipse at the human).
5. The human translates the symbolic representation to her own representation (which is represented using her psychological mental model).
6. The human stores the message in her own head.

In this process, step 2 requires Hybrid AI technology: the message is translated from a sub-symbolic to a symbolic level. What this process teaches is that the symbolic representation should be such that the human can translate it back to her own model. Because many symbolic representations are possible, we must choose the one that fits the user's mental model best (personalised explainability).

3.3.1 Semantic Anchors

We will call the symbolic representations which are used for communication between agents and humans the Human-Agent-Teaming Communication Language (HATCL) [20]. In the communication between the autonomous system and the human-agent teaming software, there is a need for grounding the knowledge from the ontologies. If the teaming software discusses an element from the ontology with the system, or vice versa, this element needs to be translated to variables that are meaningful to either side. This process of grounding a semantic concept into variables known to the autonomous system is done through so-called "semantic anchors" that are implemented inside the autonomous system (cf. fig. 6).

A fruitful combination of semantic anchors with data-driven models demands a type of *data-auditable* AI: AI that can be explained from a data usage perspective, answering questions such as: which data is being used in the latent space of a deep learning network just prior to outputting a decision? Once these data circumscriptions are clear, external knowledge can find its way into a model.

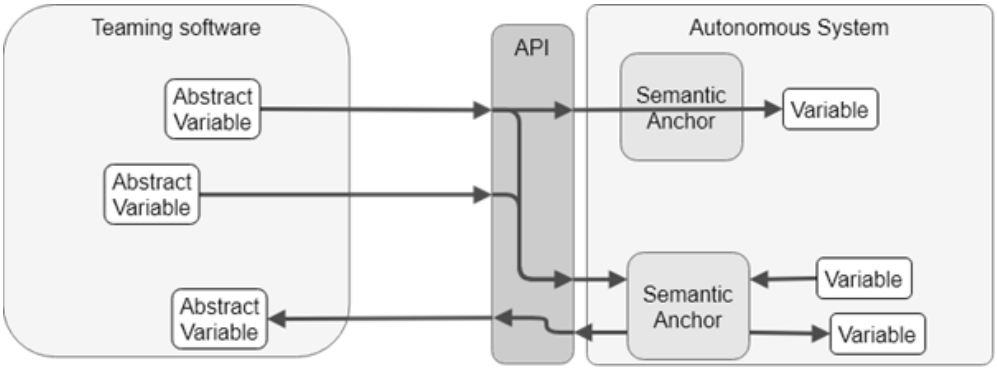

Figure 6: Semantic anchors in an autonomous system.

In general, the teaming software works at a higher abstraction level than the system. This means that semantic anchors, whose information flows from the teaming software towards the system, perform a translation into a lower abstraction level. Whereas the translation into a higher abstraction level occurs when the information flow is reversed.

An example of a simple gateway semantic anchor could be when the autonomous system contains an instance of 'car' that adheres to the definition of a car present in the ontology. As such, the level of abstraction of that instance is already appropriate to be used in the teaming software. The semantic anchor would then simply be a gateway of that exact instance, without performing any translation.

An example of a more complex semantic anchor could be when the definition of car is not known as such by the system. Perhaps the system only has a camera-feed with pixels that represent that car but not its position or velocity. The semantic anchor may then contain an operation that extracts the position of that car as well

as its velocity vector to obtain a complete description of a car according to the ontology.

3.3.2 Increasing Explainability by Structured ML Approaches

In the chapter about Controllable AI, machine learning (ML) approaches were introduced that incorporate high-level knowledge about phenomena and mid-level features of which the phenomena are composed. An advantage of such rules and features is that they can be easily explained to a user. For instance, consider a robot which is collaborating with a human. The robot has a camera and its video AI interprets that the human is waving, because the AI observes that the human's arm is moving sideways in the air. It can communicate 'Did you ask for help? I noticed that you waved your arm'. Such communication is much more human-like than using terms of changing pixels or movement in the image itself.

