CYBERNETICA

Co-funded by the European Union

ECCC

Risks and controls for artificial intelligence and machine learning systems

Report

Version 1.0

May 27, 2024

D-16-432

PUBLIC

Date	Version	Description
May 27, 2024	1.0	Translation from Estonian to the English language.

Project leads: Liina Kamm (Cybernetica AS)
Hendrik Pillmann (RIA)

Authors: Dan Bogdanov
Paula Etti
Liina Kamm
Andre Ostrak
Tanel Pern
Fedor Stomakhin
Maria Toomsalu
Sandhra-Mirella Valdma
Anto Veldre

Cybernetica AS, Mäealuse 2/1, 12618 Tallinn, Estonia.

E-mail: info@cyber.ee, Website: https://www.cyber.ee, Phone: +372 639 7991.

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Cybersecurity Competence Centre. Neither the European Union nor the European Cybersecurity Competence Centre can be held responsible for them.

Table of Contents

1 Introduction

1.1 Purpose

The Estonian society has adopted digital services for improving work efficiency. Our digital state is renowned for its low administrative overhead. Transactions between state agencies take place over the X-Road data exchange layer. Both the public and the private sector have adopted digital identity solutions. For Estonia, a digital society is an object of constant development.

Rapid advances in computing power have taken the development of artificial intelligence technology to a qualitatively new level. Artificial intelligence systems capable of generating text, images, sounds, music, and video based on a natural language description have made the technology accessible to a wide population, leading to an increasing belief that information technology will enable the development of a new generation of systems capable of performing such tasks better than humans.

Artificial intelligence systems are being developed in Estonia and the rest of the world by both public and private sector institutions. The purpose of this report is to support the implementation of this technology by providing guidance in ensuring cybersecurity, fulfilling of legal requirements, and societal safety.

The report is written for a broad audience. It will be most useful for small and medium-size organisations and private individuals who may not have legal, information security, or artificial intelligence experts on their staff. These users will be able to utilise the quick-reference guide at the end of the report for AI system risk assessment and choice of measures. Our goal is for everyone to use AI lawfully, safely, and without harming the society and environment.

More mature organisations employing quality management systems and more labour-intensive risk management processes will be provided with guidance on the application of artificial intelligence. They will be given recommendations on which standards and reports to follow to ensure an adequate level of maturity.

1.2 Definitions and abbreviations

AGI
: Artificial general intelligence.

AI
: Artificial intelligence.

AI system
: Artificial intelligence system.

AI HLEG
: EU High-Level Expert Group on AI.

API
: Application programming interface.

ASI
: Artificial superintelligence.

BERT
: Bidirectional Encoder Representation from Transformers.

CaaS
Compute as a service.

CNN
Convolutional neural network. A model architecture used in image recognition.

CPU
Central processing unit.

CUDA
Compute Unified Device Architecture, a toolkit developed by the Nvidia Corporation for accelerated general-purpose computing.

DPO
Direct preference optimisation. Fine-tuning method.

FLOP
Floating-point operation. Computational resources required for model training is measured in floating-point operations.

GAN
Generative adversarial network. Model architecture used in image synthesis.

GPT
Generative pretrained transformer. AI model architecture.

GPU
Graphics processing unit.

IaaS
Infrastructure as a service.

AI technology
The study and development of artificial intelligence.

IPO
Identity preference optimisation. Fine-tuning method.

LLM
Large language model. Artificial intelligence model used for natural language processing, distinguished by the large number of parameters involved.

LSTM
Long short-term memory. Model architecture widely used in language models before the adoption of transformers.

ML
Machine learning.

MoE
Mixture of Experts. Model architecture.

NPU
Neural processing unit. Artificial intelligence accelerator mainly used in phones.

OWASP
Open Worldwide Application Security Project. Web community aggregating and producing web application and software security resources.

PaaS
Platform as a service.

RAG
: Retrieval-augmented generation. Method used for the deployment of artificial intelligence applications where the language models inherits additional context from a database or another external source based on a user prompt for improving response quality.

RLHF
: Reinforcement learning with human feedback. Fine-tuning technology utilising reinforcement learning.

RNN
: Recurrent neural network. Model architecture widely used in language models before the adoption of transformers and LSTM.

SaaS
: Software as a service.

SFT
: Supervised fine-tuning. AI model training method that, unlike pre-training, is supervised and is used for the further guidance of the model's work.

TPU
: Tensor processing unit. AI accelerator developed by Google. Corporation

VAE
: Variational autoencoder. Model architecture used in image synthesis.

XAI
: Explainable AI. Collection of methods for the explanation, interpretation, and validation of the work of AI models and the results of this work

1.3 Structure of the report

We begin our report with an overview of the history of AI and main AI technologies (Section 2). We will then move on to their applications, presenting examples of areas of life in which additional value is hoped to be gained from AI. The field itself has been developing rapidly during the writing of this report; hence, we will also include an overview of current trends.

Countries all across the world have begun to legally regulate artificial intelligence. Section 3 provides a review of the current state of this legislation. Section 4 focuses on the architecture of AI systems and presents three general models for the deployment of AI applications. These three deployment models form a good basis for organisations for the application of their risk assessment methodologies .

Alongside legal considerations, applications of AI technology must also take into account cybersecurity and societal safety requirements. Guidelines for relevant risk assessment measures are presented in Section 5. The existence of risks, meanwhile, also necessitates the application of mitigating measures. These are reviewed in Section 6.

Section 7 summarises the recommendations for the promotion of the application of AI systems in Estonia developed in the course of this study.

The last part of the report is the most practical and is mainly targeted at those looking for quick solutions for analysing the risks of AI systems. This part presents specific and easy-to-follow guidance for identifying and dealing with the main risks in the creation or development of an AI system. Relevant guidelines with supporting figures can be found in Section 8.

2 Overview and use cases of AI applications

2.1 History of artificial intelligence technology

Artificial intelligence (AI) is understood herein as any system capable of performing tasks seemingly employing human-level intelligence. Figure 1 presents an overview of important milestones in the history of AI. AI as a field grew out of cybernetics, the goal of which was the study of feedback systems, including biological, technological, and social systems. Although the idea and structure of artificial neurons was already proposed in the 1940s, the history of artificial intelligence is traced to a summer seminar held at Dartmouth in 1956 where the term was first proposed.

The participants of the seminar reached the conclusion that machines can be made to perform all tasks tied to human intelligence. Indeed, they considered computers to be capable of independent learning, language use, and creativity. Even though no big breakthroughs were made during the two-month seminar, over the next 20 years, its participants figured among the main promoters of AI technology. The AI systems developed in this period were capable of solving mathematical problems, playing checkers, and translating texts from one language to another.

