APPL.AI

RESEARCH STRATEGY TRUSTWORTHY ADAPTIVE AI

Authors: Albert Huizing and Cor Veenman

September 2021

PREFACE

After alternating periods of impressive progress, slow acceptance and of standstill, Artificial Intelligence (AI) is now being adopted on a large scale in many areas in society. As an umbrella term for technologies like optimization, knowledge engineering, machine learning, computer vision and more, AI appeared in various ways and terms depending on the current fashion: from artificial intelligence in the early days, to data mining, big data, data science and artificial intelligence again. Successes and growing expectations were followed by deceptions and tempered expectations. Also, skepticism and fear of AI has a long history. While in the beginning loss of jobs was the main fear of the public, nowadays loss of anonymity, biased outcomes, non-transparent decision making processes and damage to democratic values and discourse feed the mistrust in AI.

Clearly, AI can have a very high impact on our personal lives in a positive and a negative way. Innovation is needed to explore, research, and develop the promises of AI: from personalized medicines, and self-driving cars to rescue robots and accelerating the energy transition. Above all, innovation is required to accommodate our fundamental values in the whole lifecycle of development and use: from non-discrimination, fairness and safety, to the right to explanation, privacy, and human control.

The current document aims to bring TNO's ambitions together in developing trustworthy AI systems, that put human values central in AI systems of the future.

EXECUTIVE SUMMARY

TNO Appl.AI is the branding for AI activities in the TNO units. Among these activities are TNO-wide science and technology research investments to further the capabilities and propositions that TNO can offer to its customers, research partners and other stakeholders. The goal of this Research Strategy document is to guide the investment decisions with a planning horizon of up to 10 – 15 years. Such decisions obviously also depend on the pace of developments outside TNO. This research strategy and an accompanying AI landscaping study will be used as input for proposals for AI research programs with specific timelines and performance goals of capabilities. These programs will be developed by TNO, whenever possible in collaboration with research partners.

AI offers an enormous economic potential in terms of improving the effectiveness and efficiency of products and services for governments and companies and AI is expected to lead to new propositions and business models. In applications that are critical with respect to safety, security or ethics, the use of AI is highly demanding. Acceptance by society will be crucial for the market prospects of AI-enabled systems. To TNO the European values with respect to trustworthy AI are therefore guiding principles.

At TNO, AI technology developments focussed on adaptive AI are driven along four dimensions that define the complexity of deployment in the real-world: open environment, multiple purpose, high risk and teaming (see Figure 1). The current document describes three roadmaps along which TNO intends to develop its Trustworthy adaptive AI capabilities. The development along these roadmaps is traceable by intermediate milestones (see Figure 2).

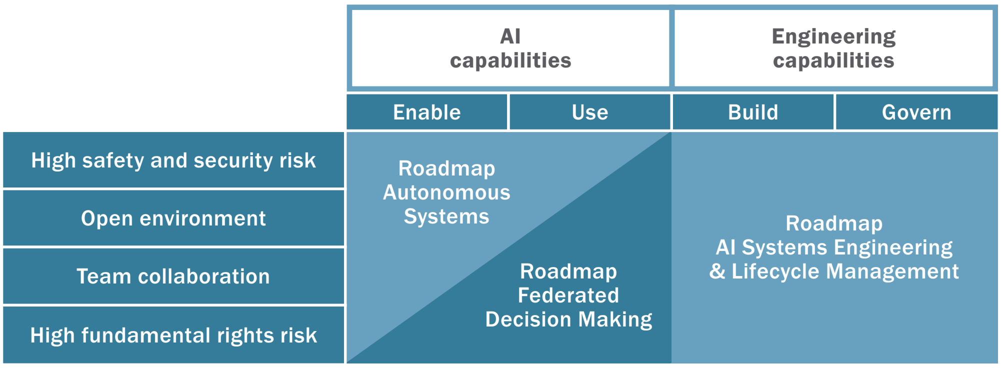

AI capabilities
Engineering capabilities
Enable
Use
Build
Govern
High safety and security risk
Open environment
Team collaboration
High fundamental rights risk
Roadmap Autonomous Systems
Roadmap Federated Decision Making
Roadmap AI Systems Engineering & Lifecycle Management

Figure 1 *Research Strategy with three roadmaps for the development of trustworthy adaptive AI to challenge the four real-world complexity dimensions. The high-risk dimension has been divided into safety and security risks and fundamental rights risks.*

ROADMAP AUTONOMOUS SYSTEMS

Autonomous systems gradually move from navigation and information collection in the environment to physical interventions and collaboration with other autonomous systems. These autonomous systems need to know their own and each other’s capabilities and limitations to balance control between the users and the systems.

ROADMAP FEDERATED DECISION MAKING

Humans making complex decisions can benefit from supportive systems. These systems should continuously and concurrently advise and learn from various types of users. Federation enables the users to remain in control and responsible.

ROADMAP AI SYSTEMS ENGINEERING & LIFECYLE MANAGEMENT

AI systems learn continuously and are thus fundamentally different to build and maintain than other (more static) software systems. This roadmap aims at developing capabilities for building, controlling, and managing AI systems with the functional capabilities developed in the other two roadmaps.

Roadmap	Milestone 1	Milestone 2	Milestone 3	Milestone 4
Autonomous Systems	Autonomous Navigation	Autonomous Information Collection	Autonomous Physical Intervention	Autonomous System Collaboration
Federated Decision Making	Reliable Decision Support	Role-based Decision Support	Trustworthy Decision Support	Federated Decision Support
AI Systems Engineering & Lifecycle Management	Engineering Trustworthy Model-based AI Systems	Engineering Trustworthy Data-driven AI Systems	Engineering Trustworthy Self-adaptive AI Systems	Engineering Trustworthy Systems of AI Systems

Figure 2 *Milestone definitions for the three roadmaps in the Appl.AI Research Strategy.*

The three roadmaps focus on capability development to support both short- and long-term programs, projects, and propositions. The outside world will have its own dynamics in challenges, limitations, and possibilities. TNO is well rooted in markets, society, government, and academics, that is, well positioned to monitor, co-design, co-develop, and adapt for future Trustworthy AI systems in high-risk application domains. In close cooperation with others TNO integrates different technologies in real life demonstrators and prototypes. Our aim is to deliver generic methods, techniques, and tools. Through publications, open source, and re-use for our stakeholders we disseminate and valorise this knowledge base.

CONTENTS

1. INTRODUCTION

1.1 APPL.AI PROGRAM

Owing to the tremendous progress in recent years, Artificial Intelligence (AI) is now a major field of investment by governments and companies with a total global corporate investment of more than USD 67.9 billion in 2020 [1]. This is also the case in the Netherlands, where organisations such as the Innovation Centre for AI (ICAI) and public-private partnerships such as the Netherlands AI Coalition (NL AIC) profile themselves with their specific instruments and programs. Recently, the Dutch government decided to reserve a budget of € 276 million from the 'national growth fund' for phase 1 of the AINED strategic investment program.

At the end of 2018, the TNO Executive Board has decided to concentrate TNO's AI research efforts in an integrated program, called Appl.AI. Appl.AI research is conducted at three levels: exploratory, capability-oriented, and application-oriented research, see Figure 1.1.

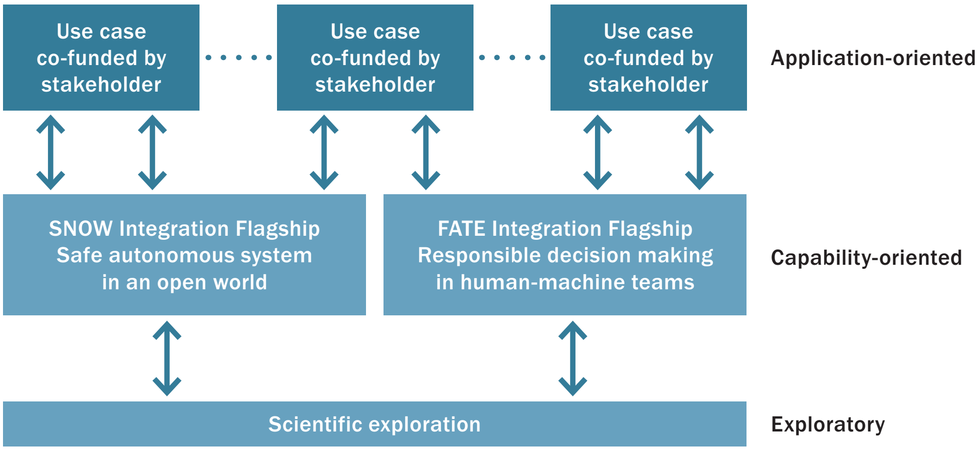

Use case co-funded by stakeholder
Use case co-funded by stakeholder
Use case co-funded by stakeholder
Application-oriented
SNOW Integration Flagship Safe autonomous system in an open world
FATE Integration Flagship Responsible decision making in human-machine teams
Capability-oriented
Scientific exploration
Exploratory

Figure 1.1 *Organisation of AI research in the Appl.AI program.*

Capability-oriented research is the core of the Appl.AI program which is currently conducted in two integration flagship projects called SNOW (Safe autoNomous system in an Open World) and FATE (Fair, Accountable, Transparent and Explainable decision making in human-machine teams). The objective of these integration flagship projects is to develop, integrate, and demonstrate generic AI capabilities. Application-oriented research is performed in use cases in collaboration with stakeholders. The use case projects in the Appl.AI program provide real-life challenges for current and future AI. Examples of use cases are diagnostics of production printers, human-machine teaming for defence applications, explainable and responsible systems for lifestyle advice of diabetes patients, bias-free skills matching for job vacancies, and automated shuttle buses. Exploratory research is primarily carried out by PhD students at universities under guidance of TNO-appointed university professors. The Appl.AI research at these three levels support, strengthen and inspire each other with exchange of information, methods, technologies, and experiences.

1.2 DOCUMENT PURPOSE AND SCOPE

The purpose of this document is to define a research strategy that will align and focus TNO investments in AI research in the Appl.AI program with the objective of creating a distinctive and defendable technology position at TNO that is valuable for TNO's stakeholders and for the TNO units involved. The goal of this research strategy is to develop capabilities for trustworthy adaptive AI systems. The main part of this research strategy consists of three roadmaps that specify milestones and capabilities for three research lines: Autonomous Systems, Federated Decision Making, and Systems Engineering & Lifecycle Management. The roadmaps Autonomous Systems and Federated Decision Making build on the results achieved in the integration flagships SNOW and FATE, respectively. The roadmap AI Systems Engineering & Lifecycle Management is a new research line that complements the other two roadmaps by focussing on engineering capabilities that are needed to build and govern trustworthy adaptive AI systems instead of AI capabilities for these systems.

The roadmaps in this document guide future TNO investment decisions with a planning horizon of up to 10 – 15 years. These decisions also depend on the pace of developments outside TNO. For example, some capabilities that are important for reaching the milestones might be made available by research partners or open sources. Specific timelines and performance goals are not included in this document since they not only depend on external developments but also on (previous) investment decisions. This research strategy and an AI landscaping study will be used as input for proposals for AI research programs with specific timelines and performance goals of capabilities that will be developed by TNO, possibly in collaboration with research partners.

It is not the purpose of this document to provide a research strategy that covers all AI-related research activities in the TNO units or to set the goals for all TNO's propositions on AI and direct all research towards these goals. The AI domain and its many application areas are too dynamic for such a formal top-down approach.

To keep the research strategy and roadmaps up to date with developments in and outside TNO, this document will be revised every year.

1.3 ROADMAPPING PROCESS

The Appl.AI research strategy was commissioned by the Appl.AI Steering Group which consists of Henk-Jan Vink, Managing Director of the TNO Unit ICT (Information and Communication Technology), and Hendrik-Jan van Veen, Director of Science, TNO unit DSS (Defence, Safety and Security). The roadmapping process was facilitated by Marcel-Paul Hasberg, TNO unit DSS [2].

The internal stakeholders of the roadmaps in this document are the TNO units as represented by their Directors of Science and Market Directors. The external stakeholders are TNO's current and future clients, co-developing knowledge organisations and funding organisations.

The core team for the creation of the research strategy consisted of
- Albert Huizing (Lead Scientist Appl.AI Program)
- Cor Veenman (Lead Scientist Appl.AI Program)

Input and comments on early versions of the research strategy were provided by the Appl.AI Management Team and Steering Group:
Judith Dijk, Frans van Ette, Anita Lieverdink, Hendrik-Jan van Veen, Henk-Jan Vink.

Parts of the roadmap AI Systems Engineering & Lifecycle Management were provided by Frank Benders (Principal Systems Engineer, Integrated Vehicle Safety) and Michael Borth (Senior Research Fellow, Embedded Systems Innovation). In addition, feedback on earlier versions of this document was provided by individual TNO scientists, by the TNO units after an online presentation, and by the Appl.AI Scientific Advisory Board. Extensive input and feedback on an advanced draft was provided by Jaap Lombaers (Director Knowledge Management and Partnerships).

1.4 ROADMAP FRAMEWORK

Figure 1.2 shows the framework for the roadmaps in this document linking technology development with user driven market needs [3]. This is an established roadmapping approach with markets setting needs and technologies driving capabilities to be able to offer propositions that satisfy those needs.

The markets and needs identified in this document are derived from the Product Market Combinations (PMCs) established in TNO's units. The propositions are the (future) products and services that TNO can offer to its customers in these markets, based on the capabilities that TNO develops along the roadmaps in this document. Since the main purpose of this research strategy is to align future technology research, the key milestones of the roadmaps concern the evolution of the TNO's capabilities over time.