3.3.3 Storyboards: Visual Explanations

For the verification of a hypothesis of an AI agent, a user may want to check on which information the hypothesis was based. Of course this can be done using the lower-level data, but this is hard to interpret. AI agents that operate on images (camera, spectrograms, etc.) are able to communicate their hypotheses in the form of images, possibly with metadata such as overlays (boxes, texts, regions). For instance, a hypothesis about an observed scenario, such as IED placement (see above), may be shown to the user by one snapshot for each of the consecutive activities that were observed and linked (e.g., standing next to a road, digging, placing something in the ground, walking away).

3.4 Responsibility

Broadly speaking, responsible AI concerns the use of AI in domains with societal and personal impact. Examples are the use of recommender systems for loan applications, the use of assistive systems for the decision whether a convict should benefit from parole, or the use of an autonomous recruiting system to pre-filter applicants to job openings [21, 22].

Arguably, the three main research themes in responsible AI are fairness, confidentiality and transparency. Fairness addresses the problem that the AI models may be biased and lead to discriminating decisions. Confidentiality addresses the problem of hiding sensitive information in shared datasets. Finally, transparency addresses the problem regarding the whole process of training and applying AI systems.

Confidentiality in data processing is not primarily achieved by AI, in general. Rather, confidential responsible AI is centred on providing computational techniques and cryptographic infrastructures (e.g. multi-party computation, federated learning, homomorphic encryption), such that machine learning can operate on privacy-secured data [23, 24]. These techniques are out of the scope of this paper.

3.4.1 Fair Responsible AI

The goal of fairness is to avoid discriminating sensitive categories during decision making in a sensitive societal application domain [22]. For instance, a fair AI system should not suggest lower wages to women or should not favour specific ethnic groups in suggesting loans and insurance policy schemes.

The notion of open world, within the context of fair AI, is slightly different than the one used in for example controllable AI. We could imagine that the training task, during which an AI model is built from historical data, is seen as the closed world, whilst the deployment of an AI to reason and decide about newly unseen data is seen as the open world.

The main issue faced in fair AI is that if the historical data contains biases and the AI is trained on such data, then the AI and its decisions will be biased as well. Once deployed in the open world, the AI's biased decisions and classifications will affect society.

Moreover, the new unseen data, which will be processed in a biased manner, would subsequently become historical data at the next training phase, and thus further increase the historical bias. For instance, if a biased AI suggests that non-white people are more likely to deal drugs, then more non-white people will be inspected and possibly arrested. Consequently, their data would enrich the existing historical dataset. In this sense, the interaction between the AI and the user is mono-directional, with the user simply receiving the outcome of a classification/decision.

3.4.2 Symbolic Fairness

The vast majority of AI used in the societal domain is based on pattern recognition via machine learning techniques [25, 26, 27, 28]. These techniques are extremely capable of finding correlations between data features; this is why biases and discrimination are propagated from historical data to the AI.

There has been a large body of work addressing fairness in AI, both in "hard" technical and "soft" jurisdictional terms [29, 30]. With respect to technical fairness, the mitigation of bias and discrimination is commonly done by either obfuscating sensitive features in the data (process fairness, [24, 26, 31]), by introducing additional costs/performance metrics during training (outcome fairness, [26, 27, 28]), or by appending further decisional thresholds to a biased model (post-processing fairness, [24]). All these intervention methods necessarily rely on a quantitative definition of fairness, e.g. disparate impact [27] or statistical parity [26]. For instance, disparate treatment is a measure of fairness adopted during outcome fairness which aims at generating, for different sensitive groups, the same volume of false positive and false negative classification errors.

Technical fairness is far from being the silver bullet for fair responsible AI. Not only because two fairness measures could be orthogonal to each other – improving a specific fairness measure might diminish another one [24, 27] – but also because technical fairness reasons on the same level of the AI under investigation: data, its processing, its correlation finding, and its classification. Technical fairness assumes a generic connotation that goes across domains of applications. This context-independence of technical fairness somehow misses context-specific aspects [26].