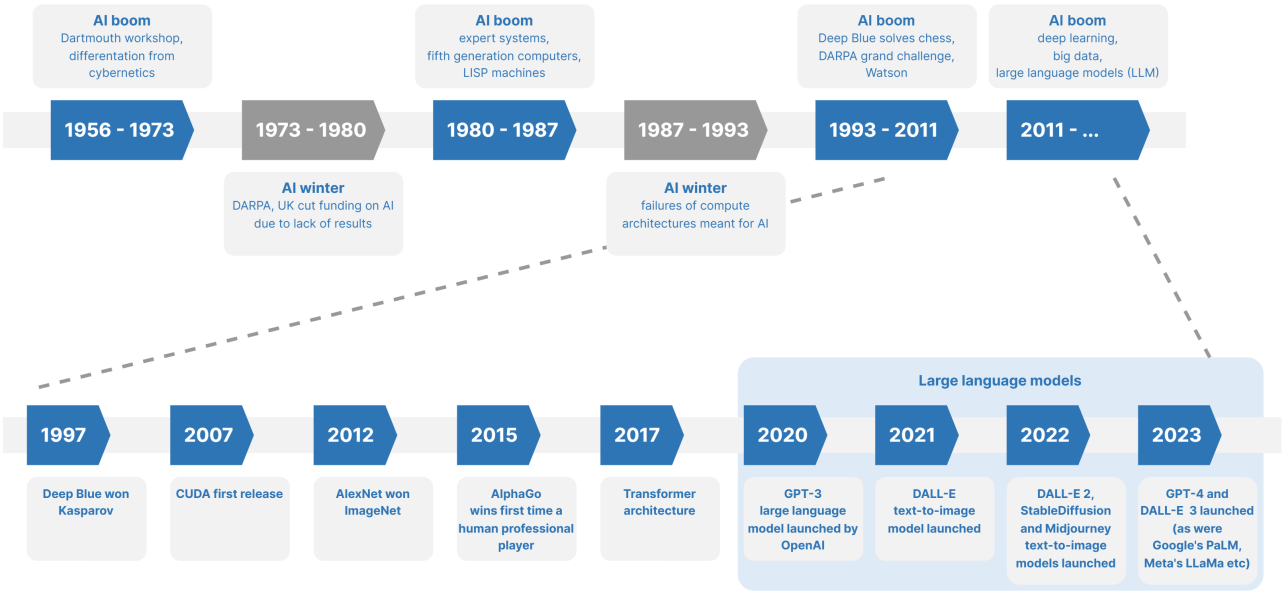


Figure 1. History of AI development

1958 saw the birth of the high-level Lisp programming language that became the main language of AI software for the next three decades. The seemingly major advances and solutions developed in this period fell rather short in reality, though. Translation programs employed literal translation and remained thus unable to relate the meaning of phrases. Programs for proving mathematical theorems or playing checkers were only capable of reviewing a limited number of states and failed to solve more complex problems.

Problem-solving was demonstrated in small play-environments called microworlds. Perhaps the most famous of the microworlds were virtual blocks worlds that the user could manipulate

using English-language commands, e.g. via the SHRDLU language parser. Even though genetic algorithms and the basic principles of artificial neural networks were already proposed in the late 1960s, little progress was made with these algorithms due to their low level of optimisation and insufficient computational power.

The hopes raised by the emergence of the first AI systems led many researchers to make promises that could not be fulfilled. This led to disappointment among the backers of AI research and a decline in AI research and development in the 1970s. Both the UK and the US significantly cut AI funding for universities,and the US Defense Advanced Research Projects Agency (DARPA) stopped funding AI projects altogether. This era from 1974 to 1980 is called the first AI winter.

In spite of the funding cuts, the development of AI still continued, but instead of solving large and complex problems the focus now turned to systems concentrating knowledge provided by experts in different fields and using this for the solution of narrower problems. Such so-called expert systems were used in e.g. medicine and analytical chemistry. Expert systems were also successfully studied by Estonian researchers (including Enn Tõugu and Leo Võhandu).

The success of expert systems led to renewed public interest in artificial intelligence in the early 1980s. One of the first commercial rules-based systems was R1, a system that assisted clients in configuring computers in accordance with their requirements. In 1981, the so-called Fifth Generation Computer Systems project was announced in Japan. The project involved a decade-long plan for the development of intelligent computers. This also created renewed interest in artificial intelligence in the US and the UK.

The new AI boom peaked in the second half of the 1980s. Large American corporations created working groups focusing on AI systems. The focus once again turned to artificial neural networks and their training using back propagation algorithms. Mathematical and statistical optimisation methods, as well as specialised languages and software were increasingly employed for the development of AI algorithms. The best-known AI-specific languages were the all parts of the Lisp family of programming languages. Special computers – Lisp machines – were developed to run programs written in these languages more efficiently.

In spite of the large advances made, 1987 marked the beginning of a second AI winter. The maintenance and updating of specialised artificial intelligences was complicated; they were also unable to independently handle previously unfamiliar inputs, leading to them quickly becoming obsolete. IBM and Apple produced ever higher-performance general-purpose desktop computers. Special-purpose machines (including Lisp machines) lost their usefulness. The fifth generation computer project failed to yield the hoped-for results. Thus, 1991 should have seen the completion of artificial intelligence capable of holding everyday conversations with the user; it would take decades before this goal was finally reached. Disappointed in the limited capabilities of expert systems, DARPA again drastically reduced funding for AI systems research.

Subsequent development of AI technology was increasingly founded on exact mathematical methods developed in the past. The focus once again moved to rigorous logic and solutions were sought from control theory, a subfield of cybernetics. At the same time, researchers also began to utilise probability models and fuzzy logic enabling them to describe relationships and conditional probabilities of features and, unlike pure logic, express lack of knowledge and uncertainty in forecasts.

The 1990s saw the rise of data mining and machine learning algorithms. Systems were no longer described only by programmers and experts: the computers became capable of independent learning through the analysis of large datasets. AI technology and probability methods were tied

together by Bayesian networks allowing the conditional probabilities linking different variables to be expressed in the form of directed graphs. A new paradigm emerged in AI that saw artificial intelligences as agents receiving signals from the environment and attempting to optimise their behaviour for the achievement of certain goals. The greatest achievement of AI technology in the 1990s could be considered to be the victory by the chess-playing system Deep Blue over the reigning chess world champion Garry Kasparov on May 11th, 1997. By this point, AI systems also began to be utilised in everyday services, especially web-based solutions. Natural language processing was thus employed by the Google PageRank search algorithm, also created in 1997. The algorithm ranked the pages displayed after user queries; this is considered one of the critical pieces of functionality setting Google apart from other existing search engines.

Natural language processing was also employed in speech synthesis models, such as DECtalk, used as his speech synthesiser by Stephen Hawking, as well as the slightly more complex Bell Labs TTS (Text-to-Speech system), capable of synthesising speech in several different languages. For nearly 20 years, starting from the early 1990s, machine translation as a field was dominated by statistical models developed at IBM. Meanwhile, hidden Markov models became predominant in speech recognition. The main approach to face recognition in the 1990s consisted in the use of eigenface algorithms employing linear algebraic methods for the analysis of facial features.

In spite of the advances made by artificial intelligence systems, the term AI was still frowned upon at the end of the 1990s. Researchers avoided the term, preferring to speak of statistical methods, machine learning, and control theory instead. The end of the second AI winter is not clearly defined, but it is generally agreed to have ended by 2005 when the Stanford-built self-driving car Stanley covered the 212 km DARPA Grand Challenge trail in the Nevada desert in less than seven hours. This was a major step forward, considering that during the previous year's ten-hour event none of the competing vehicles were able to cover more than 12 km. Two years later, DARPA repeated the competition in a city setting. The winner of this challenge was the Carnegie Mellon University Boss robot which covered 96 km in less than six hours in these conditions.

In 2011, IBM demonstrated their question-answering system Watson on the US TV-show Jeopardy! (Also popular in Estonia under the name Kuldvillak). In two consecutive shows, Watson competed against two human players (one of whom was Ken Jennings, regarded as one of the best Jeopardy! players in history) winning both games by a good margin. Watson's success was founded on ideas derived from a variety of language models and large computing power, enabling the system to be trained on large datasets. Error analysis was continuously carried out throughout the training, and the program was constantly improved. Nevertheless, Watson's performance was not completely flawless. For instance, during the Final Jeopardy! round of the first show, Watson gave the answer 'Toronto' to a question about US cities.