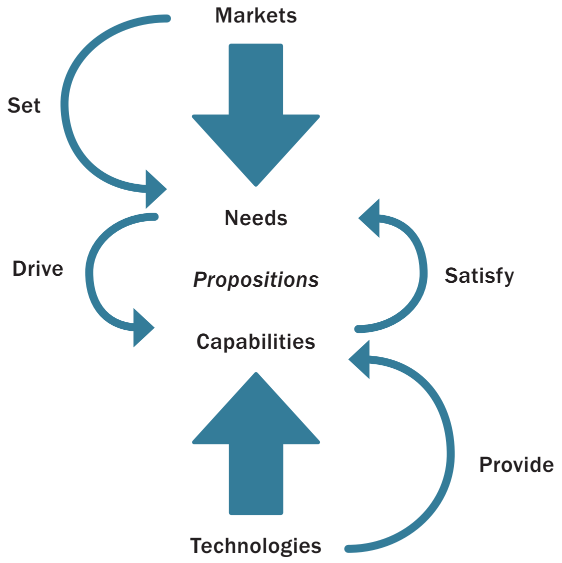


Figure 1.2 *Framework for the roadmaps.*

1.5 DOCUMENT STRUCTURE

This document is organised as follows. Chapter 2 introduces the overall vision for the research strategy and the rationale for partitioning it in three roadmaps contributing to a common goal but each with a distinct focus. Chapter 3, chapter 4, and chapter 5 define the roadmaps for Autonomous Systems, Federated Decision Making, and AI Systems Engineering & Lifecycle Management, respectively. Chapter 6 concludes the document. A list of acronyms and a glossary of terms can be found at the end of the document.

2. TRUSTWORTHY ADAPTIVE AI

This chapter first describes the need for new capabilities along four dimensions to achieve the full potential of AI while meeting the requirements for trustworthiness. To focus the development of new capabilities, we differentiate between two AI system classes that cover a wide variety of use cases. We further introduce four perspectives of different groups of stakeholders on the development of capabilities for trustworthy adaptive AI. Finally, we explain the structure of the roadmaps.

2.1 DIRECTIONS IN AI RESEARCH ALONG FOUR DIMENSIONS

AI offers an enormous economic potential in terms of improving the effectiveness and efficiency of products and services for governments and companies, but also in finding solutions to grand societal challenges such as climate change and healthcare for an ageing population [4]. In recent years, AI has achieved remarkable success in specialized tasks such as speech recognition, machine translation, the detection of tumours in medical images, and the prediction of the 3D-structure of a protein from its amino acid sequence. Despite these successes there are also some clear signs of the limitations of current AI in real-world applications. For example, biases in AI-enabled face recognition and fraud prevention have shown that prejudice in AI systems is an actual problem that must be solved. Furthermore, accidents with self-driving vehicles demonstrate that AI cannot yet be trusted in safety-critical applications.

The objective of the Appl.AI Research Strategy is to create a distinctive and defendable technology position that is valuable for TNO's stakeholders and for the TNO units involved. This will be achieved by conducting AI research along four dimensions [5]:

- **Environment**

Current AI can operate successfully in carefully controlled environments such as mass production lines in factories. However, several application domains that are being served by TNO have to deal with dynamic open environments in which environmental conditions can quickly change, unforeseen situations occur, and, as a consequence, limited amounts of (labelled) data are available for training machine learning algorithms. By developing capabilities that enable AI to operate effectively and safely in open environments new opportunities can be created for businesses and government to improve their products and services.

- **Purpose**

One could design an AI system for a specific purpose and tune decision rules or parameters until the AI system performs satisfactorily. It will however save development costs and time-to-market to have a more generic approach in which the goal of a task as intended by the user can be transferred to the AI system and the AI system uses its generic problem-solving capability to optimally execute the task. The specification of this goal needs to capture not only the intentions and preferences of the user, but also operational constraints, laws and regulations, ethics, and societal values.

- **Risk**

No human and no AI system is perfect. There is always a probability of errors that for instance may be caused by difficult circumstances or malfunctioning hardware or software. In AI applications such as systems providing online product recommendation the impact of errors on humans is generally small. Many companies commercially exploit these low-risk applications with the current generation of AI. However, there also is an increasing number of cases in which the application of AI has led to serious concerns about high risks of physical and mental harm and negative impact on fundamental rights including human dignity, non-discrimination, and protection of privacy. The EU recognizes the importance of addressing this problem and has published a proposal for an AI Regulation with a risk-based approach [6]. When in force, providers of high-risk AI systems in the EU will need to comply, otherwise administrative fines are incurred up to 30 M€ or 6 % of global annual turnover, whichever is higher. Prospective providers of high-risk AI systems will therefore need AI technologies, methods, and management processes that conform with this AI Regulation.

- **Collaboration**

Current AI systems mainly interact with humans and other systems in a pre-determined way acting like smart stand-alone tools employed to solve specific problems. This predetermined interaction and fixed task allocation between humans and AI tools helps to manage expectations and minimize risks, but it also limits the effectiveness of combined human intelligence and machine intelligence in complex and dynamic environments. Effective collaboration between humans and AI systems 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 systems that learn to understand and interpret human abilities, and for self-improving forms of collaboration.

Figure 2.1 summarizes the general direction in AI research along four dimensions that TNO will follow to build a distinctive technology position.

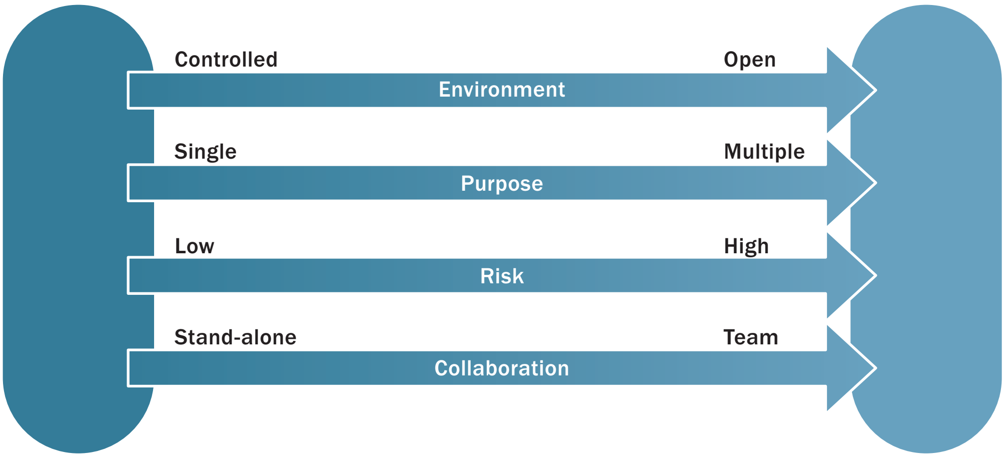


Figure 2.1 *AI research direction along four dimensions.*

2.2 TRUSTWORTHY ADAPTIVE AI

The previous section described the direction for AI research in TNO along four dimensions:

- From a controlled environment to an open environment
- From single purpose to multiple purpose
- From low-risk applications to high-risk applications
- From stand-alone to collaboration in a team of humans and AI systems

The first question then is what the general properties of an AI system should be to meet this overall research direction. To enable operation in an open environment an AI system should be able to adapt to changing environmental conditions and unexpected situations. Furthermore, AI systems should be able to adapt to different user requirements and purposes without a major redesign effort. In addition, an AI system must be able to assess and manage risks in complex dynamic situations. Finally, the transition from stand-alone operations to operations in a team also requires a capability to adapt to different team compositions.

Adaptive AI systems that during operation adjust themselves to different environments, purposes, risks, and team compositions will generally perform better and more efficiently than fixed and rigid AI systems that do not adapt themselves. However, the drawback of adaptivity during operations is that the behaviour of an adaptive AI system is more difficult to direct, predict and explain to the user. Directability, predictability and explainability of the behaviour of an AI system are crucial to users to gain trust in AI systems [7]. This particularly is the case for high-risk applications of AI systems where errors may have a big impact on people's lives, livelihoods, and fundamentals rights. This need for trustworthiness has been recognized in Europe by the High-Level Expert Group (HLEG) on AI as reflected in the ethics guidelines for trustworthy AI that were published in 2019 [8].

In summary, TNO aims to build a distinctive technology position by initiating research on trustworthy adaptive AI systems for multiple purposes involving high risks, operating in teams in open environments.

2.3 AI SYSTEM CLASSES

To focus our research of trustworthy adaptive AI, we propose to differentiate between two AI system classes that cover a wide variety of use cases: autonomous systems that can act without constant human interaction in an open environment, and systems for federated decision making in human-machine teams, see Figure 2.2. These two system classes roughly correspond with the two main categories of high-risk AI systems (AI systems intended to be used as safety component of products, and AI systems with mainly fundamental rights implications) that have been identified in the proposal for AI regulation in the EU [6].

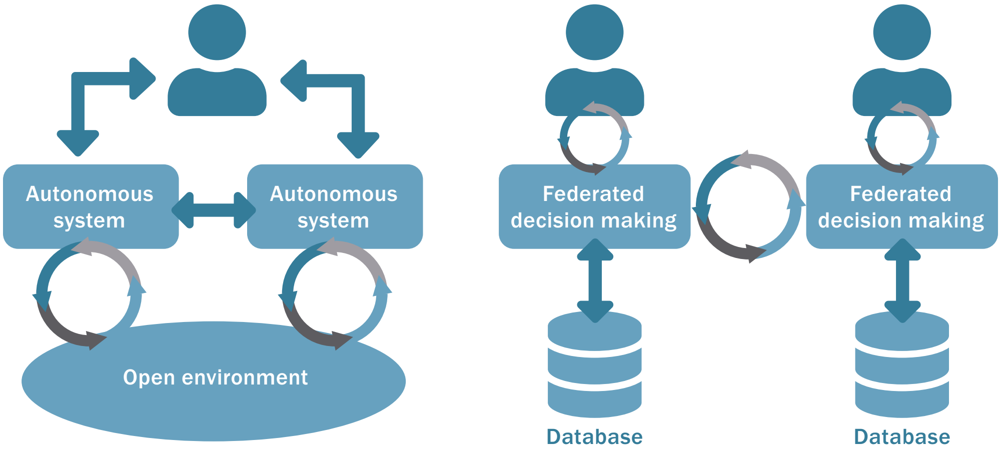


Figure 2.2 *Two AI system classes: (left) autonomous systems acting under human oversight in an open environment and (right) systems for federated decision making in a human-machine team.*

Autonomous systems such as automated vehicles and robots are employed to reduce the risk of physical injuries to humans, in applications where reaction time is critical and in jobs where skilled human workers are scarce. The challenge of AI for autonomous systems in an open environment is to conduct various tasks effectively and safely without direct human intervention for an extended period of time. The second system class exploits the complementary capabilities of humans and AI systems in decision making to become more effective at conducting tasks while at the same time assuring compliance with laws, ethics, and societal values. To achieve this, such an AI system needs to collaborate with its users and other AI systems. The difference in focus for AI in these two system classes is illustrated by Figure 2.2 with AI for autonomous systems focusing on the interaction loop of these systems with their environment and AI for federated decision making concentrating on the interaction between humans and systems, and the interaction between systems.

2.4 STAKEHOLDER PERSPECTIVES

Much of the attention in scientific journals, conferences, and media coverage of AI is devoted to technologies (e.g., algorithms, knowledge representations, data sets, and computing hardware) that (potentially) enable new functionalities or improved performance of products and services. However, technologies rarely are enough to create significant benefits and it is not sufficient for research in trustworthy adaptive AI to only view it from the perspective of technology suppliers at universities, research and technology organisations, and start-up companies. The proposed AI Regulation clearly indicates the need for considerable attention to and research into the roles of system providers that build AI systems to place on the market or to put into service, of users that employ AI for professional or personal purposes, and of authorities that are responsible for the governance of trustworthy AI during its full lifecycle. Figure 2.3 illustrates these four perspectives (build, enable, use, and govern) and the associated groups of stakeholders (system providers, technology suppliers, system users, and authorities). The stakeholder group that uses AI systems asks for functional capabilities and the stakeholder group of enablers creates and supplies these functional capabilities. The stakeholder group that builds AI systems manages and executes the processes that develop and maintain AI systems during their lifecycle. Oversight of the management and execution of the development and maintenance of AI systems is provided by the stakeholder group with a governance role.