If we consider two fair classifiers, one to decide which drug to be administered to patients and one to decide which loan scheme to be given to customers of a bank, we can state that both classifiers could be made technically fair by means of the same technical fairness approach and metrics. However, if two individuals are present in both context datasets, they could appear "equal" in different ways. It is the very notion of "equality" that goes beyond the technicalities of machine learning and its training process and goes towards concepts which become hard to encode in equations. In other words, the lack of representation of context-dependent, yet

generic, notions of equality constitutes the semantic gap of responsible fair AI [24, 25, 26].

"Soft" fairness – arising from jurisdictional social sciences research – addresses fairness on a higher level than the one within which technical fairness is situated. This is intuitive, as fairness of treatment is a concept that goes beyond AI. Laws, norms and behaviours evolve, influence each other and somehow frame societies. Whilst some aspects of soft fairness are universal, e.g. constitutional articles specifying that a country refutes any form of discrimination, others are much more context-specific, e.g. labour or financial laws.

Encoded laws and norms can be used both as technical fairness or as a constraint that the whole training/usage of AI processes should satisfy (cf. fig. 7).

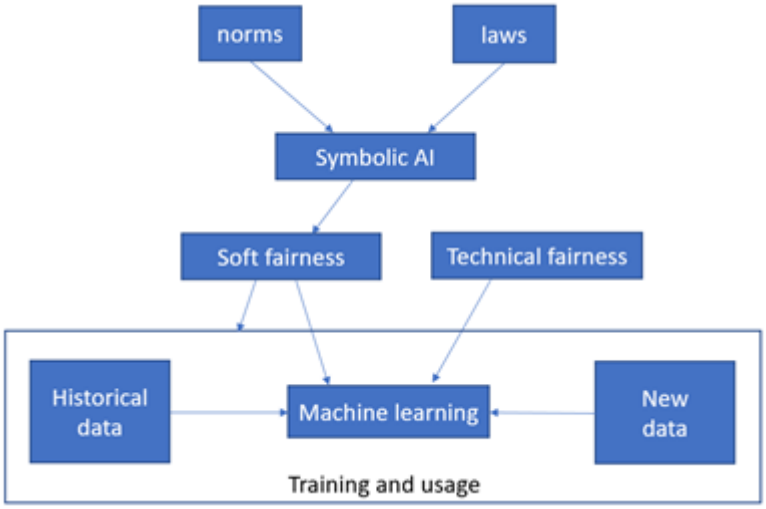

Figure 7: Encoded laws and norms.

3.4.3 Transparency

Another contribution that Hybrid AI can provide to responsible AI - more specifically causal reasoning [24, 32] - is its use around the training and usage process described in figure 7.

In a nutshell, causal graphs are graphical representations of the dependencies between random variables (e.g. input data features) and target variables (e.g. the output feature to predict in a dataset). Causal graphs could potentially be used to highlight the dependencies between sensitive attributes and output features - i.e. bringing transparency in process fairness - and, subsequently, intervene on removing unwanted dependencies efficiently. Similarly, causal graphs could be used to explain the processing of a machine learning model, in a sort of digital twin type of relationship. Such an approach could be used for instance by auditing bodies - e.g. what TNO could become in the responsible fairness domain - who would provide "non-discriminating" certifications to assistive decision making tools. Along these lines, the information represented in a causal graph could be used to effectively provide post-processing fairness measures.

Hybrid AI would then occur by learning causal graphs from the dataset or model at hand. An interesting consequence of learning causal graphs is that it would be possible to track the evolution of an unfair classifier and biased historical dataset: it

would, therefore, be possible to quantitatively study the impact the negative feedback has on the dataset and model.