One of the greatest breakthroughs of the artificial intelligence era came in 2012 when the AlexNet convolutional neural network won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a large margin. AlexNet was not the first convolutional neural network; the architecture was first proposed by Yann LeCun back in 1989. The breakthrough was catalyzed by training algorithms optimised for specialised graphics processing units enabling the training of larger and deeper neural networks than ever before. The ImageNet database contained 15 million images from more than 22 000 categories. In the following ImageNet contests, all winning ideas were based on convolutional neural networks and AlexNet's result was improved multiple times. Today, the ImageNet challenge is considered to have been solved.

After the AlexNet breakthrough, neural networks have been subject to active development.

Alongside convolutional neural networks, significant attention was also garnered by large language models, recurrent neural networks, long short-term memory models. This, in turn, led to the rapid development of speech recognition and synthesis and translation models. Artificial intelligence was widely adopted in medicine, industry, and finance. Recurrent networks began to see use in time series analysis, robotics, and games. Notably, the AlphaGo system received great attention after defeating a professional human player at Go in 2015.

As at the time of this report, the main public attention is directed to generative AI models capable of communicating in human language, answering questions, seemingly logical reasoning, generating images and music, and assisting programmers in writing code. While the concept of generative machine learning models is hardly new, the main achievements related to deep generative neural networks date to the previous decade. Generative adversarial models and variational autoencoders were introduced in 2014, both of which are important tools for image synthesis. Generative adversarial models allowed synthesising high-resolution images of human faces for the first time.

In 2015, it was demonstrated that the methods of statistical physics can be used for training generative diffusion models. Perhaps the biggest step forward, however, came in the form of attention mechanism transformers, the basic architecture of which was proposed by Google in 2017. Transformers are at the core of a number of well-known generative language models, such as GPT and BERT, as well as the GitHub Copilot code completion tool.

Transformers enable the construction of parallelisable models with long context windows that can be trained unsupervised on large datasets. Unsupervised models can also be retrained for specific tasks through transfer learning. This is a vital feature, for a time- and resource-consuming universal model only has to be trained once in such case. This model can then later be easily adapted to a specific problem using a much smaller dataset and far fewer resources.

Image synthesis, or more specifically, text-to-image models also use transformers, but their architecture is generally more complex. DALL-E 3 and Stable Diffusion use an autoencoder for encoding images; the encoded data are used for training diffusion models, in turn made up of convolutional neural networks.

2.2 Artificial intelligence algorithms and taxonomies

The term 'artificial intelligence' is very broad and encompasses methods with large differences in complexity, explanatory power and depth, as well as areas of use and training algorithms. On a higher level, artificial intelligence algorithms are divided into rule-based systems, traditional machine learning algorithms, and neural networks.

2.2.1 Rule-based systems

Rule-based systems are the simplest artificial intelligence systems. In general, these systems consist of rules created by human experts that the computer can then follow to solve problems seemingly requiring human intellect. For example, rule-based systems are good at solving certain types of logical thinking exercises and puzzles (e.g. so-called Einstein's puzzles and zebra puzzles).

2.2.2 Machine learning

Machine learning means that the computer learns to solve a task based on existing data (which could include machine-readable representations of sensors, previous events, etc.). Machine learning utilises mathematical optimisation methods which the program uses for finding a maximally accurate solution to the initial problem. This allows the system to solve tasks where the solution algorithm is difficult for a human to describe using precise instructions.

Machine learning methods can be categorised in various ways. For example, from the perspective of applications and training data, machine learning can be divided into supervised and unsupervised machine learning and reinforcement learning.

2.2.2.1 Supervised and unsupervised machine learning, reinforcement learning

In **supervised machine learning**, the goal of the training algorithm is to create a model capable of predicting values or vectors, also known as labels, based on the input received. In unsupervised learning, the model being trained is presented with training data which includes both inputs and the corresponding labels. The model can continuously compare its predictions with correct labels and use the comparison results for improving its prediction capacity. Supervised machine learning is used in almost all fields where machine learning is utilised, such as medical research, image, text, and voice recognition or processing, and the training of search engines and spam filters.

Supervised machine learning tasks are divided into classification and regression tasks. The goal of classification models is to predict which of the two or more classes a given record belongs to. Regression models try to provide a maximally accurate prediction of the numerical value corresponding to the record.

In **unsupervised machine learning**, labels corresponding to the records either do not exist or the model cannot see them. The goal of the algorithm in such cases is to identify relationships or structure within the data without the aid of training labels. Unsupervised algorithms permit the dimensional reduction of the base data (principal component analysis) or grouping of similar records (clustering). Unsupervised machine learning methods are used e.g. in genetics for the identification of sub-populations, as well as for training generative models, such as autoencoders. Unsupervised methods are often also used prior to the employment of supervised machine learning.

Another class of methods alongside supervised and unsupervised machine learning algorithms is reinforcement learning. In the case of reinforcement learning, not every single input will be paired to an output. The algorithm will instead learn to select actions based on the environment so that the reward for these actions is maximised. For example, reinforcement learning can be used for speech processing or teaching the computer to play games. Reinforcement learning was thus used for e.g. training AlphaGo.

Transfer learning is a machine learning technique wherein information acquired for the performance of one task is also used for performing other tasks. For example, trained general-purpose language models can be used for the performance of different linguistic tasks without any additional fine-tuning of the model (see Section 2.2.4.1).

2.2.2.2 Machine learning algorithms

Linear regression (Figure 2) is one of the simplest supervised machine learning models. As a statistical model, it has actually been used for centuries. The model is used for the prediction of a real number output value from input data. As per the name, linear regression is used for modelling a linear relationship between an input and an output. The trained model is thus easily explainable, as it easy to surmise from the model itself how a change in the input value will influence the prediction.

Logistic regression (Figure 2) is very similar in nature to linear regression; in spite of its name, however, it is mainly used for classification analysis. In the case of binary logistic regression, the prediction algorithm first employs a linear function, the output of which can be interpreted as the logarithm of the probability of a label. The output is then passed to a sigmoid function that transforms the output value to a probability in the range $[0, 1]$. Logistic regression can also easily be adapted to situations where there are more than two output classes.

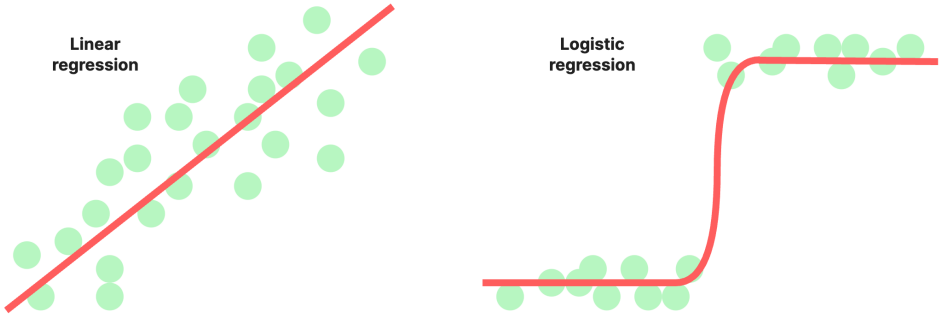


Figure 2. Linear and logistic regression

Support vector machines are supervised machine learning methods initially developed for classification tasks. The simplest support vector machine is a linear classifier tasked with finding hyper-levels demarcating records of different classes. Linear classifiers presume that data classes are linearly separable, which is, however, generally not the case. This has led to the development of a number of adaptations over time which enable support vector machines to be trained for non-linear classification, regression analysis, exception finding, and dimensionality reduction.