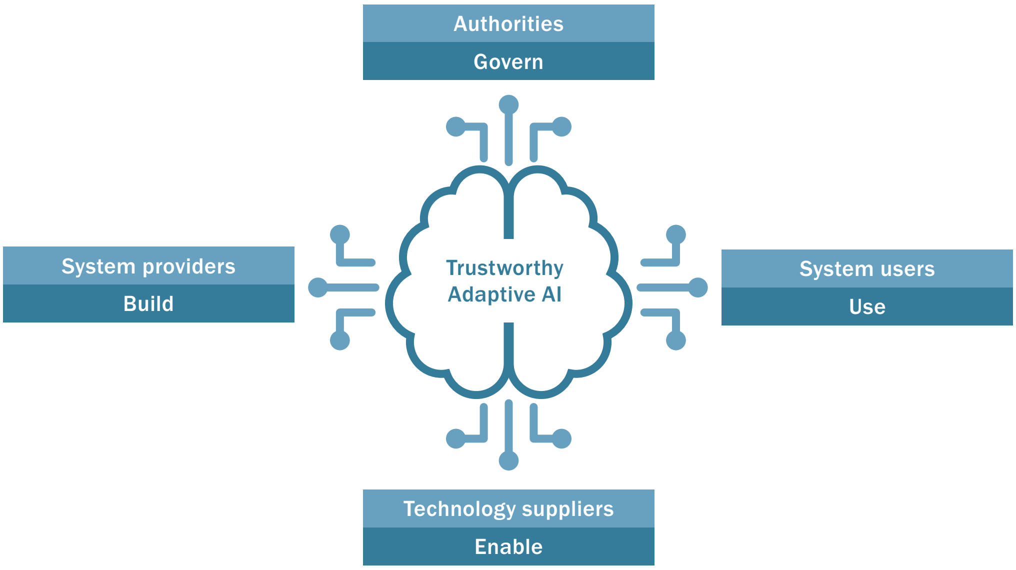


Figure 2.3 *Building, enabling, using, and governing trustworthy adaptive AI involves groups of stakeholders with different roles and responsibilities.*

The development of trustworthy adaptive AI according to these different perspectives requires researchers from many different scientific and engineering disciplines (alpha, beta, and gamma sciences) to work together. Fortunately, TNO employs researchers with backgrounds in many of the required disciplines, which makes TNO well positioned to overcome the challenge of developing trustworthy adaptive AI.

2.5 ROADMAPS

To create focus and mass in AI research at TNO, three roadmaps with a common goal of trustworthy adaptive AI systems are defined in the following chapters:

- Roadmap Autonomous Systems
- Roadmap Federated Decision Making
- Roadmap AI Systems Engineering & Lifecycle Management

Figure 2.4 shows that these three roadmaps are distinguished by their initial focus on different dimensions (purpose, environment, risk, collaboration), stakeholder perspective (enable, use, build, govern), and system classes (autonomous systems and federated decision making). The roadmap Autonomous Systems is primarily focussed on research of AI capabilities for multipurpose autonomous systems in an open environment with high safety and security risks. The roadmap Federated Decision Making is primarily focused on research of AI capabilities for systems that reduce the risk of unfair, biased, and non-transparent decisions affecting human livelihoods and fundamental rights. The roadmap AI Systems Engineering & Lifecycle Management concerns the research of engineering capabilities involving new methods, processes, and policies that are needed to build and govern trustworthy adaptive AI systems (autonomous systems as well as system for federated decision making).

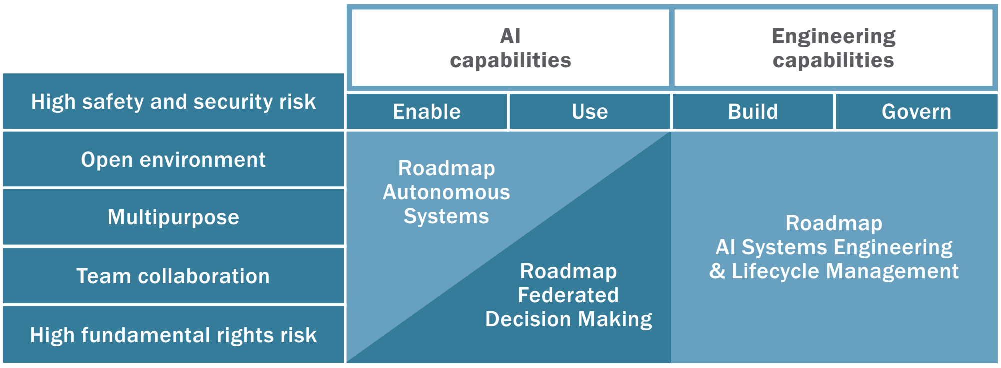

AI capabilities
Engineering capabilities
Enable
Use
Build
Govern
High safety and security risk
Open environment
Multipurpose
Team collaboration
High fundamental rights risk
Roadmap Autonomous Systems
Roadmap Federated Decision Making
Roadmap AI Systems Engineering & Lifecycle Management

Figure 2.4 *Appl.AI research strategy with three roadmaps for the development of trustworthy adaptive AI.*

For each of the three roadmaps, described in detail in Chapters 3-5, a set of four milestones and topics for the development of capabilities towards the objective of trustworthy adaptive AI has been defined, see Figure 2.5. A milestone indicates a moment in time at which several capabilities of a sufficient maturity level are integrated and demonstrated to show the progress in trustworthy adaptive AI. The sequence of milestones represents the research and development effort that is needed to reach these. Some capabilities may take much longer to develop than others and research and development periods may therefore overlap. Furthermore, capabilities generally do not contribute to just a single milestone but to subsequent milestones. Please note that Figure 2.5 should not be interpreted as a Gantt chart with dependency relationships between capabilities, but rather as a representation of the proposed topic sequence.

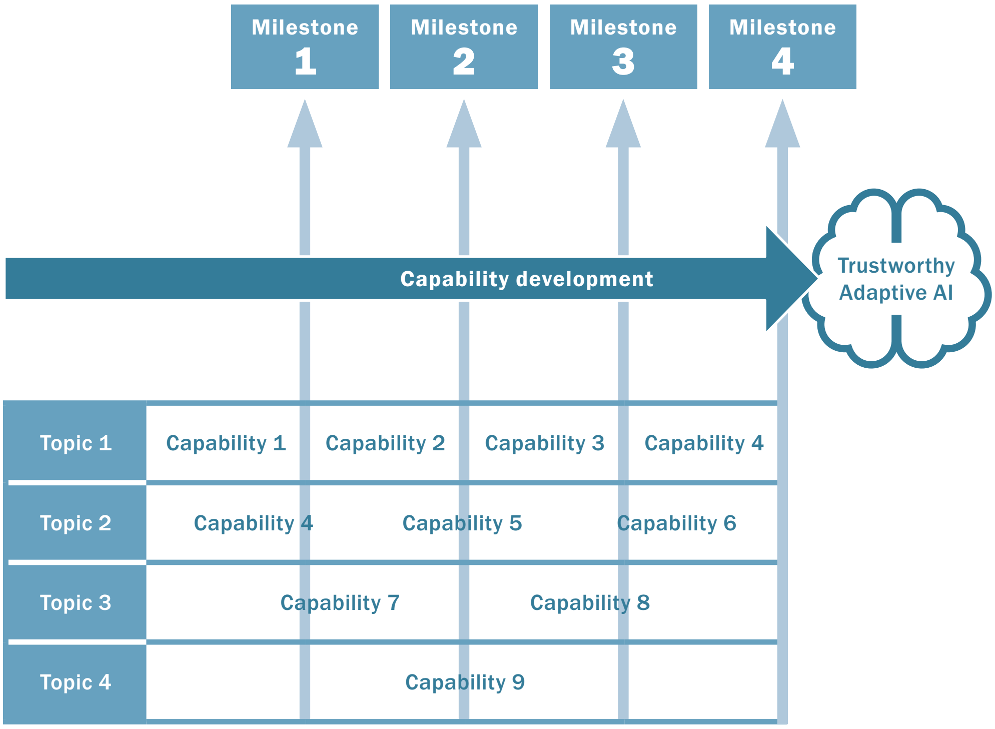


Figure 2.5 *Milestones and topics for the development of capabilities in a roadmap.*

3. ROADMAP AUTONOMOUS SYSTEMS

The roadmap Autonomous Systems specifies research of AI capabilities for autonomous systems such as vehicles, robots, and smart infrastructure, enabling these systems to conduct tasks safely, securely, and effectively in an open world without the need for constant interaction with a human operator. In the roadmap description below, we first identify the market needs for these systems and then define and elaborate on milestones for this research. Next, we describe the capabilities that contribute to these milestones.

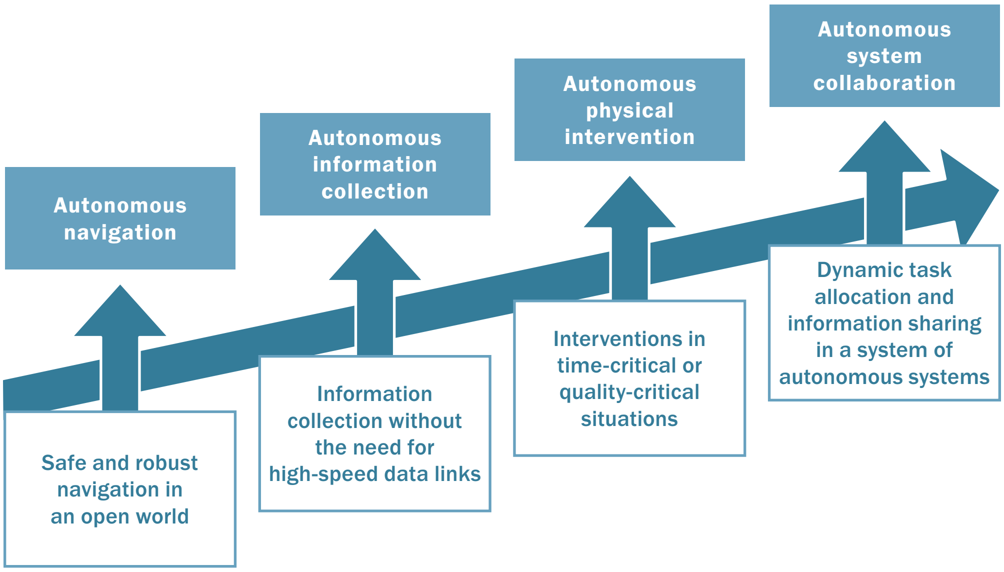
3.1 MARKET NEEDS

In several markets and application domains served by TNO, stakeholders foresee an important role for autonomous systems that can operate intelligently in an open environment under real-world conditions. The tasks that autonomous systems must be able to conduct can generally be categorised as dull, dirty, or dangerous. Depending on the application domain, expected benefits from the use of autonomous systems are:

- To keep people safe
  - Reduce physical harm and mental stress
  - Bring fewer humans in hazardous situations
- To decrease operator workload and to maintain or increase performance
  - Increase operator endurance
  - Extend operator capabilities
  - Improve utilization of equipment
  - Improve quality
- To operate remotely
  - Extend or cope with communication limits
  - Operate in remote environments, i.e., provide expertise at a distance

3.2 MILESTONES

Figure 3.1 shows the milestones for the roadmap Autonomous Systems. The roadmap starts with a milestone for autonomous systems that can navigate autonomously in an open environment, then progresses to autonomous information collection and next to physical interventions and finishing with the long-term goal of collaboration with other autonomous systems. It must be emphasized that it is not the primary goal of the roadmap to develop autonomous systems that permanently act without interaction with an operator. Instead, the goal of this roadmap is to enable autonomous systems to act without continuous operator interaction, i.e., to reduce the burden on human operators by increasing self-sufficiency of systems [9]. It is the decision and responsibility of its human operator to set the level to which an autonomous system is allowed to autonomously navigate, to collect information, to intervene in situations, or to collaborate with other autonomous systems.



Autonomous navigation
Safe and robust navigation in an open world
Autonomous information collection
Information collection without the need for high-speed data links
Autonomous physical intervention
Interventions in time-critical or quality-critical situations
Autonomous system collaboration
Dynamic task allocation and information sharing in a system of autonomous systems

Figure 3.1 *Milestones of the roadmap Autonomous Systems.*

3.2.1 AUTONOMOUS NAVIGATION

Autonomous systems such as vehicles and robots that are remotely controlled by an operator have existed for more than a century. Typically, operators control the actuators on such mobile systems remotely through a wired or wireless connection. By maintaining a safe distance between the remote system and the operator, the operator can stay at a safe distance from hazardous conditions that are for example due to the presence of toxic gases or explosive substances. This type of remotely controlled mobile systems requires continuous attention of a skilled operator. Developments in global positioning systems (GPS), sensors, and autopilots have enabled waypoint navigation in which the mobile system travels autonomously along a pre-programmed or remotely controlled path without the need for a remote operator to continuously control the actuators. Using information from onboard sensors, the mobile system automatically avoids collisions with obstacles. However, such a system still requires the remote operator to control the sensors and actuators that are needed for tasks beyond mere navigation, such as information collection and intervention tasks that may be much more complex than navigation itself.

Applications of mobile systems with autonomous navigation capabilities include:

- Transportation of persons and goods
- Inspection and monitoring of factories and building sites
- Telepresence

Challenges for mobile systems with autonomous navigation capabilities include robust navigation in unstructured environments, dealing with communications latency and limited data rates, enabling situation awareness for the remote operator, and avoidance of collisions with moving objects and persons.

3.2.2 AUTONOMOUS INFORMATION COLLECTION

The next milestone in the capability development of autonomous systems is to enhance a mobile system with an autonomous information collection capability. This implies that in addition to avoiding obstacles, the mobile system can also plan its trajectory and control its sensors in such a way that the required information is gathered by the mobile system. If physical intervention or social interaction with humans in the environment is needed, this is conducted remotely by an operator. The benefit of an autonomous information collection capability is a significant reduction of the operator workload in controlling the sensors and the trajectory of the mobile system when compared with the workload using a mobile system that only has autonomous navigation capabilities. This allows a single operator to supervise multiple mobile systems simultaneously which for instance enables cost reduction or a faster execution of the mission.

Applications of mobile systems with autonomous information collection capabilities include:

- Inspection of industrial plants and offshore platforms
- Area, perimeter, or building surveillance
- Reconnaissance of hazardous sites
- Contraband detection

Challenges for mobile systems with autonomous information collection capabilities include specification of which information to collect, recognition of the operational context, and dealing with novel objects and situations.