4. Engineering

4.1 Knowledge Engineering

Knowledge management and engineering are essential in all phases of use of a Hybrid AI system: from knowledge acquisition to exploitation and sharing. Knowledge acquisition requires establishing links between learned structures and activations in a neural network (which are the outcomes of machine learning algorithms) and human-understandable concepts (as formulated using knowledge graphs, for example). These links should be bidirectional and continuously updated. As the knowledge of the world constantly changes, it is necessary to add, replace, update and remove (instances of) concepts, based on a temporal model of knowledge. Knowledge models include concepts, properties, relationships and rules, as well as provide causal relationships between actions and effects.

4.2 Design Patterns

Hybrid AI systems can be implemented in various ways. To be able to discuss the different forms of hybrid AI a common notation can be used. In this paper we adopt the architectural patterns of Frank van Harmelen and Annette ten Teije [33], who proposed several ways on how to combine symbolic AI with Machine Learning. Van Harmelen builds different architectural design patterns using two different elements: ovals for algorithms and boxes for their input and output. The oval algorithms can be SR for symbolic reasoning and ML for data-driven machine learning. The input and output in his scheme are symbolic relational structures (‘sym’) or data. These are presented in rectangles. Typical architectural patterns he defines are given in the figures below.

The first architectural pattern he defines (cf. fig. 8) is classical symbolic reasoning, where symbolic relations are used in combination with reasoning to derive new symbolic structures. This design pattern relates to the *first wave of AI*.

Figure 8: Classical symbolic reasoning system.

The second architectural pattern (cf. fig. 9) is classical machine learning, where data is used to train a neural network to obtain data-based results. This design pattern relates to the *second wave of AI*.

Figure 9: Classical machine learning system.

A first hybrid AI pattern is learning with domain knowledge as prior (cf. fig. 10). An example of this pattern are Logical Tensor Networks (see above), where the prior symbolic knowledge can be used to train networks with fewer training data obtaining more robustness against noise.

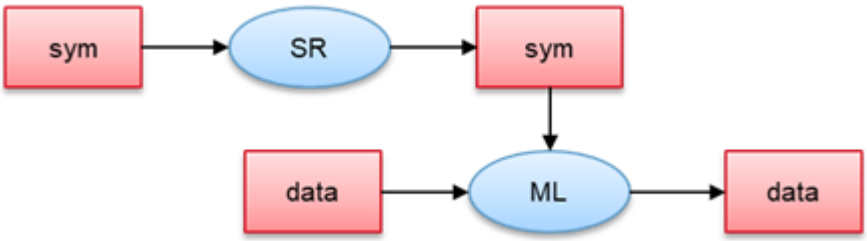

Figure 10: Learning with domain knowledge as prior.

A second hybrid pattern is ontology learning, where a symbolic structure is learned from data. This ontology is then used in the next step for reasoning (cf. fig. 11). This can for example be used for ontology learning from text.

Figure 11: Ontology learning.

Another way to combine both symbolic AI and machine learning is to use the symbolic knowledge as a prior for machine learning, such as with Logic Tensor Networks, or as a meta-reasoner over the machine learning system.

Yet another way, as described in the paper, is to intertwine the machine learning and symbolic AI in one system. These ways can obviously also be combined, for example in creating explanations.

5. Examples

In the following sections, we highlight some applications that were and are being developed at TNO. Many more will be developed in the near future for various domains, such as healthcare, smart industry, agriculture, etc.

- Drones can learn a world model (i.e., detect objects and their positions), then navigate with rule-based planning & control. Object properties are known from an ontology. In a hybrid approach locations of objects can be predicted from experience and the world model can be updated. This up-to-date world model is then used for real-time waypoints determination.

- Robotic systems consist of many components with limited resources, incl. time. Learning on the fly and adapting to a constantly changing environment is crucial. Robots can perform self-assessment to determine their level of performance, both logically and physically (wear & tear), by comparing learned properties with a reference model of themselves.

- A robotic system should have self-assessment, i.e. it can assess whether it is able to perform its task adequately. One way to tackle this is to have knowledge about itself, the configurations and current setting and it can use that knowledge to know whether it can do a certain task or not, and to take actions based on its information. A similar set-up can be true for the outer world; if a system knows about its surroundings in terms of knowledge, it helps to improve situational awareness.