Support vector machines are used in image and text classification, but also in e.g. biology. The main weakness of support vector machines is their difficult explainability and higher computational complexity in training.

Decision trees (Figure 3) are supervised hierarchical data structure-based models utilised for regression and classification analysis as a series of recursive decisions. The tree consists of test nodes and end nodes or leaves. In the test nodes, the input is subjected to tests which are used for choosing the next branches. Leaves return the output corresponding to the input based on the tests performed.

Decision-making can be envisioned as a series of yes/no questions were each new question depends on a previous one and the final predicted value depends on each single answer. Decision trees are easily explainable and intuitively understandable models which has made them historically extremely popular.

The naive Bayes method is a classification algorithm utilising the Bayes theorem for the pre-

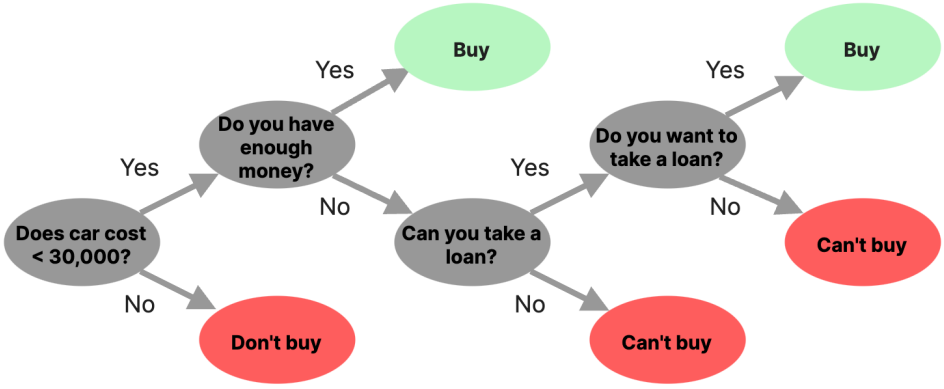


Figure 3. Decision tree for a car purchase

diction of the most probable labels based on an input. This method presumes that the input features used for training the model are independent of each other. Nevertheless, the naive Bayes method has been historically popular due to its sufficient power, as well as simple explainability and trainability. Unlike many other machines learning algorithms, the solution of the naive Bayes method does not have to be found in iterative steps, as the formula for assessing the highest probability can be presented in an explicit form.

The k-nearest neighbour algorithm (Figure 4) is a supervised algorithm that can be used for solving both regression and classification tasks. As per the method's name, predictions are made based on k nearest neighbours where k is a positive integer. In the case of classification tasks, the algorithm determines which class has the highest representation among the k nearest neighbours. In the case of regression, the predicted value is the average of the values of k nearest neighbours. The predictions can be modified by assigning weights to the neighbours based on their distance from the original record. Distances between different points can be measured using different metrics based on the initial problem.

The nearest-neighbour method is popular, as there is no need for pre-training: predictions are made based on the training data. The model is also easily explainable. The main drawback of the model is seen in the fact that the method is a local one, i.e. predictions are based on a few individual records while the rest of the training dataset is ignored.

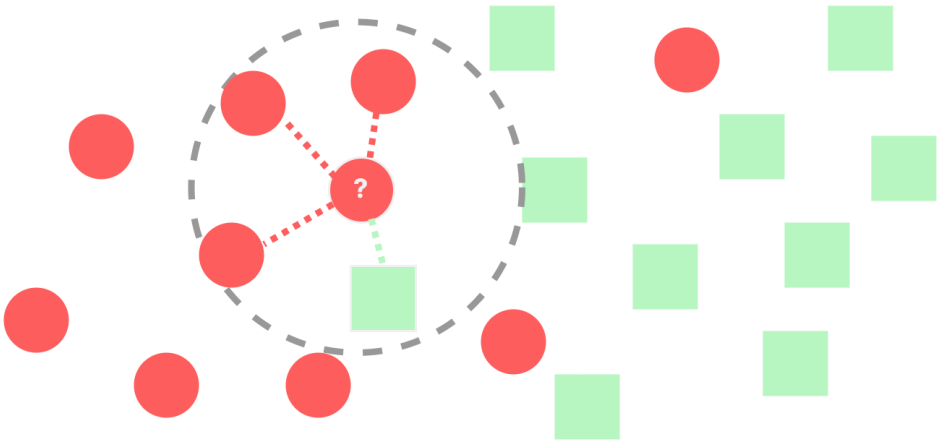


Figure 4. The k-nearest neighbour algorithm analyses the nearest neighbours of the unidentified record

Principal component analysis is an unsupervised algorithm that allows translating data to a more easily explainable coordinate system using linear transformations. Principal component analysis is often utilised the dimensional reduction of the dataset. This is especially useful in situations where many features found in the dataset are strongly correlated to each other. First principal components are vectors that maximally represent the variance of the data upon mapping. Mapping the data onto the first principal components also enables the clustering of the data to be studied visually.

The k-means method or k-means clustering method is an unsupervised machine learning algorithm that divides the data records into k different clusters where k is a positive integer. The k-means method should not be confused for the k-nearest neighbour method which is a supervised methodology. Whereas, in the case of the k-nearest neighbour method, predictions can be made by only looking at the nearest points to the record. The k-means method looks for an optimum clustering for all points which makes training much more difficult and the interpretation of the output requires identifying all the records that were clustered together. Clusters can be used for identifying relationships within the dataset. Clustering yields the centre of each cluster which can be used in e.g., signal processing as a representative cluster point. The method can also be used for automatic feature learning which allows input data to be translated to a form suitable for other machine learning methods.

Hidden Markov models (Figure 5) are statistical algorithms modelling Markov processes, i.e., series of possible events where the probability of each following event only depends on the state of the process after the previous event. Markov process states are not observable in a hidden Markov model. The only things that are observable are the events directly influenced by the hidden states/events. The goal is to use the observable events to study the hidden states and event.

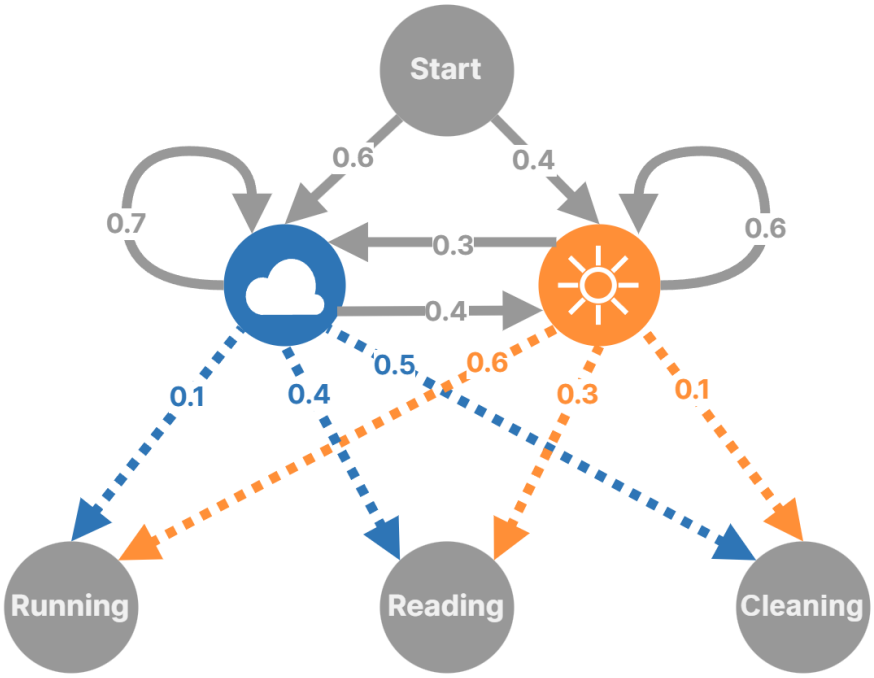


Figure 5. Example of a hidden Markov model of activities for different weather conditions

Ensemble methods (Figure 6) are techniques combining different machines learning models. Combined models are often better and more stable than individual models by themselves. Var-

ious methods exist for the combination of models: bootstrap aggregating or bagging, stacking, boosting. The best-known ensemble methods, such as decision forests and gradient-boosted trees combine different decision trees. Diffusion models have also been used for the generation of neural network parameters [1]. Ensemble learning is also known as meta-learning.