3.2.3 AUTONOMOUS PHYSICAL INTERVENTION

Mobile systems with autonomous information collection capabilities generally do not change their operational environment as they are passive observers. For tasks that require the system to make physical changes in the real world without the possibility or need of actuator control by an operator, e.g., due to the absence of a communication channel, systems are needed with an autonomous physical intervention capability. Applications include:

- Remote maintenance and repair of equipment
- Bomb disposal

Challenges for systems with autonomous physical intervention capabilities include the specification of intervention tasks and learning to manipulate novel objects or to manipulate known objects in novel ways.

3.2.4 AUTONOMOUS SYSTEM COLLABORATION

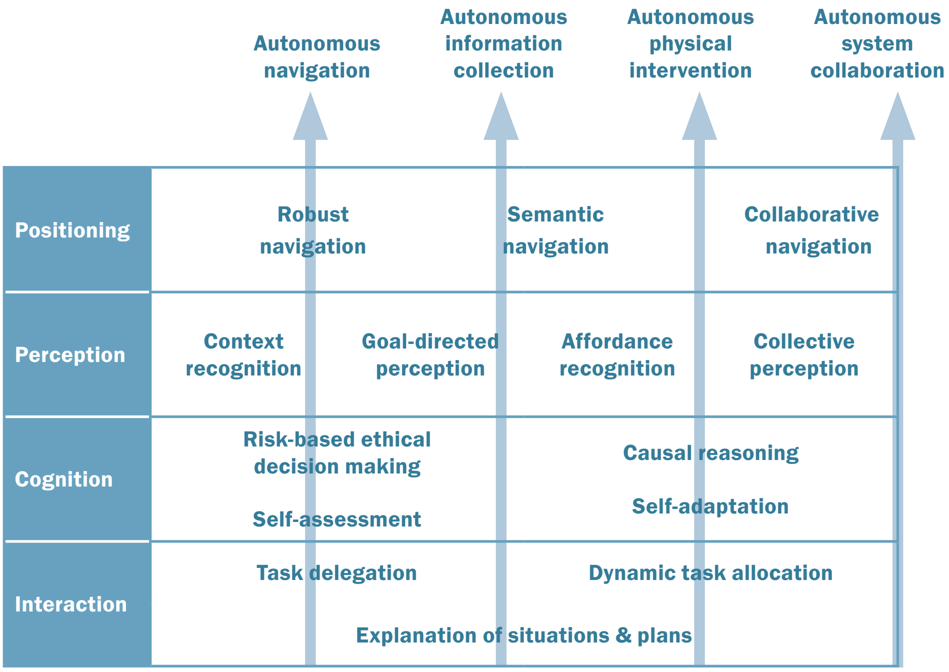
Using multiple autonomous systems with similar or differing capabilities can accelerate task execution in time-critical situations or improve performance and safety due to a better situation awareness and more diverse system capabilities. This requires multiple (heterogeneous) autonomous systems to collaborate when needed and to act independently of each other when collaboration is not feasible. Applications of collaborative autonomous systems include:

- Connected and cooperative automated driving
- Search and rescue

Challenges for collaborative autonomous systems include dynamic allocation of tasks in a changing environment, coordination of collaborative and joint tasks, and sharing of relevant information in environments with limited communications bandwidth.

3.3 CAPABILITIES

Figure 3.2 shows capabilities that contribute to the milestones shown in Figure 3.1. These capabilities generally do not contribute to just a single milestone but also to subsequent milestones.



	Autonomous navigation	Autonomous information collection	Autonomous physical intervention	Autonomous system collaboration
Positioning	Robust navigation		Semantic navigation	Collaborative navigation
Perception	Context recognition	Goal-directed perception	Affordance recognition	Collective perception
Cognition	Risk-based ethical decision making Self-assessment		Causal reasoning Self-adaptation	
Interaction	Task delegation Explanation of situations & plans		Dynamic task allocation	

Figure 3.2 *AI capabilities that contribute to the milestones in the roadmap Autonomous Systems.*

3.3.1 NAVIGATION CAPABILITIES

Robust navigation

Autonomous vehicles and robots that operate outside use GPS and maps to locate themselves and plan routes and trajectories. However, in unstructured environments, built-up areas and inside buildings autonomous systems cannot always rely on maps and GPS for accurate positioning, and they must have other ways of navigating. A well-established technique for navigating without maps and GPS in buildings is Simultaneous Localization and Mapping (SLAM) that solves the two-fold problem of mapping of an unknown environment and self-localization in this environment using camera images [10]. However, other sensors may have to be used to complement optical cameras in environments where visibility is poor. Fusing information from multiple sensors with varying degrees of accuracy is one of the main challenges for robust navigation.

Semantic navigation

Semantic navigation capabilities enable an operator to direct an autonomous system to a location that is described semantically (with natural language) instead of geometrically (by its geographical position) [11]. For example: ‘go to the right at the end of the corridor and enter the third room on the left’. In semantic navigation, a semantic map of the environment is combined with a geometric map. Semantic mapping involves complex semantic concepts of the environment, such as objects and their relations, by putting together knowledge representation and knowledge inference.

Collaborative navigation
Collaborative navigation concerns multiple autonomous systems orchestrating their localisation and navigation [12]. For collaborative navigation, timely and effective information exchange between autonomous systems is required. This information for example consists of estimated and predicted accuracy of navigation, and estimated control delays. The navigation information that is provided by other autonomous systems is combined with the navigation information of the autonomous system itself to compose improved semantic and geometric maps.

3.3.2 PERCEPTION CAPABILITIES

Context recognition
For safe and effective execution of tasks in a complex dynamic environment, autonomous systems must be able to recognize the operational context [13]. Since a change in context may have a considerable impact on safety, security, and performance, it may require adaptations in the planned course of action and the internal configuration of the autonomous system. Deriving a succinct and informative characterization of the operational context from sensor observations and data of the environment is difficult because of the high-dimensional nature of sensor data such as camera images, non-linearities due to saturation of receivers, thermal noise, and interference.

Goal-directed perception
To collect information in an open environment, an autonomous system must be able to locate and recognize objects with an accuracy depending on the goal of the task. If at any point objects cannot be localized or recognized with the required accuracy, the system must change its plan to achieve the goal [14]. For example, if an object cannot be recognized with sufficient accuracy because of a low resolution or occlusion by obstacles, the autonomous system may decide to move to get closer to the object, or to change the line-of-sight to get an unobstructed view of the object.

Affordance recognition
In open environments, mobile systems will encounter novel objects that may need to be manipulated. The recognition of object affordance (i.e., how an object can be used) is a key capability to be able to manipulate objects successfully, particularly when these objects are closely spaced [15]. Segmentation of the environment in different objects based on sensor data such as camera images is needed to enable the recognition of object affordance.

Collective perception
In a system of autonomous systems, the position and sensors settings of each autonomous system can be optimized with respect to the joint goal of information collection. With collective perception capabilities, sensor data from multiple positions and perspectives can be combined to extract more and higher quality information from the environment than can be obtained with a single autonomous system [16]. Furthermore, computational power for demanding perception functions that is needed but not available in one autonomous system may be provided by the other autonomous systems. Collective perception capabilities will need an advanced strategy for information exchange and communications to reduce the need for high-speed data links with limited and predictable latencies.

3.3.3 COGNITION CAPABILITIES

Risk-based ethical decision making
In environments where autonomous systems and humans interact, an autonomous system may have to choose between courses of action that could lead to different (and possibly even life-threatening) consequences for humans. A well-known example is the ‘trolley problem’ where a self-driving vehicle must make a moral choice between two options that lead to human injuries or deaths of different people [17]. Although the trolley problem is useful in highlighting ethical dilemmas that autonomous systems may face, it assumes completely certain outcomes of courses of action which is never the case in the real world.
Risk-based ethical decision making is a promising approach for autonomous systems, particularly for high-risk applications [18]. Here, risk is defined as the product of the probability and the severity of an unwanted event [19]. Augmented utilitarianism is an approach for risk-based ethical decision making that enables autonomous systems to make ethical decisions by computing risks during run-time [20].

Causal reasoning
Physical interventions in hazardous conditions carry an inherent safety risk. For example, opening the wrong valve in a chemical plant can lead to explosions or the release of toxic gases in the atmosphere. Reasoning about the cause of events in the environment and the effect of physical interventions in that environment is an important precondition for safe deployment of systems with an autonomous physical intervention capability [21]. Collaboration in a system of autonomous systems will also benefit from a causal reasoning capability because this will help to anticipate the behaviour of other autonomous systems in the system.

Self-assessment
A mobile system that navigates autonomously in an open world should deal with unexpected situations in which the system is not able to conduct the task satisfactorily, i.e., the system is not competent in that situation. It is therefore important that the autonomous system can assess its own performance and internal health state (self-assessment) during operations and compare it to the expected performance [22]. If there is a significant observed and/or predicted discrepancy in performance, the autonomous system should inform the operator well in advance and, if needed, transit to a safe state to avoid unsafe situations.

Self-adaptation
Changes of the goal, the task constraints or the environment may during operation require change of the system configuration or of other settings of the system. Moreover, errors or malfunctioning components that have been diagnosed by the self-assessment capability of an autonomous system may also require change of system configurations or settings. An autonomous system should therefore have a self-adaptation capability optimizing the system configuration and settings based on the goal and constraints of the task and the internal health state [23].

3.3.4 INTERACTION CAPABILITIES

Task delegation
To enable an operator to assign an information collection task to an autonomous system without detailed specification of how the task should be conducted, the system should be able to interpret the goal and constraints of the task at an abstraction level that is easily understood by the operator. For this purpose, the operator and the autonomous system should share a common language and a knowledge model or ontology representing their common understanding of the operational environment, resources, goals, tasks, and constraints [24].

Dynamic task allocation
When an autonomous system physically intervenes in a dynamic environment under supervision of a human operator, the allocation of tasks between the operator and the system must constantly be assessed. Dynamic task allocation capabilities include monitoring the performance and workload of the operator and the autonomous system, and dynamically re-allocating tasks if needed.
An important challenge in this adaptive autonomy approach is to decide when a task needs to be re-allocated because this depends not only on the situation but also on the skills and capabilities of the operator and the autonomous system [25]. In a system of autonomous systems, dynamic task allocation becomes even more important because of the different and changing conditions, safety risks, and workloads that each of the autonomous systems may be confronted with.

Explanation of situations and plans
An important characteristic of many of today's successful AI algorithms is that it is very difficult, even for AI experts, to understand how they arrive at a certain decision, or proposed a specific course of action and not another one. Autonomous systems should be capable of explaining to the operator in understandable language the situation and how they arrived at their plans [26]. This explanation capability will gradually build the trust of the operator that the system knows in which situation it is and why it is proposing certain plans.

4. ROADMAP FEDERATED DECISION MAKING

The roadmap Federated Decision Making (FDM) develops interactions between users and AI systems for high-risk applications [6]. In the roadmap description below, we first identify the market needs. We define and elaborate on milestone systems as steps towards multiple systems that concurrently and continuously learn with and from their respective users and each other enabling federated decision making. We then describe the capabilities needed to enable the milestone systems.

4.1 MARKET NEEDS

Decision support systems (DSS) are used for data-intensive tasks involving high risk decisions. The market needs systems that are:

- Less labour intensive
- More reliable
- More flexible
- More effective

More specifically, there is a need for decision support systems that enable

- Learning with partially labelled and biased datasets
- Learning with multiple federated datasets and users
- Online robust and adaptive learning
- Explanation to various types of users
- Adoption of ethics, legal and policy conditions

4.2 MILESTONES

Figure 4.1 shows the milestones of the roadmap Federated Decision Making.

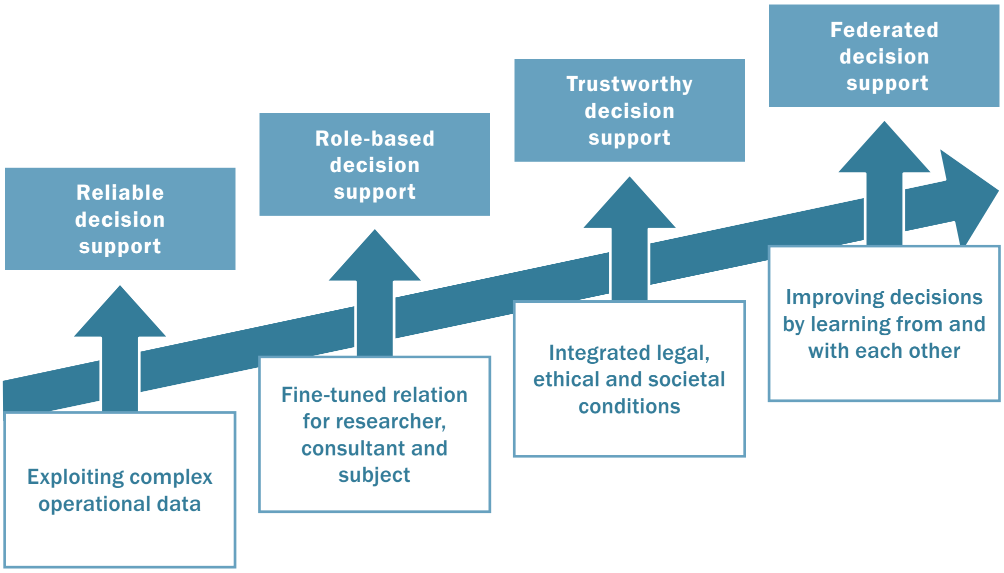

Reliable decision support
Role-based decision support
Trustworthy decision support
Federated decision support
Exploiting complex operational data
Fine-tuned relation for researcher, consultant and subject
Integrated legal, ethical and societal conditions
Improving decisions by learning from and with each other

Figure 4.1 *Milestones of the roadmap Federated Decision Making.*

4.2.1 RELIABLE DECISION SUPPORT

The AI system uses data to learn and support complex decision processes. It exploits complex dependencies between predictors that humans cannot oversee, while identifying the most important predictors. It is challenging that nowadays the best performing AI systems mostly are still black boxes. Moreover, training datasets may have specific biased distributions, which become problematic under certain use case conditions. Decision support systems have a long history of application in industry, marketing, and medicine. Governmental organizations are increasingly using decision support systems in applications with high demands regarding Ethical Legal and Societal Aspects (ELSA).