- By analysing video streams, people and objects can be detected. When coupled to a knowledge model of relationships among people and objects, a situation can be understood and identified as dangerous or friendly, for example.

- The analysis of heterogeneous news sources results in a huge amount of identified persons, objects, locations and events. By integrating them into a knowledge graph and using semantic reasoning, more useful information can be extracted and understood, such as relationships between countries and political groups.

- Robotic systems consist of many components with limited resources, incl. time. Learning on the fly and adapting to a constantly changing environment is crucial. Robots can perform self-assessment to determine their level of performance, both logically and physically (wear & tear), by comparing learned properties with a reference model of themselves.

- Decision support systems have the purpose to support humans in their decisions. For example in the health domain, it is important to have explainable systems in which we have a human-in-the-loop. In these systems the human-in-the-loop helps to either improve explanations (for example in using a memory to remember which explanations were effective and which not) or update a machine learning model with the expert knowledge.

- Creating ontologies is a tedious task. NLP / ML algorithms can extract relations between words from large bodies of texts. With these relations an ontology can be created using data-driven methods. A human expert is, however, needed in the loop to judge whether the relations are correct and relevant. In the Agri domain, we automatically created ontologies comparing several algorithms and qualitatively and quantitatively measured performance. A next step is to use it as a quick start in an ontology creation session. The algorithms can also be used in other domains.

6. Conclusions

So far, we have motivated the research regarding Hybrid Artificial Intelligence and explained what the concept entails, how it could be realised and which benefits to expect, along with some example applications. It appears that there is a need for Hybrid AI for building future AI systems that fulfil the requirements of truly intelligent autonomous systems that are as flexible, understanding and context- and human-aware as we would want them to be.

Although the idea of combining various approaches of Artificial Intelligence research is certainly not a new one, it has not materialised until now. Each camp was eager to pursue their own agenda, but the time is ripe to tackle the bigger problems as we realise the shortcomings of moving on in isolation. Recently, a growing number of publications and events can be witnessed that point in the direction of Hybrid AI.

For instance, the books *Rebooting AI* [11] and *The Book of Why* [12] created a big impact on the AI community. Conferences are being organised, such as the *Combining Machine Learning and Knowledge Engineering in Practice* symposium (AAAI-MAKE), the *Cognitive Computation Symposium: Thinking Beyond Deep Learning* (CoCoSym), a Machine Learning track at the *Extended Semantic Web Conference* (ESWC) or a *Workshop on Semantic Deep Learning* at the *International Semantic Web Conference* (ISWC). In short, the concepts and implementations of Hybrid AI are getting significant attention fast.

Naturally, the concept of Hybrid AI is too broad to allow for only one solution. There are many ways of combining knowledge representation with learning and reasoning in context. The design patterns, as outlined above, are a way of describing potential combinations. They are, however, just a first step towards building HAI systems easily and efficiently. Therefore, they will need to be developed further in a method for engineering Hybrid AI systems from scenarios and requirements to prototypes and products. Finally, a set of tools should be developed for supporting the implementation of Hybrid AI systems, based on these patterns and methodology. A good toolset would enable many developers to create Hybrid AI systems based on proven approaches.

7. References

[1] Pearl, J. (2018). Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution, 1–8. https://doi.org/10.1145/3159652.3176182

[2] Garcez, A. d'Avila, Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. Retrieved from http://arxiv.org/abs/1905.06088

[3] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People, (2012), 1–58. https://doi.org/10.1017/S0140525X16001837

[4] Valiant, L. G. (1984). A theory of the learnable. Communications of the ACM, 27(11), 1134–1142. https://doi.org/10.1145/1968.1972

[5] Marcus, G. (2018). Deep Learning: A Critical Appraisal. ArXiv Preprint ArXiv:1801.00631, 1–27. Retrieved from http://arxiv.org/abs/1801.00631

[6] Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. Retrieved from http://arxiv.org/abs/1710.00794