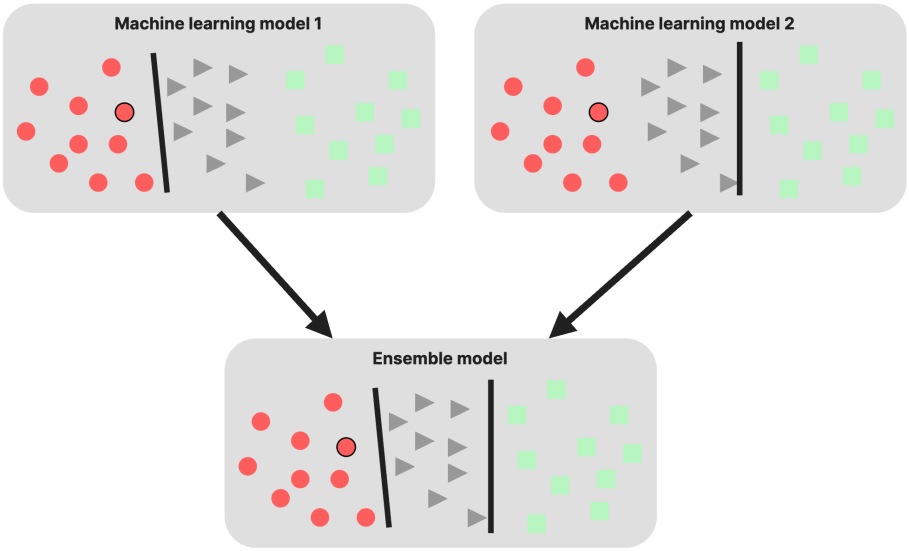


Figure 6. Ensemble methods combine different machine learning models

2.2.3 Artificial neural networks

Artificial neural networks are machine learning models that attempt to imitate the operation of the human brain. Neural networks consist of layers of nodes, the behaviour of which should be similar to the neurons found in the brain. Even though the first neural networks were built as early as in the 1950s, they only saw real success about a decade ago with the creation of the first convolutional neural networks capable of achieving better results in image processing and face recognition than any other existing algorithm.

Increases in computing power and reductions in related costs have created the conditions for training large and complex neural networks which has led to a kind of race in both research and implementation of such systems. Today, models based on neural networks are capable of solving tasks that were considered impossible a mere few years ago. Neural networks are generally difficult to explain and the trained models are seen as black boxes. As a result, more easily explainable machine learning models of similar predictive capacity are often preferred to neural networks. The study of the explainability of neural networks is an active field of research.

2.2.3.1 Neural network architectures

Fully connected neural networks (Figure 7) are one of the first neural network architectures ever developed. A fully connected network is made up of a series of fully connected layers which in turn consist of linear nodes, the outputs of which are subjected to non-linear activation functions.

Convolutional neural networks are neural networks comprising one or several hidden convolutional layers. Whereas fully connected layers comprise linear nodes or weights corresponding to each input value, a convolutional layer is made up of small kernels/filters making the layers

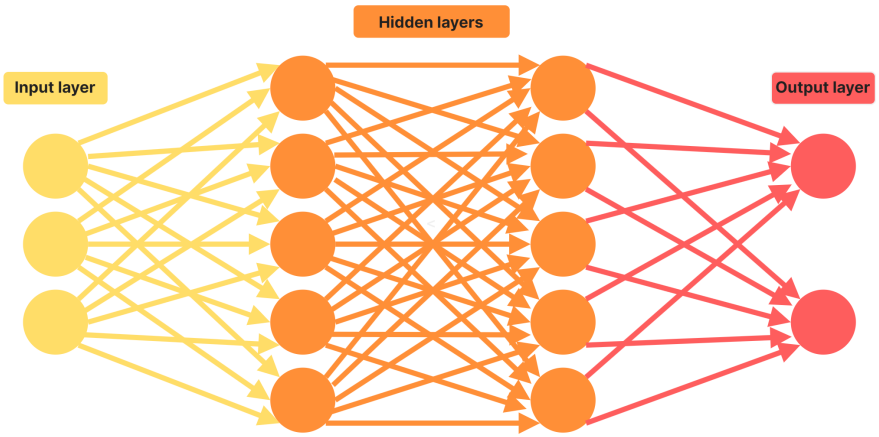


Figure 7. Artificial neural networks consist of different layers and nodes.

smaller and enabling them to be used for the creation of deeper (i.e. involving more layers) neural networks.

The best-known application of convolutional neural networks is artificial vision. In the case of face recognition, a convolutional network can identify different features layer by layer, beginning with lines and angles, followed by eyes and the mouth, and ending with the complete human face. Convolutional networks had a real breakthrough in 2012. with AlexNet which beat other contestants in the ImageNet Large Scale Visual Recognition Challenge by a huge margin. From then on, convolutional networks have been the main tool of artificial vision. Convolutional networks are also successfully used in text processing and, to a lesser extent, other specialty tasks.

Both fully connected and convolutional neural networks are examples of feed-forward networks where the output of a hidden layer is the input for the next layer, i.e. information only flows in a single direction through the network layers. In case information can also flow in a cyclic manner within the neural network, i.e. a layer's output is fed back into the network and can influence later inputs to the same layer, this kind of neural network is called a recurrent neural networks. Recurrent neural networks are mainly used for the analysis of data series, as they can keep track of the preceding inputs within the same series when processing a training input. Recurrent neural networks are widely used in, e.g., language models, text generation, speech recognition, artificial vision, video labelling.

The training of recurrent neural networks can be rendered unstable by the 'explosion' or 'vanishing' of gradients during backpropagation. To mitigate this problem, long short-term memory (LSTM) neural networks have been adopted as a subset of recurrent networks. At the core of LSTM are cells with input, output, and forget gates controlling the flow of information through the cell in order to prevent gradient explosion or vanishing during backpropagation.

Transformers are deep learning models using attention mechanisms for the analysis of sequential data. Transformers came into the limelight in 2017 when it was shown that, when applied to natural language processing, they are capable of identifying the context corresponding to a token based on the preceding sequence without the iterative analysis of this sequence. An input of a certain length is analyzed as a whole and an attention mechanism is used to identify the signals most relevant to each token in the preceding sequence of tokens. This enables the models to be trained in parallel, thus reducing computing costs compared to e.g. LSTMs.