4.2.2 ROLE-BASED DECISION SUPPORT

Role-based systems differentiate between types of users being (domain) researchers, consultants and (data) subjects in supporting decision making, where fairness, transparency, and data confidentiality are key factors. Depending on their role, users have the ability to configure the system, to obtain additional and fine-tuned information and to have a continuous learning relation. Obtaining useful knowledge and enforcing context conditions even using black-box models should sacrifice only limited system performance. Examples of this class of systems include systems to support researchers who want to gain new knowledge in their domain, such as systems for discovering new cyber threats, and systems for fine grained optimization of energy demand and supply, and to support consultants in recruitment, and law enforcement, and subjects in preventive health apps.

4.2.3 TRUSTWORTHY DECISION SUPPORT

Decision support systems with integrated measures for bias, accountability, transparency, explainability, and confidential data. Importantly, the ethical, legal, and societal conditions have impact on bias, transparency, explainability, accountability at the same time, while these are often approached separately. This is important for high-risk decision making with high demands on accountability and transparency, such as in governmental processes, banking, insurance, health, law enforcement, defence, sustainable energy systems, and the environment [6].

4.2.4 FEDERATED DECISION SUPPORT

At this level, multiple systems concurrently and continuously learn from their respective users and each other, enabling federated decision making. Key challenge is to improve the federated decision making without being overly dependent on each other. The learned AI model and user data involved must be treated confidentially. Applications range from systems connecting preventive healthcare, curative health care and recovery after treatment, systems enabling banking and insurance companies to collaborate against fraud and illegal transactions to collaborating multimodal energy systems and collaborating robot systems for production and logistics.

4.3 CAPABILITIES

In order to reach the milestone for federated decision making systems, many capabilities need to be developed by TNO and other science and technology parties. Many of these capabilities involve the combination of learning from data with the use of expert domain knowledge.

4.3.1 SCOPING

The potential list of capabilities to develop for the FDM roadmap is endless. We constrain the capability development to the intelligence needed to serve the ethical, legal, and societal demands in high-risk decision making tasks, where collaboration with the user helps in optimizing the information transfer in both ways: making the user better informed about justification of proposed advices and improving knowledge transfer from user into the system about the domain. Certainly, also a wide range of interfacing modalities could be supported, such as facial expressions, gestures or haptic based interfaces. These types of capabilities are considered out of scope.
Further, AI systems can aid in decision tasks of many learning paradigms. Recommending items in webshops, music or movie libraries, aims at suggesting relevant items typically based on similar query items and similar user profiles, while reinforcement learning aims at reward optimization in collaboration or competition. The most common learning paradigm is supervised learning, which is the main target for the FDM roadmap. Supervised learning starts with training data consisting of feature vectors describing the properties of the data subject together with a corresponding target label. The goal is to learn a model that describes the relation between the feature vector and the target labels that can be used as predictor for feature vectors for which the target label is not yet known.
Also, for supervised learning the potential list of capability enabling technologies is large. Different data domains, such as databases, videos, or text corpora, require dedicated tools. The incorporation of domain and expert knowledge, and the amount of learning data and its dimensionality put constraints on prediction model complexity. We will develop the capabilities based on typical use cases and their data and modelling requirements as articulated in TNO units.

Figure 4.2 shows capabilities that contribute to the milestones shown in Figure 4.1. Exact timelines for the development of these capabilities have not been established yet: this will be done in future programs and project plans. Moreover, a capability for a certain milestone generally needs further development beyond that milestone.

The capabilities have been grouped in four capability clusters: fairness, interaction, confidentiality, and adaptivity. These clusters link the evolution along the four dimensions from Section 2.1 to the trustworthiness requirements from the High Level Expert Group guidelines for federated decision making.

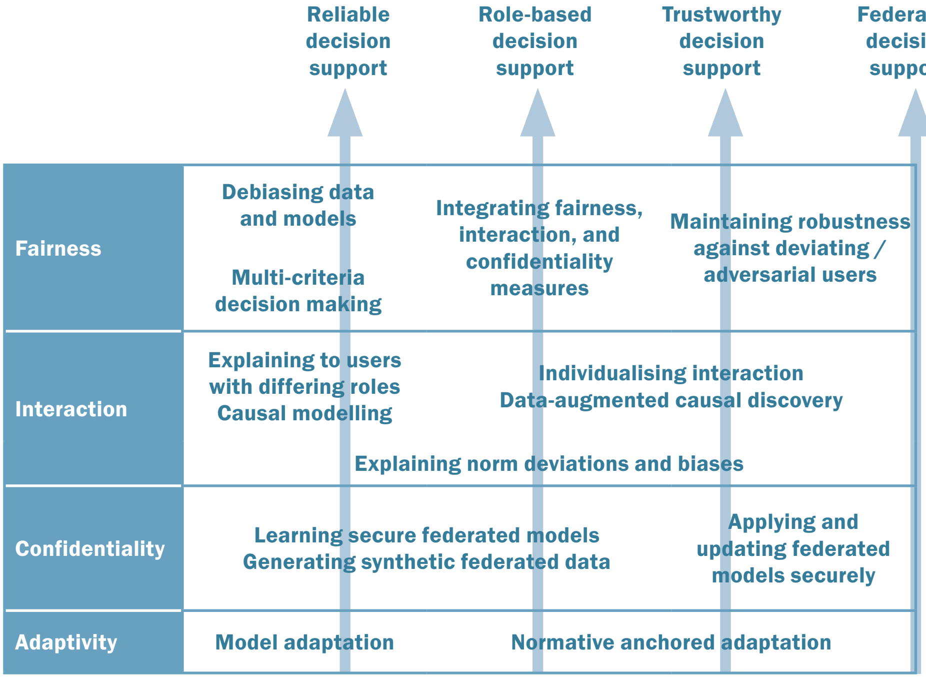

Reliable decision support
Role-based decision support
Trustworthy decision support
Federated decision support
Fairness
Debiasing data and models
Multi-criteria decision making
Integrating fairness, interaction, and confidentiality measures
Maintaining robustness against deviating / adversarial users
Interaction
Explaining to users with differing roles
Causal modelling
Individualising interaction
Data-augmented causal discovery
Explaining norm deviations and biases
Confidentiality
Learning secure federated models
Generating synthetic federated data
Applying and updating federated models securely
Adaptivity
Model adaptation
Normative anchored adaptation

Figure 4.2 *Capabilities for the roadmap Federated Decision Making*

4.3.2 FAIRNESS CAPABILITIES

The fairness capabilities deal with implementing ethical, legal, and societal requirements in AI systems. These capabilities range from dealing with historical data properties such as biases to implementing legal and ethical constraints or balancing the interest of multiple stakeholders.

Debiasing data and models
Data-driven algorithms can discern hidden patterns in data and use these to generate predictions (utility). The exploitation of patterns based on direct or indirect (proxies) discriminative information should be avoided, however. Methods are needed that are able to suppress the use of such discriminative information, and, more general speaking, to balance utility with fairness. Direct discrimination can easily be prevented by leaving out explicit information about gender, age or ethnic background, However, completely avoiding the influence of proxies is generally not possible without rendering the model useless. The capability to balance fairness and utility is needed in all phases of the development process of AI systems [27].

Multi-criteria decision making
The consequence of minding ethical, legal, and societal requirements in AI systems calls for a way to balance multiple requirements and objectives. Relationships can be linear, non-linear, additive, and otherwise dependent. Systems must allow for the requirements to be dealt with in a flexible way such that the interests of multiple stakeholders can be (re)configured [28].

Integrating fairness, interaction, and confidentiality measures.
This is not a fairness capability per se. The different capability clusters must be integrated as required by a given use-case, since capabilities for fairness, interaction and confidentiality are mutually dependent. This makes the integration a capability by itself.

Maintaining robustness against deviating / adversarial users
A learning system must be robust to unintentionally and intentionally deviating inputs such as errors, deviations from norms, or adversarial manipulations. Most demanding are those users, or AI systems acting as surrogate users, who game the system in order to suggest falsified trends and changing conditions. The more transparent the system is, the better it is possible to exploit weaknesses in adaptability. This capability is strongly linked to the adaptivity capabilities.

4.3.3 INTERACTION CAPABILITIES

Interaction capabilities include the capability to provide to the different user roles advanced justifications and explanations of the advices by the system. Special interest is devoted to causality. Making AI systems use causality is a challenging topic with promising results lately.

Explaining to users with differing roles

Transparency in decision making by AI systems is of paramount importance as put forward in practically all documents on regulations for AI systems especially systems taking high risk decisions. The documents hardly elaborate on the meaning. In order to test and evaluate the transparency of existing AI systems as well as to develop methodologies ourselves, we need explicit and objective definitions and measures articulated separately for the three defined anchor user-roles being (domain) researcher, consultant, and subject. The researcher is interested in an overall view of the AI system abilities, and the consultant in the generation of individualised advice for the subjects concerned. The subject typically has a more intimate and continuous relation with the system which requires yet another type of explanation. The current models strongly vary in complexity and a priori transparency of the inference process. In order to gain the trust of the user, she should be offered insight in the relation between her data properties and the derived advice by the system.

Causal modelling

True justifications of AI system advices should be based on causal relations between properties (data) of the subjects and the resulting advices. When causal relations are known a priori, the correspondence between the AI system and this information need to be tested to enable causal explanations. Also, some potential causes expressed as correlations in the AI system can be ruled out with causal reasoning. Specific techniques are needed to integrate causal information into AI systems.

Explaining norm deviations and biases

Before being robust to deviations from the learned model and imposed ethical, legal and societal norms, the system must be able to recognize its limitations and explain these exceptional situations to the user in a way appropriate for that user's role. Also, biases in the data recognized by the AI system need clarification or justification.

Individualising interaction

With the role differentiation as described in Section 4.2.2, the system still has rather rigid user models. Especially the subject role needs to be individualised because divergence in background knowledge and personal conditions is relatively large. Building a personalized model of the user's expectations and abilities is essential for understanding the user and effectively expressing information towards the user. In several application domains of AI such as personalized health and job seeking the user's trust in the system can be increased by using natural language for the information exchange between the user and AI system [29].

Data-augmented causal discovery

In domains with many variables potentially many interactions can drive the prediction models. In such situations the knowledge of causal relationships is lacking. Moreover, in many cases for the researcher one of the main reasons for using the AI system is to learn about the underlying phenomena: for instance, how diabetes develops, or which suspects tend to reoffend. Discovering the causal relationships from available data is an upcoming and promising field [30].

4.3.4 CONFIDENTIALITY CAPABILITIES

We consider two main capabilities related to confidentiality. Especially in collaboration scenarios, where multiple confidential databases are used, secure federated modelling capabilities need to be developed. To enable flexible development of AI systems without the use of the real confidential data, synthetic data generation is a helpful tool.

Learning secure federated models

In high-risk applications, data typically concern personal characteristics and properties to be used by the administrator of these data and only for the purpose for which these were collected. When working in collaboration, other parties typically are not allowed to see and process these personal data. Secure federated learning models allow for learning from each other's data sets, either in peer-to-peer or client-server setting [31]. Confidentiality conditions need then to be defined, implemented, and tested.

With an increasing number of collaborating parties and amount of data per party, the computational and communication demands to the systems involved become higher. This is especially the case if secure encryption layers are implemented to achieve proper levels of security, for instance when multi-party computation (MPC) technologies, such as homomorphic encryption, are deployed. Sophisticated alignment of local computation and MPC technologies is then required to enable appropriate scale of implementation and security.

Generating synthetic federated data

In some cases, the administrator of the data wants to use these data for other than the original purpose and want to share these with other parties to have algorithms developed externally or to even publicize the data. To maintain confidentiality, creating a derived synthetic dataset with the same syntactical and mostly the same statistical properties is then needed [32].

If several databases are to be shared simultaneously, the best solution is clearly not to create independent synthetic datasets, because covarying variables between databases cannot be identified. Creating integrated synthetic datasets solves this problem. It will likely require secure federated learning models (see previous capability) to create such integrated synthetic data.

Applying and updating federated models securely

With multiple federated AI systems that concurrently advise and communicate with their respective users, the challenge remains to keep the overall system state consistent and current, taking confidentiality of learning ánd applying the learned prediction models into account.

4.3.5 ADAPTIVITY CAPABILITIES

Adaptivity aims at adapting to changes in the environment and to trends. It deals with making the system of AI systems consistent using the distributed collected data over time, while imposed norms keep the systems in line, to remain compliant.

Model adaptation

Model adaptation in feedback or online learning scenarios deals with adapting to changing subject properties and contextual conditions, also referred to as concept drift [33]. Such adaptations are necessary to consistently represent the overall state of knowledge, but data errors, outliers, anomalies, and adversarial attacks should be recognized as such and should not lead to adaptations. The system should therefore set limits beyond which it should not adapt to states that deviate too much from the normal patterns [34].