[7] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[8] Launchbury, J. (2016). A DARPA Perspective on Artificial Intelligence. https://www.darpa.mil/attachments/AIFull.pdf

[9] Russell, S. J., Norvig, P., & Davis, E. (2010). Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River, NJ: Prentice Hall. http://aima.cs.berkeley.edu/

[10] Hutt, R. (2016). What are the 10 biggest global challenges?, World Economic Forum Annual Meeting, https://www.weforum.org/agenda/2016/01/what-are-the-10-biggest-global-challenges/

[11] Marcus, Gary, Davis, Ernest; Rebooting AI, Pantheon (September 10, 2019), ISBN-13: 978-1524748258]

[12] Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus and Giroux.

[13] Pearl, J., Mackenzie, D. (2018). The Book of Why. New York: Basic Books. ISBN: 978-0-465-09760-9

[14] Thomas M. Mitchell. 1997. Machine Learning (1 ed.). McGraw-Hill, Inc., New York, NY, USA.

[15] Y. LeCun, Y. Bengio, G. Hinton, “Deep learning”. Nature 2015, 521:436-444.

[16] Bottou, L. (2014). From machine learning to machine reasoning: An essay. Machine Learning, 94(2). https://doi.org/10.1007/s10994-013-5335-x

[17] DARPA Broad Agency Announcement Science of Artificial Intelligence and Learning for Open-world Novelty, March 2019

[18] Sabour, Sara; Frosst, Nicholas; Hinton, Geoffrey E. (2017-10-26). "Dynamic Routing Between Capsules". arXiv:1710.09829

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin (2017): Attention is all you need. https://arxiv.org/abs/1706.03762

[20] van der Vecht, B., van Diggelen, J., Peeters, M., Barnhoorn, J., & van der Waa, J. (2018). Sail: A social artificial intelligence layer for human-machine teaming. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10978 LNAI, 262–274. https://doi.org/10.1007/978-3-319-94580-4_21

[21] COMPAS software: https://www.equivant.com/compas-classification/

[22] Solon Barocas and Moritz Hardt and Arvind Narayanan, “Fairness and Machine Learning”, fairmlbook.org, 2018

[23] Archer, David W., Dan Bogdanov, Yehuda Lindell, Liina Kamm, Kurt Nielsen, Jakob Illeborg Pagter, Nigel P. Smart, and Rebecca N. Wright. "From Keys to Databases—Real-World Applications of Secure Multi-Party Computation." The Computer Journal 61, no. 12 (2018): 1749-1771.

[24] Konečný, Jakub, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. "Federated optimization: Distributed machine learning for on-device intelligence." arXiv preprint arXiv:1610.02527 (2016).

[25] Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness." In Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214-226. ACM, 2012.

[26]Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. "Learning fair representations." In International Conference on Machine Learning, pp. 325-333. 2013.

[27] Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment." In Proceedings of the 26th International Conference on World Wide Web, pp. 1171-1180. International World Wide Web Conferences Steering Committee, 2017.

[28] Verma, Sahil, and Julia Rubin. "Fairness definitions explained." In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1-7. IEEE, 2018.

[29] Hacker, Philipp. "Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law." Common Market Law Review55.4 (2018): 1143-1185.

[30] Butterworth, Michael. "The ICO and artificial intelligence: The role of fairness in the GDPR framework." Computer Law & Security Review 34.2 (2018): 257-268.

[31] Resheff, Yanai Elazar, Moni Shahar and Oren Shalom, “Privacy and Fairness in Recommender Systems via Adversarial Training”, in Proc. of the 7th International Conference on Pattern Recognition Applications and Methods (2018)

[32] Pearl, Judea. The Seven Tools of Causal Inference with Reflections on Machine Learning. Technical Report, Communications of the Association for Computing Machinery, 2018.

[33] van Harmelen, F., & ten Teije, A. (2019). A Boxology of Design Patterns for Hybrid Learning and Reasoning Systems. JOURNAL OF WEB ENGINEERING, 18(1-3), 97-123. https://doi.org/10.13052/jwe1540-9589.18133