Unlike LSTMs, in the case of long inputs, transformers lack the capacity to keep track of the

entire preceding series and can only track a certain segment of the series which can prove problematic in analyzing long texts. Transformers generally consists of an encoder and a decoder, the first of which analyzes the input ant the second generates the output step-by-step. The coder and decoder can be used both simultaneously and separately. For instance, GPT is a purely decoder-based and BERT a purely encoder-based models; there are, however, models such as T5 that employ both an encoder and a decoder.

Transformers are used in the training of both supervised and unsupervised, as well as hybrid models. Large language models, such as BERT and GPT, are first trained unsupervised on a large set of texts. The model will then be trained on a smaller, labelled dataset for a specific task. Transformer-based models have achieved almost complete dominance among language model in the recent years. In other fields, however, no similar success has been observed. For instance, in artificial vision, convolutional neural networks are still preferred to transformers, even if attention mechanisms are already employed in these.

An **autoencoder** is an unsupervised neural network comprising both an encoder and a decoder. Input is received by the encoder and transformed to another form, while the decoder attempts to reconstruct the original input from the transformed input. The trained encoder can then be used for dimensional reduction of the input data and the decoder, for the generation of new data. In most cases, the generative capacity of an autoencoder is limited, as the proximity of the decoder's inputs does not guarantee the similarity of outputs.

Variational autoencoders (VAE) have been proposed for use in data generation, e.g. image synthesis. VAEs differ from ordinary autoencoders in that the encoder maps an input to a distribution, rather than a single point, e.g., by outputting a normal distribution mean value and covariance matrix, whereas the decoder will be given a random vector from this distribution as an input which it will then try and use for reconstructing the encoder's original input. Unlike autoencoders, trained VAE decoders will usually generate similar outputs for proximate inputs.

A **generative adversarial network** (GAN) is a generative model where two neural networks – a generative and a discriminative one – contest with each other for training the model. Both neural networks are trained simultaneously. The generative model receives an input from a simple distribution and attempts to use this to generate an output from a complex descriptive distribution, whereas the discriminative model attempts to distinguish the outputs of the generative model from real data, the distribution of which the generative model is attempting to imitate. GANs can be used in, e.g., image synthesis where the generative model is generating images of humans while the discriminative model is attempting to distinguish the real images from the generated ones. Generative adversarial networks are also used in speech and text synthesis.

Diffusion models are generative models based on Markov processes. Diffusion models are somewhat similar to autoencoders, in that they comprise a forward process where noise is added to real data step-by-step, and a reverse process attempting to recreate the original input through the gradual removal of the noise. In general, noise used for training diffusion models is generated using a normal distribution; after the addition of a sufficient amount of noise the original input will disappear completely and the output will only consist of random noise.

If a lot of noise is added to the input at once it will be extremely difficult to predict the original input, but it turns out that when noise is added in sufficiently small increments, the most recent addition of noise can be predicted and removed using, e.g., a neural network for the prediction. The trained model can be sequentially applied to a completely random input and used to generate an output similar to real data.

The efficiency of such training stems from the knowledge that if noise is generated from a normal

distribution and added sequentially, then all of the added noise also originates from a normal distribution. The sum of noise from several increments can thus be simultaneously added to the original input during training, and the neural network can be asked to predict only the small amount of noise added in the latest step.

Diffusion models originate from statistical physics. In 2015, it was demonstrated that they can also be used for image synthesis. Subsequent study of these models has given rise to the realisation that diffusion models are more powerful and stable yet less resource-intensive than, for example, generative adversarial networks that were previously the best image-generating models. Today, diffusion models and transformers are the main components of text-to-image models, such as DALL-E 3 and Stable Diffusion.

2.2.4 Large language models

Large language models (LLMs) are generally transformer-based text synthesis models, distinguished by the large number of parameters and amount of training data used. Non-transformer based language models also exist. Various architectures, such as RetNet [2], RWKV [3], and Mamba [4] have been developed that can also be used for the creation of language models, offering solutions for the weak sides of transformer architecture. A large part of recent innovation in machine learning and artificial intelligence has been related to the development of LLMs and the adoption of products (such as ChatGPT) built on LLMs.

According to one hypothesis, should artificial general intelligence (AGI) prove possible at all, it can only be developed on the basis of multimodal large language models [5]. Demis Hassabis from the AI developer DeepMind has opined that 'multimodal foundational models are going to be key component of AGI'[1]. Conceptions and definitions of AGI vary, however, and some claim the necessary level of technology has already been reached [6].

2.2.4.1 Training

As with all other artificial intelligence models, the model architecture needs to be trained after being established. The training of LLMs usually involves several steps, none of which are, however, strictly required. The training process of LLMs and the choices made in the process are closely tied to the deployment models of the AI applications founded upon the LLM.

Pre-training is the first, unsupervised stage of training where the model is fed text sequences containing masked elements and is instructed to predict these elements. The selection of the masked elements is automatic. Pre-training is the most compute-intensive process involving huge amounts (∼trillion tokens) of unlabelled, low-quality data, usually acquired through web crawling. Pre-training yields a pre-trained model that can generate a continuation to an input based on what it has learned from the training data. This continuation may not necessarily be useful: when the pre-trained model is asked a question it can generate an answer to the question, or it may generate a continuation or follow-up questions.

Supervised fine-tuning (SFT) is the second stage of training meant to tune the model for a specific purpose. For instance, in the case of chatbots, it is specifically preferred that the system generate answers, not other kinds of outputs. Training data used for fine-tuning are often, although not always, assembled and labelled by humans. Their quality is higher and quantity

[1] The Guardian: 'Google says new AI model Gemini outperforms ChatGPT in most tests'. https://www.theguardian.com/technology/2023/dec/06/google-new-ai-model-gemini-bard-upgrade Visited December 11th, 2023

much lower (∼tens of thousands sample pairs) compared to pre-training data.

Reinforcement learning with human feedback (RLHF) is the third, reinforcement-based phase of training where the model is tuned to human preferences. A reward model is created for this purpose, which is then applied to the fine-tuned model for the evaluation of its outputs. The reward model is trained using a dataset created with human assistance where each query is mapped to 1 or more $(good\_answer, bad\_answer)$ pairs where the goal is to maximise for each pair the difference between the reward model's evaluation of the good and bad answers. After the reward model has learned to distinguish the desirable answers from the undesirable ones, it will be employed to additionally fine-tune the model that has already undergone SFT during reinforcement learning.

Direct preference optimisation and **identity preference optimisation** (DPO, IPO) are alternative approaches to fine-tuning where, similarly to RLHF, a dataset of human preferences is used for preference learning. The two approaches are distinguished by the fact that unlike RLHF, DPO and IPO do not require the employment of a reward model because the LLM itself can fulfil the role of the reward model [7, 8], using the difference between the evaluations of good and bad answers as the loss function. Whereas a model that has only been pre-trained can give irrelevant or dangerous answers, SFT and RLHF/DPO/IPO as parts of the training process enable using human supervision to train the model to make it more secure and more compatible with user and business requirements.

2.2.4.2 Inference and context learning

A prompt is a user input token used by a generative image or language model for the generation of an output. This process is called **inference**. A prompt is usually made up of a natural-language text. The prompts used by LLM-based chatbots are combined with a pre-prompt containing additional information on the context of the conversation, the user, and the language model. Among other things, this is crucial for ensuring that the chatbot's output is based on its role as a chatbot responding to questions, rather than generating a continuation to the user input. A pre-prompt can also be used for providing information about the outside world, such as the date, time, user name, contents of a document or text file, and other features of the user or the environment.