Normative anchored adaptation

Ethical, legal, and societal conditions represent another limitation to AI system adaptation.

Some norms are stricter than others, which means that the 'cost' of violations will differ and will depend on the interplay with other context factors and violations. This capability is related to the multi-criterion decision making capability.

5. ROADMAP AI SYSTEMS ENGINEERING & LIFECYCLE MANAGEMENT

The roadmap AI Systems Engineering & Lifecycle Management develops engineering and management capabilities to build and govern trustworthy AI systems that can adapt to changing operational conditions. In the roadmap description below, we first link the market needs for adaptive and trustworthy AI systems to those for systems engineering and lifecycle management, and then define and elaborate on milestones for the latter. Next, we describe the capabilities that contribute to these milestones.

5.1 MARKET NEEDS

Grand challenges to society such as climate change, an ageing population and shifts in global economic and military power, generate a demand for systems that can adapt quickly to changes in user requirements or tasks, in the operational context, and in the environment. By their adaptability, AI systems potentially meet these demands. Furthermore, we expect AI to increase the effective lifetime and efficiency of systems, advancing sustainability goals and circularity.

Not only will such AI systems in their original ‘engineered’ state differ from their non-AI counterparts, but they will also change within their (prolonged) lifetime to maintain or even increase the fitness and effectiveness for their purpose and to address new operational conditions and risks such as adversarial attacks that were not foreseen in the engineering phase and that need to be managed during their operational lifecycle.

Both the system engineering and the system lifecycle management disciplines are ill-prepared for these challenges. Systems and software engineering methods as well as system governance and lifecycle management principles that have been developed for conventional systems cannot deal with AI systems that contain data-driven or self-learning components during operations [35] [36]. Consequently, there is a need for new principles, frameworks, and methods for AI systems engineering and lifecycle management that ensure that AI systems remain trustworthy during their full lifecycle.

In addition, AI software engineering issues must be addressed that are specific to data-driven AI systems that use machine learning (ML). This so-called-technical debt of ML exists at the system level instead of the software code level. Some examples of these ML issues include the difficulty in specifying the intended system behaviour, unstable data dependencies, and hidden feedback loops [37].

5.2 MILESTONES

Figure 5.1 shows the milestones for the roadmap AI Systems Engineering & Lifecycle Management. These milestones represent progress in the degree of adaptation to changing customer needs in a dynamic environment by introducing data-driven components and on-line learning during the lifecycle. A higher speed of AI system adaptation may lead to risks regarding reliability, human safety and fundamental rights, affecting the user’s trust in the system. The aim of the roadmap is to develop capabilities for building and governing AI systems that can adapt fast to changing operational environments while at the same time maintaining trust and minimizing risks caused by adaptation.

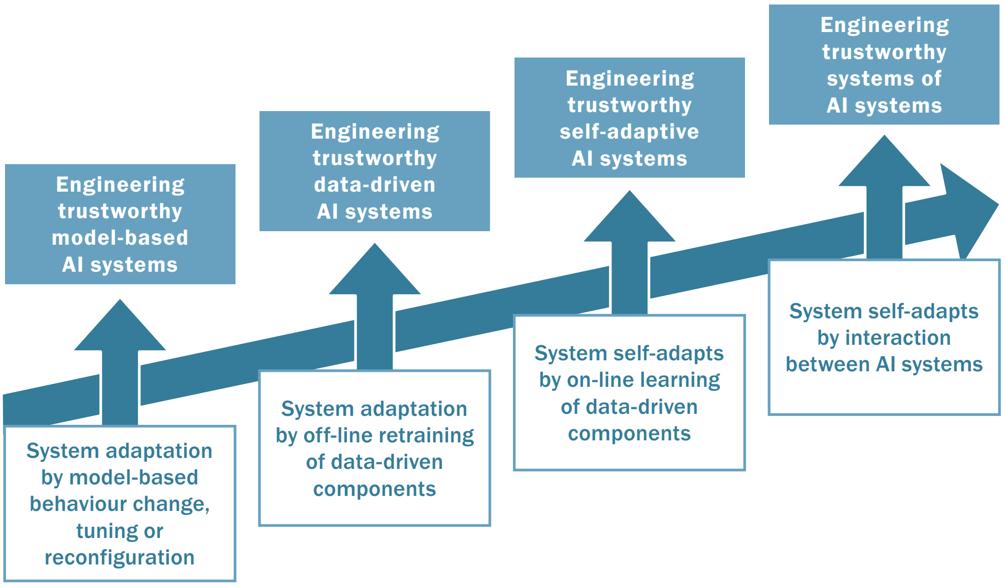

Engineering trustworthy model-based AI systems
Engineering trustworthy data-driven AI systems
Engineering trustworthy self-adaptive AI systems
Engineering trustworthy systems of AI systems
System adaptation by model-based behaviour change, tuning or reconfiguration
System adaptation by off-line retraining of data-driven components
System self-adapts by on-line learning of data-driven components
System self-adapts by interaction between AI systems

Figure 5.1 *Milestones of the roadmap AI systems engineering & lifecycle management.*

5.2.1 ENGINEERING TRUSTWORTHY MODEL-BASED AI SYSTEMS

A model-based AI system is driven by a MAPE-K (Monitoring-Analysis-Planning-Execution / Knowledge) feedback loop that uses operational data for decision making based on knowledge. As Figure 5.2 illustrates, it is designed by human engineers who use typical engineering artifacts, like descriptions of user requirements, but also models encoding the domain knowledge and specifying the system's intended behaviour. These models are transformed at design-time into the AI system's decision making core such as a rule-based decision engine or a model-based probabilistic reasoning engine. Feedback from operational use of the system can be employed by the system engineer to update the model-based AI system.

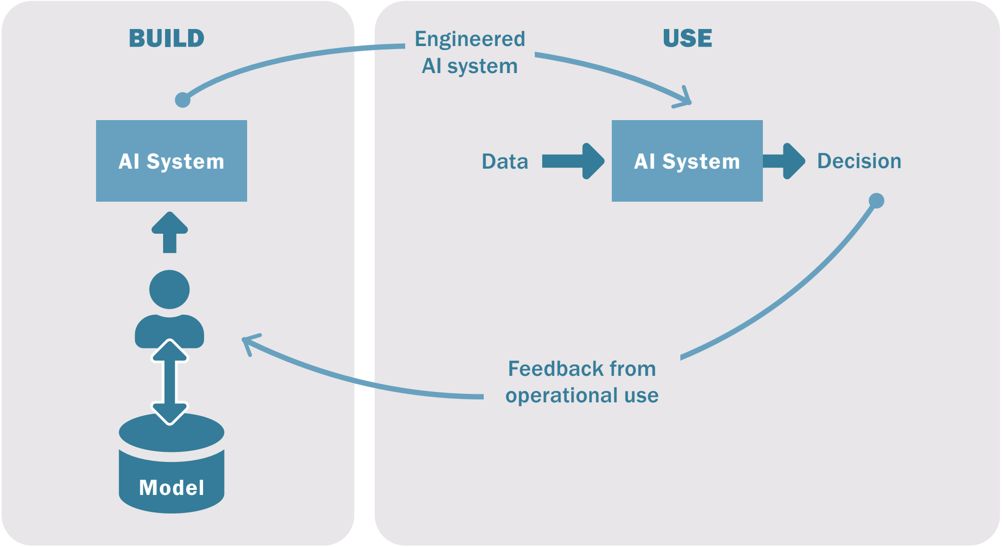


Figure 5.2 *Model-based AI system.*

Benefits of this approach are that the system behaves in a predictable way and that the behaviour can be analysed and explained in terms of the models and domain knowledge. In line with Model-Based Systems Engineering (MBSE), this allows to establish a single-point-of-truth for the system at which it is validated and verified including its future adaptations, providing for traceability and trustworthiness. The approach is, on the other hand, challenged by (i) the scale of the needed effort, and (ii) a lack of understanding on how best to design smart behaviour.

Regarding (i), we see that traditional MBSE as a non-automated activity is not suited to cope with the exploding number of states that may occur in AI-based systems resulting from decision making that takes new incoming data and its history into account. Regarding (ii), we see that smart behaviour emerges via interdependent interactions of control and AI components that differ in their locality,

immediacy, and time-horizons, resulting in a complexity that renders the design of these interactions as a wicked problem.
Both challenges are aggravated if the design, validation, and verification activities need to be repeated many times to realize later adaptations, resulting in considerable efforts and costs.
Consequently, we meet this milestone once we extend MBSE such that it

- allows us to (better) design and analyse intelligent system behaviour as it emerges from the interactions of system building blocks, traditional control, and AI-based decision making.
- allows us to (better) design and analyse model-based AI systems on a scale that fulfils the needs of the industry, authorities and other stakeholders.
- allows us to reduce the efforts in the design, verification, and validation of adaptations to model-based AI systems.

5.2.2 ENGINEERING TRUSTWORTHY DATA-DRIVEN AI SYSTEMS

A data-driven AI system can accelerate the adaptation to feedback from the operational environment by using examples of intended system behaviour to train the system (under supervision of an engineer) to improve an existing function or add new functionality, see Figure 5.3. Newly collected data can be employed to adapt the system quickly to changing environments or new user requirements. For example, a fruit picking robot can be quickly adapted to picking different fruits by training the fruit recognition algorithm with images of those fruits.

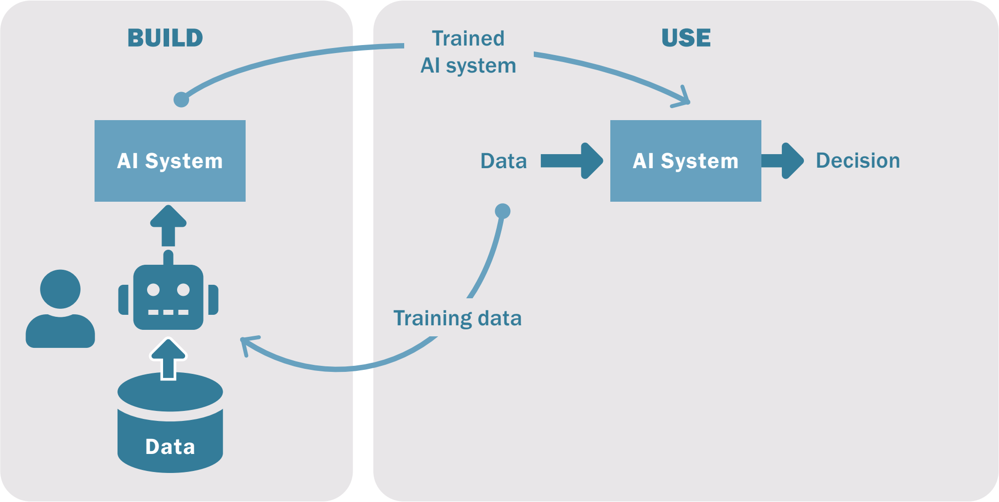


Figure 5.3 *Data-driven AI system.*

This data-driven approach enables applications of AI systems for which it is tedious or too difficult in terms of complexity or efforts to specify the intended system behaviour in advance. Given appropriate data, the use of ML can (i) speed up the development of the AI system by several orders of magnitude in comparison to modelling efforts and (ii) eliminate part of the development efforts as some system components' behaviour is learned from data rather than hand-crafted by an engineer.
A significant drawback of a data-driven approach is that the system behaviour tends to be difficult to interpret and explain due to the black-box nature of many ML algorithms. Here, the advantage of data-driven AI systems of not needing behaviour specifications is at the same time a disadvantage of these systems as this makes it difficult to verify that the system meets the requirements from users and authorities. Moreover, the behaviour of a data-driven AI system can be particularly difficult to predict in the case of unexpected or even malicious input, e.g., in case of adversarial attacks. This is due to a lack of understanding how the performance of ML-based system components changes under such circumstances and how this affects systems in their totality. As this lack of understanding complicates the analysis, verification, and validation of systems, it is yet another reason why it is difficult to realize trustworthy AI systems with this approach.

Finally, change management over the full lifecycle of a data-driven AI system is difficult because data dependencies typically prevent a modular approach. Change, no matter how local it is, will often affect established trust in the system, resulting in a repeat of processes, including learning and validation, to rebuild trust. These repeats can cancel earlier efficiency gains.
Consequently, we meet this milestone once we extend MBSE and ML such that these

- allow us to (better) analyse, validate, and verify intelligent system behaviour.
- allow us to engineer the ML processes that enable trustworthy data-driven AI systems.
- allow us to reduce the efforts in realization, verification, and validation due to system adaptations.

5.2.3 ENGINEERING TRUSTWORTHY SELF-ADAPTIVE AI SYSTEMS

A further acceleration of the adaptation of AI systems to changes in the environment can be achieved by learning from data collected in that environment without going through the system engineering process again, see Figure 5.4. Benefits of this approach for self-adaptive AI systems are that (i) continuous and fast adaptations to changing conditions in operational environments can be achieved, (ii) adaptations are customized for a system and the system's context, and (iii) it becomes easier to keep the system under changing conditions fit for purpose.

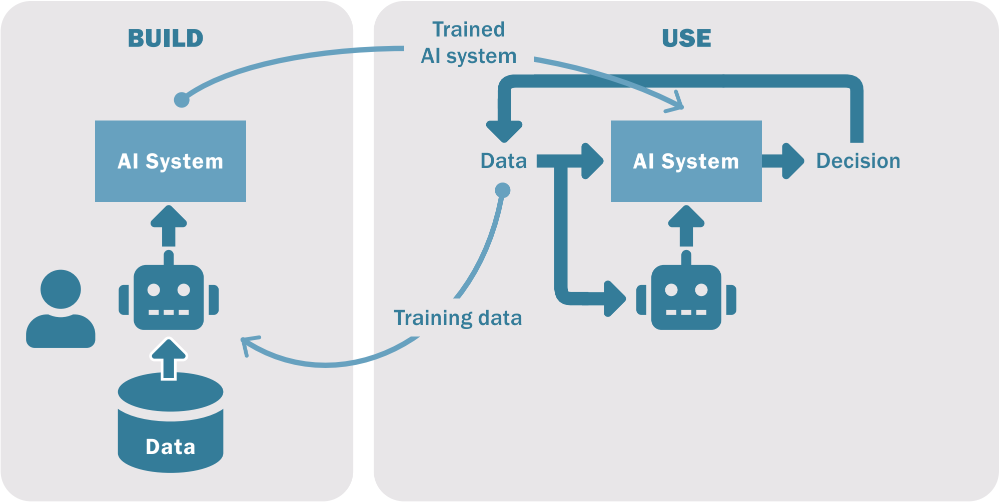


Figure 5.4 *Self-adaptive AI system.*

A drawback of this approach is the difficulty to verify and validate that the system continues to meet its specifications while it is continuously adapting itself. In fact, an AI system that continuously learns and changes after it is brought to market does not comply with the current proposal for AI Regulation that demands that an AI system undergoes a new conformity assessment whenever a change occurs which may affect the compliance of the system with this regulation [6].

Next to that, in self-adaptive AI systems ML faces most, if not all challenges it has in data-driven systems (identified in Section 5.2.2).

Consequently, we reach this milestone once we

- realize self-adaptation and learning processes that are predictable in their impact and can thus be trusted to be beneficial and safe.
- find ways to customize validation and verification processes to specific types of self-adapting systems.

The combination of these two points essentially merges their benefits, but also reflect the needs to combine technological progress made with model-based and data-driven AI approaches.

5.2.4 ENGINEERING TRUSTWORTHY SYSTEMS OF AI SYSTEMS

We expect the next level of AI systems with capabilities beyond the ones described so far to be 'systems of AI systems', for example swarms or fleets of systems, that self-adapt by learning both on an individual level as well as from each other and that jointly learn how to orchestrate their combined efforts for maximal effect.

Potential benefits of this envisioned level are (i) increased efficiency and effectiveness of adaptation, (ii) the increased ability to solve ill-defined adaptation challenges, for instance during exploration of unknown environments; and (iii) increased robustness and resilience of the overall systems of AI systems, even against drastic change or attacks, as the diversity of the existing systems ensures that individual failures are not catastrophic and new challenges can be met from many angles. In other words, 'system of AI systems' benefit from their additional built-in redundancy.

We still know little about design principles for systems of AI systems but clearly any capability to design and to analyse them will bring us closer to this milestone.

5.3 CAPABILITIES

Figure 5.5 shows capabilities that contribute to the milestones shown in Figure 5.1. Capabilities generally do not only contribute to a single milestone but also to subsequent milestones.

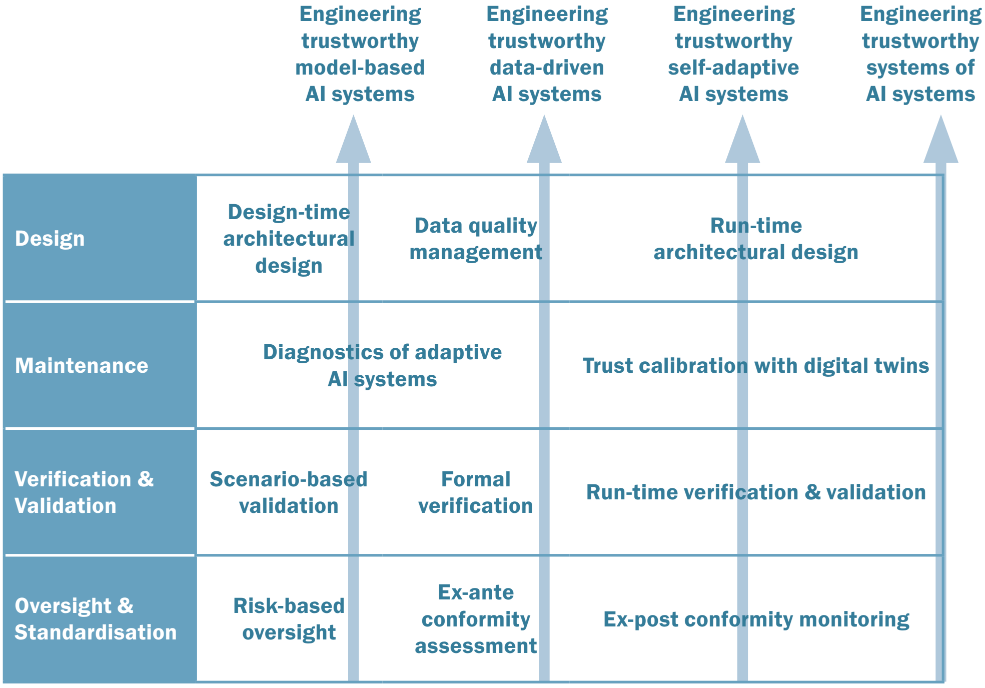

Engineering trustworthy model-based AI systems
Engineering trustworthy data-driven AI systems
Engineering trustworthy self-adaptive AI systems
Engineering trustworthy systems of AI systems
Design
Design-time architectural design
Data quality management
Run-time architectural design
Maintenance
Diagnostics of adaptive AI systems
Trust calibration with digital twins
Verification & Validation
Scenario-based validation
Formal verification
Run-time verification & validation
Oversight & Standardisation
Risk-based oversight
Ex-ante conformity assessment
Ex-post conformity monitoring

Figure 5.5 *Capabilities for the roadmap AI systems engineering & lifecycle management*

5.3.1 DESIGN OF AI SYSTEMS

Design-time architectural design

In the system design phase (design-time) the initial hardware and software architectural design of the AI system is made. Choices are made related to how the system can be made safe and secure by design. In this phase, functional and interface decompositions are important deliverables ensuring that verification of the system can be undertaken easier, and that more functional and non-functional (quality) requirements can be accommodated in the architectural design. If the system is safety-critical, functional redundancy (in hardware and software) and (health) monitoring of the functionality must be emphasized. Several architectural design patterns must be evaluated to make the appropriate design decisions meeting quality requirements. Architectural designs need to be assessed regarding their capability to mitigate faults and their resilience to adversarial attacks. Since AI systems can initiate/introduce different functional failures, mitigation strategies should be designed, integrated, and verified. Methods like Fault Tree Analysis (FTA), Attack Tree Analysis (ATA), Failure Mode, Effect, (and Diagnostic) Analysis (FMEA/FMEDA) need to be focussed specifically on AI and included in the development process.

Data quality management

For data-driven AI, management of the quality of the data used for training and assessment of the algorithms is essential [38]. Ensuring that the data sets are complete to cover system usage in its (potentially new) Operational Design Domain (ODD) is essential. In case the ODD is less known or understood in the system design phase it must be investigated if the system is still able to operate in its ODD as was initially planned. How to derive the ODD boundaries from the data set for training is a challenge. This also holds for the cases where (partially) simulated data sets are used to train the AI algorithms

Run-time architectural design

How to verify the system's real-time self-awareness whether it is still operating within the ODD, nearing its borders, or even trespassing these is a research topic. Especially how to indicate that the system is almost leaving the ODD (since for safety critical applications, the operator or system should then start taking mitigating actions, e.g., transition of control to the operator). When an AI system is learning or collaborating with other (AI) systems, these ODD boundaries might change or even become less well defined to that AI system. AI systems must monitor and control their ODD boundaries to ensure their trustworthiness and explainability.

5.3.2 MAINTENANCE OF AI SYSTEMS

Diagnostics of adaptive AI systems
Diagnosing errors, failures, and insufficient performance of complex systems is difficult due to the many potential root causes. Complex systems are therefore typically designed with an internal health and performance monitoring capability that supports an efficient diagnostics process during operations and maintenance. In addition, maintenance engineers are typically supplied with tools that assist them in quickly analysing root causes of errors in a system and giving them advice concerning the best way to repair the system [39]. The introduction of AI-based diagnostics improves the adaptivity of the system to changing conditions, but it may limit the usefulness of current diagnostics tools and architectures. Therefore, the capability to design architectures and tools for diagnostics of complex adaptive AI systems is needed to ensure the system's performance, safety, and trustworthiness during its lifecycle.

Trust calibration with digital twins
AI systems that operate in an open world must be able to adapt to changing conditions and user requirements. If these systems do so by learning from the data collected during their operations, it is hard for their users to keep track of the system's capabilities and limitations. This can lead to overtrust (the user having higher trust in the AI system than it deserves) or to undertrust (the user not trusting enough the AI system) [40]. A digital twin is a virtual representation of a physical system that enables system lifecycle management tasks such as diagnostics and predictive maintenance [41]. A digital twin of a learning AI system could be used calibrate trust during the system's lifecycle allowing the user to explore the capabilities and limitations in a virtual environment without the risks that are associated with inappropriate use in the real world. The digital twin can also be used to analyse uncertainties in the trained AI system and to investigate if the datasets are suitable to train for new operational conditions (due to extension of the ODD).

5.3.3 VERIFICATION & VALIDATION OF AI SYSTEMS

Scenario-based validation
The validation of a complex AI system is challenging because it is difficult to assess with case-by-case testing if the system meets the user/stakeholder expectations/requirements. One way to overcome this challenge is to test the system (under strict supervision because of the risks involved) during operations in the real world. However, this approach requires many hours of real-world testing to expose the system to rare events that may lead to unacceptable behaviour. On the contrary, a scenario-based validation avoids long and expensive real-world testing by constructing relevant events and situations in Extended Reality (XR) [42]. XR, which covers immersive technologies that can merge physical and virtual worlds, can also be used to formulate and validate ethical goals for AI systems [20]. In many cases a risk-based approach is followed to ensure system reliability/resilience/safety to the appropriate and acceptable level since extensive verification and validation is not always possible (e.g., when an adversarial attack is detected, and system updates are required urgently).

Formal verification
In many low-risk applications such as smart speakers and product recommendation, AI uses ML algorithms such as deep neural networks to automatically learn functions from a set of labelled training data. Increasingly, this data-driven approach is also used for high-risk AI systems such as self-driving vehicles and surveillance systems. The reason for this popularity is from the fact that specifications are not needed: a set of labelled training data determines the system behaviour. However, the absence of specifications makes it difficult to prove that the AI system has been built correctly. Recent publications show that it is feasible to formally verify some classes of (small) neural networks [43]. Extending this capability to more classes and larger neural networks will be very valuable in the development of trustworthy AI systems.

Run-time verification and validation
Formal verification of data-driven AI system components such as neural networks require a significant computational effort. In the initial development, this effort ensures that the software components meet the specifications and behave correctly. In leaning AI systems, however, data-driven components adapt continuously to new conditions and run-time verification is needed to guarantee continued compliance with the specifications [44]. A typical approach in run-time verification is to derive a monitor from the specifications and to use that monitor to check again the behaviour of the components. Run-time checking of the system boundaries and operational condition (e.g., whether the system is still operating in the ODD and not leaving it) is essential to ensure trustworthy behaviour of the system.

5.3.4 OVERSIGHT OF AI SYSTEMS

Risk-based oversight
It is hard to build trust between a human and an AI system without a shared model of the risks of physical or mental harm or infringements of fundamental rights [45]. Even the concept of a human being, which should be central in a trustworthy AI approach, is often absent from the current generation of AI systems. Building and governing trustworthy AI system will enormously benefit from a capability to model these abstract concepts in a way that can be employed by AI systems in their decision making process. It is essential that these models are easily interpretable by and explainable to engineers, users, and authorities.

Ex-ante conformity assessment
The General Data Protection Regulation (GDPR) introduced by the EU in 2018 has had a big impact on the way products and services are developed that deal with personal data. However, many businesses and governments struggle with the interpretation of the GDPR and the consequences for their products and services. As a result, many consultancy agencies now offer their help to assess if data-driven products and services adhere to the GDPR. The AI Regulation proposed by the EU may have an even larger impact than the GDPR. It even includes a requirement for an ex-ante capability to assess the conformity of high-risk AI systems [45]. Note that this regulation might be adapted over time initiating more complicated assessment procedures and additional restrictions during the lifecycle of the system.

Ex-post conformity monitoring
After market introduction of a learning AI system, it will adapt to changing environment and user requirements. For high-risk AI systems this may lead to non-conformity with the AI Regulation and therefore an ex-post monitoring capability is required that can detect potential non-conformities.

6. CONCLUDING REMARKS

In this document we have set out the research directions for the TNO Appl.AI program to create a distinctive technology position in the field of AI at TNO.

The goal is to be able to create AI systems that are increasingly adaptive to the complexity of the real-world. Adaptation is especially necessary to be able to operate in open environments, with multiple purpose tasks, for high-risk applications and in federated collaboration. These systems should be in line with upcoming regulations and ethical guidelines for high-risk AI systems: Trustworthy AI systems.

This Research Strategy addresses the system classes of autonomous systems and federated decision making. These system classes have been explored within TNO for the last two years and cover most high-risk AI applications. For both system classes we have defined a roadmap for the coming 10-15 years. A third roadmap addresses AI systems engineering and life cycle management of these AI system classes.