Models are unable to distinguish a prompt from a pre-prompt, a fact exploited by numerous prompt injection techniques. As the pre-prompt is easy for the user to acquire through a well-crafted prompt, it should not contain information that the user should not have access to. In the case of transformer architecture, the prompt along with the pre-prompt must fit into the model's **context window** which is measured in tokens and contains the (pre-)information necessary for generating an output. Another, more complex form of this approach is retrieval-augmented generation (RAG) wherein the language models creates a database query based on the user prompt and API information found in the pre-prompt, and uses the results of this query for generating a response. This also solves of the problem of the user-provided data being too large to insert into the context window using a prompt. Model architectures with unlimited prompt length also exist, e.g. Mamba [4] and RWKV [3].

Whereas simpler language models require retraining or fine-tuning for each new task, the language knowledge and generalisation capacity of LLMs mean that, in many cases, articulating the task and adding a few examples to the prompt is all it takes [9]. Given that information related to the task is fed into the model's context window, this approach is called **in-context learning**.

In-context learning is divided into numerous sub-methods: few-shot learning where the prompt is supplemented several examples alongside the instructions, one-shot learning where a single example is provided, and zero-shot learning where the query is made without providing any examples. The more parameters the language model contains, the fewer examples have to be normally added to the prompt for the successful completion of the task.

2.3 Applications of artificial intelligence

Image synthesis means the automatic generation of an image with predetermined features, e.g. based on a verbal description (or another image and a verbal description). Image synthesis subfields include, in an order of increasing granularity, inpainting, outpainting, style transfer, deep learning-based noise removal, video synthesis, and refinement. These days, image synthesis generally employs generative adversarial networks [10] and, increasingly, diffusion models [11, 12].

The purpose of **artificial vision** is the artificial extraction of information from images. This comprises class segmentation and instance segmentation, labelling, and object recognition. Artificial vision generally utilises convolutional neural network (CNN) and transformer-based deep learning models [13, 14]. Common use cases include the monitoring of cattle and agricultural equipment, monitoring of road conditions and the surroundings by self-driving cars or delivery bots, face recognition, and augmented reality.

The purpose of **speech synthesis** is to generate human-understandable speech from a given text. Primitive speech synthesis models operated by sequentially linking pre-recorded phonemes or words, but today, transformer-based neural networks are generally used for this task [15, 16]. Speech synthesis is used in chatbots, automated message delivery, screen readers, computer game localisation, and dubbing. Subfields of speech synthesis include speech style transfer, i.e., imitation of the tone and patterns of sample speech.

In contrast to speech synthesis, the purpose of **speech recognition** is the extraction of information from human speech. Speech recognition includes speech transcription, in the case off which textual information is extracted from the speech. Whereas past speech recognition models employed statistical methods, today's systems are mainly built upon neural networks based on CNNs and transformers [17]. Speech recognition is used in smart homes and hands-free devices for voice instructions and dictation.

Natural language processing is a broad field comprising the generation and classification, as well as the interpretation of texts. Text generation generally means predicting the next token, with previous tokens providing the context for the prediction. Text classification and interpretation are used in semantic search where candidate phrases found in a document or text excerpt are compared not based on keyword matching but semantic proximity. Deep learning networks comprising recurrent neural networks (RNN) and long short-term memory (LSTM) were previously used in speech synthesis. A major breakthrough in the field came with the emergence of large language models (LLMs), the architecture of which is generally transformer-based [18, 19, 20]. LLMs are used in, e.g., copywriting, chatbots, neural machine translation, emotion analysis, and code generation.

General data processing and analysis. Machine learning methods are also used in data analysis in other applications. These include various classification, cluster analysis, and discrete or continuous feature prediction tasks, such as predicting stock price movements, processing of brain signals collected by a brain-computer interface, or cluster analysis based on clients' consumption habits. Depending on the nature of the task, both deep learning neural networks and

statistical machine learning methods can be used.

2.4 Areas of use of artificial intelligence

The technologies discussed above have found use in many walks of life: e-governance, the private sector, education and research, healthcare, and unspecified personal uses. We will next take a look at some of these fields and applications.

E-state and e-governance. The AI strategies published by the Estonian Ministry of Economic Affairs and Communications foresee widespread adoption of AI in the public sector. The natural language processing-based virtual assistant Bürokratt enables communication with public sector services via a chat window. The Estonian Parliament's digital stenographer Hans uses speech recognition to transcribe speeches made in the plenary hall. Several ministries have used the Texta text analysis toolkit for auditing their documentation. Ilme, a service provided the National Archives of Estonia, allows using artificial vision to find people similar to user-uploaded images in historical photos.

Education. Artificial intelligence has numerous uses in education, e.g. consider the Education technology compass published by the Estonian Education and Youth Board (HARNO)[2]. The educational non-profit Khan Academy uses a chatbot based on GPT-4 for the personalisation of studies. The Duolingo language learning application contains a similar GPT-4-based interactive chatbot solution; natural language processing methods are also employed by the Lingvist language learning application.

Research. Artificial intelligence and machine learning have been used by researchers for both discovering new knowledge and finding and systematising existing information [3]. The SemanticSearch search portal uses natural language processing and artificial vision for summarising, indexing, and searching scientific publications, whereas the AlphaFold AI system developed by Alphabet has made it possible to predict the shapes of proteins with previously unknown structures. Models based on machine learning and artificial intelligence have been adopted in particle physics for data analysis and simulation design, and in biomedicine, for the development of new pharmaceuticals.

Healthcare. Artificial intelligence has been successfully employed in personal medicine, clinical research, as well as drug development[4]. Machine learning-based big data analysis methods allow using the patient's gene data for providing better treatment. Artificial vision is helpful in the interpretation of medical images and diagnosing the patient. Natural language processing and text analysis methods enable finding and organising patient data. Machine learning methods are used in drug development, e.g. in molecular simulations, prediction of therapeutic properties, as well as the generation of molecular structures and synthesis paths.

Private sector. Machine learning-based audio processing, noise removal, and audio and video stream packing techniques (Skype) are utilised in telecommunications. Artificial vision is used in e.g. robotics (Milrem, Cleveron), agriculture, identity verification (Veriff). Chatbots based on natural language processing are increasingly common in customer support.

[2]Education and Youth Board. Education technology compass. https://kompass.harno.ee/tehisintellekt Visited August 10th, 2023

[3]OECD, Artificial Intelligence in Science. https://www.oecd.org/publications/artificial-intelligence-in-science-a8d820bd-en.htm Visited August 10th, 2023

[4]National Institute for Health Development. Artificial intelligence as the foundation for personal medicine in oncology. https://www.tai.ee/et/personaalmeditsiini-uudiskirjad/tehisintellekt-kui-personaalmeditsiini-alus-onkoloogias Visited August 11th, 2023

Personal use. AI-based personal assistants, such as Google Assistant, Amazon Alexa, and Siri were common even before the emergence of LLMs and diffusion-based image synthesis models. The proliferation and increased accessibility off LLMs and diffusion-based image synthesis models has led to an evolutionary leap in this area, including the widespread adoption of the AIaaS (artificial intelligence as a service) business model. Models developed for presonal use and plugins and applications built upon these can analyze code (GitHub Copilot), read documents or web pages and extract necessary information (Bing Chat), generate texts from birthday invitations to marketing materials (ChatGPT).

Image synthesis models can be used by individuals for creating illustrations in the desired style, generate interior design ideas, increase the resolution of images or photos (StableDiffusion, Midjourney), and even identify certain species of mushrooms in the woods.