In each of the three roadmaps, the stepping stones towards the end goal are defined in the form of consecutive milestone systems. Realization of these milestone systems will require the development of clusters of capabilities. The milestone systems will demonstrate the integration of the capabilities at the then current maturity level.

For the roadmaps to become actionable, the next step is to translate these into concrete research programs and projects, assigning research capacity and further refining the technology developments to be made.

The proposed roadmaps are the basis for continuation of the current research program (ending 2022) and for starting up of a follow-up research program.

7. REFERENCES

[1] Zhang D. et al., The AI Index 2021 Annual Report, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.

[2] Hasberg, M.-P., Weima, I. van Lier L., Technology Roadmapping, TNO Report, 2012.

[3] Robotics 2020 Multi-Annual Roadmap for Robotics in Europe, Horizon 2020 Call ICT-2017 (ICT-25, ICT-27 & ICT-28), Release B 02/12/2016

[4] European Commission, White paper on Artificial Intelligence: a European approach to excellence and trust, Brussels, 19.2.2020, COM(2020) 65 final.

[5] Huizing A., Veenman C., Neerincx M., Dijk J. (2021) Hybrid AI: The Way Forward in AI by Developing Four Dimensions. In: Heintz F., Milano M., O'Sullivan B. (eds) T rustworthy AI - Integrating Learning, Optimization and Reasoning. TAILOR 2020. Lecture Notes in Computer Science, vol 12641. Springer, Cham.

[6] Proposal for a Regulation laying down harmonised rules on artificial intelligence, European Commission, Brussels, 21.4.2021 COM (2021) 206 final.

[7] Peeters, M.M.M., van Diggelen, J., van den Bosch, K. et al. Hybrid collective intelligence in a human–AI society. AI & Soc 36, 217–238 (2021)

[8] The High-Level Expert Group on AI of the European Commission. “Ethics Guidelines for Trustworthy AI”. April 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[9] Bradshaw J.M., Hoffman R.R., Woods D.D., Johnson M., The Seven Deadly Myths of "Autonomous Systems, IEEE Intelligent Systems, vol. 28, no. 3, pp. 54-61, May-June 2013.

[10] Marck J. W., Mohamoud A., van der Houwen E., van Heijster R., Indoor radar SLAM A radar application for vision and GPS denied environments, 2013 European Radar Conference, 2013, pp. 471-474.

[11] Crespo, J., Barber, R. & Mozos, O.M. Relational Model for Robotic Semantic Navigation in Indoor Environments. J Intell Robot Syst 86, 617–639 (2017)

[12] Fankhauser P. et al., "Collaborative navigation for flying and walking robots," 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 2859-2866.

[13] Alegre U., Augusto J.C., Clark T., Engineering context-aware systems and applications: A survey, Journal of Systems and Software, Volume 117, 2016, pp. 55-83.

[14] Schwartenbeck P., Passecker J., Hauser T.U., FitzGerald T.H., Kronbichler M., Friston K.J., Computational mechanisms of curiosity and goal-directed exploration, Frank M.J. (Ed.), eLife, 8 (2019), Article e41703

[15] Katz, D., Venkatraman, A., Kazemi, M. et al. Perceiving, learning, and exploiting object affordances for autonomous pile manipulation. Autonomous Robots 37, 369–382 (2014).

[16] Valentini G., Brambilla D., Hamann H., Dorigo M. (2016) Collective Perception of Environmental Features in a Robot Swarm. In: Dorigo M. et al. (eds) Swarm Intelligence. ANTS 2016. Lecture Notes in Computer Science, vol 9882. Springer, Cham.

[17] Awad, E., Dsouza, S., Kim, R. et al. The Moral Machine experiment, Nature 563, 59–64 (2018).

[18] Goodall, N.J. (2016). Away from Trolley Problem and Toward Risk Management. Applied Artificial Intelligence, 30(8), 810-821.

[19] Möller N. (2012) The Concepts of Risk and Safety. In: Roeser S., Hillerbrand R., Sandin P., Peterson M. (eds) Handbook of Risk Theory. Springer, Dordrecht.

[20] Aliman N.-M., Kester L., Werkhoven P., XR for Augmented Utilitarianism, 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), 2019, pp. 283-2832.

[21] Erdem E., Haspalamutgil K., Palaz C., Patoglu C., Uras T., Combining high-level causal reasoning with low-level geometric reasoning and motion planning for robotic manipulation, ICRA, pp. 4575-4581, 2011.

[22] Burghouts G.J., Huizing A., Neerincx M.A., Robotic Self-Assessment of Competence, May 2020, https://arxiv.org/abs/2005.01546v1.

[23] Krupitzer C., Roth F.M., VanSyckel S., Schiele G., Becker C., A survey on engineering approaches for self-adaptive systems, Pervasive and Mobile Computing, Volume 17, Part B, 2015, pp 184-206.

[24] van Diggelen J., Barnhoorn J., Post R., Sijs J., van der Stap N., van der Waa J. (2021) Delegation in Human-Machine Teaming: Progress, Challenges and Prospects. In: Russo D., Ahram T., Karwowski W., Di Bucchianico G., Taiar R. (eds) Intelligent Human Systems Integration 2021. IHSI 2021. Advances in Intelligent Systems and Computing, vol 1322. Springer, Cham.

[25] Giele, T.R.A., Mioch, T., Neerincx, M.A., Meyer, J.J.C., Dynamic Task Allocation for Human-Robot Teams, Proc. 7th International Conference on Agents and Artificial Intelligence, Lisbon, Portugal. January 10-12, ICAART2015, 2015, pp. 117-124.

[26] Neerincx M.A., van der Waa J., Kaptein F., van Diggelen J. (2018) Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance. In: Harris D. (eds) Engineering Psychology and Cognitive Ergonomics. EPCE 2018. Lecture Notes in Computer Science, vol 10906. Springer, Cham.

[27] Caton, S., Haas, C., Fairness in Machine Learning: A Survey. ArXiv, abs/2010.04053 (2020)

[28] Cui Y., Geng Z., Zhu Q., Han Y., Review: Multi-objective optimization methods and application in energy saving, Energy (125) 681-704 (2017).

[29] Laranjo, Liliana & Dunn, Adam & Tong, Huong Ly & Kocaballi, A. Baki & Chen, Jessica & Bashir, Rabia & Surian, Didi & Gallego, Blanca & Magrabi, Farah & Lau, Annie & Coiera, Enrico. Conversational agents in healthcare: A systematic review. Journal of the American Medical Informatics Association (2018).

[30] Shen, X., Ma, S., Vemuri, P. et al. Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology. Sci Rep 10, 2975 (2020).

[31] Lo S.K., Lu Q., Wang C., Paik H.-Y., Zhu L., A Systematic Literature Review on Federated Machine Learning: From a Software Engineering Perspective. ACM Comput. Surv. 54, 5, Article 95 (2021).

[32] Surendra H., Mohan S., A Review of Synthetic Data Generation Methods for Privacy Preserving Data Publishing. International Journal of Scientific & Technology Research (6) 95-101 (2017).

[33] Gama J., Žliobaitė I., Bifet A., Pechenizkiy M., Bouchachia A., A survey on concept drift adaptation. ACM Comput. Surv. 46, 4, Article 44 (April 2014).

[34] Chandola, V.; Banerjee, A.; Kumar, V. "Anomaly detection: A survey". ACM Computing Surveys. 41 (3): 1–58 (2009).

[35] Fischer L., Ehrlinger L., Geist V., Ramler R., Sobiezky F., Zellinger W., Brunner D., Kumar M., Moser B., AI System Engineering—Key Challenges and Lessons Learned. Machine Learning and Knowledge Extraction. 2021; 3(1):56-83.

[36] McDermott T., DeLaurentis D., Beling P., Blackburn M., Bone M., AI4SE and SE4AI: A Research Roadmap, Insight, vol. 23, no. 1, pp.8–14, 2020.

[37] Sculley D., Holt G., Golovin D., Davydov E., et al., Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems, 2015.

[38] Wilkinson M., Dumontier M., Aalbersberg I., et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

[39] Borth M., Barbini L., Probabilistic health and mission readiness assessment at system-level, Proceedings of the Annual Conference of the PHM Society, vol. 11, 2019.

[40] de Visser, E.J., Peeters, M.M.M., Jung, M.F. et al. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. Int J of Soc Robotics 12, 459–478 (2020).

[41] Pileggi P., Lazovik E., Broekhuijsen J., Borth M., Verriet J., Lifecycle Governance for Effective Digital Twins: A Joint Systems Engineering and IT Perspective, 2020 IEEE International Systems Conference (SysCon), 2020, pp. 1-8.

[42] Elrofai H., Paardekooper J.-P., de Gelder E., Kalisvaart S., Op den Camp O., StreetWise – Scenario-based safety validation of connected and automated driving. Helmond, Netherlands: TNO report, 2018.

[43] Sun, X., Khedr, H., Shoukry, Y.: Formal verification of neural network controlled autonomous systems. In: Hybrid Systems: Computation and Control (HSCC) (2019)

[44] Sánchez, C., Schneider, G., Ahrendt, W. et al. A survey of challenges for runtime verification from advanced application domains (beyond software). Form Methods System Design 54, 279–335 (2019)

[45] Greenberg, A.M., Deciding Machines: Moral-Scene Assessment for Intelligent Systems, Human-Machine Shared Contexts, 2020

[46] IEEE Standard Glossary of Software Engineering Terminology, in IEEE Std 610.12-1990, vol., no., pp.1-84, 31 Dec. 1990

[47] Guide to the Systems Engineering Body of Knowledge (SEBoK) version 2.4, https://www.sebokwiki.org/wiki/Guide_to_the_Systems_Engineering_Body_of_Knowledge_(SEBoK)

A GLOSSARY

Adaptability	The ease with which a system or component can be modified for use in applications or environments other than those for which it was specifically designed.
Adaptivity	The ability of a system to adapt to different scenarios, environments, and conditions.
AI system	Software that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with [6].
Autonomous system	A system that can conduct tasks without constant supervision by a human.
Capability	The ability to achieve a desired effect under specified (performance) standards and conditions through combinations of ways and means (activities and resources) to perform a set of activities.
Certification	The process of confirming that a system or component complies with its specified requirements and is acceptable for operational use [46].
Cognition	The ability to interpret the task and environment such that tasks can be effectively/efficiently executed even with environmental or task uncertainty. The ability to distinguish motion caused by perturbations from those by intention.
Interaction	The ability of the system to interact socially, cognitively, and physically with the user of the system, the environment, and the people in the environment.
Mobile system	A vehicle or robot that can propel itself in the real world under control of a remote human operator or on-board AI
Operational design domain	A description of the specific operating domain(s) in which an automated function or system is designed to properly operate.
Perception	The ability of the system to perceive the human intention of the operator, the environment (threats, mass, and size of objects recognition of obstacles, threats) and people in the environment.
Purpose	Intended purpose means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation [6].
Reliability	The ability of a system or component to perform its required functions under stated conditions for a specified period of time [46].
Risk	The statistical expectation value of an unwanted event which may or may not occur, i.e., the product of the probability and the severity of an unwanted event [19].
Robustness	The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions [46].
Safety	The expectation that a system does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered [47].
System provider	A natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge [6].
Technology supplier	A natural or legal person, public authority, agency or other body that develops and supplies AI components which are not a final product or service but part of an AI system.
Telepresence	A set of technologies which allow a person to feel as if they were present, or to give the appearance of being present, at a place other than their true location.
User	Any natural or legal person, public authority, agency, or other body using an AI system under its authority, except where the AI system is used during a personal non-professional activity [6].
Validation	The set of activities ensuring and gaining confidence that a system is able to accomplish its intended use, goals and objectives (i.e., meet stakeholder requirements) in the intended operational environment. The right system was built [47].

Verification — Verification is a set of activities that compares a system or system element against the required characteristics. This includes, but is not limited to, specified requirements, design description and the system itself. The system was built right [47].

B ACRONYMS

AI	Artificial Intelligence
Appl.AI	Applied Artificial Intelligence
ATA	Attack Tree Analysis
CSO	Corporate Science Office
DSS	Decision Support System
DSS	Defence, Safety & Security
ELSA	Ethical Legal Societal Aspects
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
FATE	Fair, Accountable, Transparent and Explainable
FDM	Federated Decision Making
FMDEA	Failure Mode Effect Diagnostic Analysis
FMEA	Failure Mode Effect Analysis
FTA	Fault Tree Analysis
GDPR	General Data Protection Regulation
GPS	Global Positioning System
HLEG	High Level Expert Group
HTSM	High Tech Systems and Materials
ICAI	Innovation Centre for Artificial Intelligence
ICT	Information and Communications Technology
MAPE-K	Monitoring-Analysis-Planning-Execution / Knowledge
MBSE	Model-Based Systems Engineering
ML	Machine Learning
MPC	Multi-Party Computation
NLAIC	Netherlands Artificial Intelligence Coalition
ODD	Operational Design Domain
PMC	Product Market Combination
SLAM	Simultaneous Localization and Mapping
SNOW	Safe autoNomous system in an Open World
TNO	Organisation for Applied Scientific Research
XR	Extended Reality

This paper can be found on the following links:
1) https://www.tno.nl/nl/aandachtsgebieden/artificiele-intelligentie/use-cases/
2) https://www.tno.nl/en/focus-areas/artificial-intelligence/use-cases/

innovation
for life

TNO.NL

21-12340 September 2021