2.5 Explainability in machine learning

The emergence of deep learning methods and increases in the complexity of machine learning models have given rise to questions regarding the explainability of the models. Explainability of a model means the ability to provide a human-understandable explanation of the relationship between the model's output and input. EU data protection regulations consider the transparency of the used artificial intelligence technology vital for situations where automated decisions are made using machine learning models [21]. This can be achieved through the explainability of the model.

Explainable AI (XAI) has been proposed as a solution facilitating movement towards more transparent artificial intelligence and thus avoiding limitations on the adoption of AI in critical areas [22]. As at the time of this report, there is as of yet no global consensus regarding the desirable threshold of algorithmic explainability [23].

Explainability is closely tied to the issues of transparency and trustworthiness of AI systems. The systematic definition of explainability requirements is thus a vital step in the development of transparent and trustworthy artificial intelligence systems [24]. The OECD has found [25] that, in order to ensure transparency and explainability, AI actors should provide meaningful information, appropriate to the context, and consistent with the state of art:

- to foster a general understanding of AI systems;
- to make stakeholders aware of their interactions with AI systems;
- to enable those affected by an AI system to understand the outcome and
- to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the result.

Real-world interpretations of the explainability requirements have also been studied [24]. These studies have led to the finding that the explainability of AI is, *inter alia*, fostered by the establishment of systematic definitions and the formalisation and quantification of explanations and performance indicators [22]. Four components of explainability have been proposed [24]:

- addressees – to whom to explain?
- aspects – what to explain?
- context – in what kind of situation to explain?
- explainers – who explains?

An explainable model is more trustworthy, it is easier to develop, test, and audit; it is also easier to detect any biases and explain abnormal behaviour. Explainability is vital in medicine where, e.g., an image model designed to detect tumours is more trustworthy, if the prediction is accompanied by an explanation of which features of the image (contrast, shape) led to the detection of the presence (or absence) of a tumour. Likewise, someone receiving a negative response to their loan application from a bank might be interested to learn what they need to do for the bank to give them a loan (a so-called counterfactual explanation). A profanity filter highlighting the input words contributing the most to the classification of a message as obscene will be easier to develop and test than one that does not.

Explainability is not always required. Where the risks are low and the problem itself has already been studied in depth, it can prove superfluous. As a rule, there is also a trade-off between the capabilities and explainability of the model [26]. Whereas, in the case of a linear regression model, the relationship between the output and the input can be gleamed from merely looking at the regression coefficients, more complex and more powerful models, such as deep neural networks, are a kind of a 'black box' [27] for humans where the model's prediction or decision principles are no longer identifiable on the basis of the model's structure and parameters.

Explainability can be divided into intrinsic and post-hoc explainability. In the case of intrinsic explainability (also known as transparency), the model's complexity is limited in order to prevent it from becoming a black box and to maintain the explainability of its parameters over the entire model from the start. Models with a simple structure, such as decision trees and simple regression models, are considered self-explainable. Where the task at hand calls for the employment of a more complex model, post-hoc methods are used for increasing its transparency.

Post-hoc methods are generally model-agnostic – they do not depend on the architecture of the model, nor do they presume the possession of an overview of its internal components. Post-hoc explanations treat all models, including those that are self-explainable due to their simplicity, as black boxes. So-called local post-hoc explainability methods demonstrate how much and in which direction small individual changes in input features will shift the model's output, or what are the smallest necessary changes in input features required for the model to predict another class. Global post-hoc explainability methods allow understanding the intermediary layers of an already-trained model: thus, OpenAI has created Microscope[5], a collection of visualisations, than can be used to acquire an overview of the intermediary layers of different image models, the neurons contained therein, and their properties. It also allows studying which pictures within the input dataset activate the neuron in question the most.

2.6 Global trends

2.6.1 Faster and larger

Increasing model sizes. Just as computing power, the size of neural networks has also undergone an exponential growth. In 1989, Yann LeCun's team used a convolutional neural network to identify numbers in images. The network consisted of two convolutional and one fully connected layer, for a total of fewer than ten thousand trainable parameters. The AlexNet model introduced in 2012 comprised of five convolutional and three fully connected layer, with as many as 61 million parameters.

With the spread of transformer architecture, the number of trainable parameters kept increasing (Figure 8): The BERT-base and GPT-1 language models (2018) already contained ∼110 million,

[5] OpenAI Microscope https://microscope.openai.com Visited December 10th, 2023

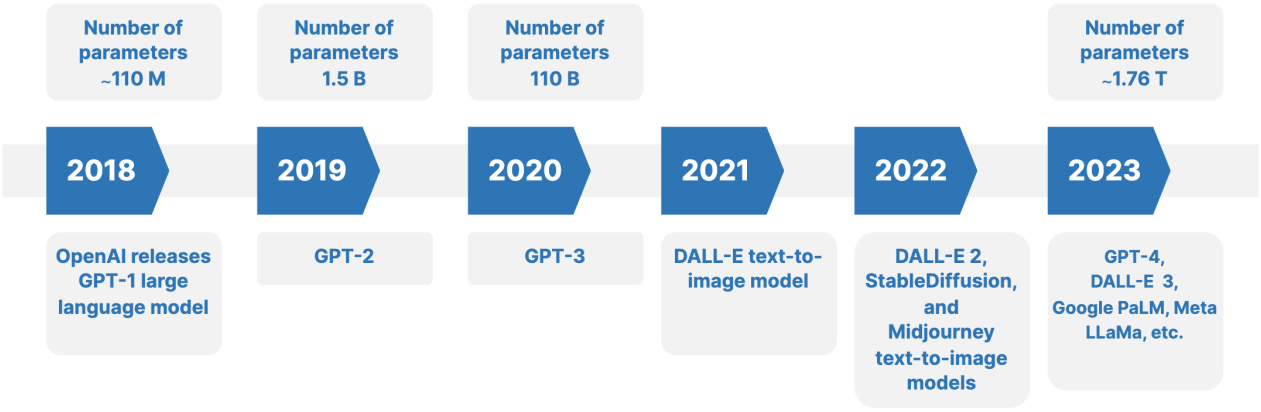


Figure 8. Growth in the number of model parameters has been exponential.

GPT-2 (2019) – 1,5 billion, and GPT-3 (2020) – 175 billion trainable parameters. The number of parameters used in GPT-4 has not been publicised but it has been speculated that it is a so-called 'mixture of experts' (MoE) model with ∼1.76 trillion parameters. The increase in the number of parameters also means increased demands for computing power and memory required for both training a model and the application of a trained model (inference). Efficient training of a model also requires ever-larger amounts of training data.

With the increase in the number of parameters, language models have begun to exhibit emergent abilities, generally understood as capabilities that exist in models with larger numbers of parameters but lacking in smaller ones [28]. For instance, larger language models are capable of summarising and translating texts, generating code, finding patterns in a text and understanding humour, while smaller models are limited to answering to simpler questions or generating seemingly grammatically correct text. Some have also attempted to explain such ostensibly emergent abilities with the better memorisation capacity and improved steerability via prompts characteristic to larger models [29]. Until the adoption of model weight quantisation and model pruning, such features were thought to appear in a language model from ∼7 billion parameters, even though certain emergent properties had been observed in the 1.5 billion parameter GPT-2. Today, however, it has become clear that smaller or compressed models may also possess such abilities to a certain extent.

A language model with a higher number of parameters requires a larger training dataset for the efficient utilisation of these parameters. Larger English-language training sets comprise trillions of tokens, whereas the size of Estonian datasets does not exceed several billions. This means that a language model trained based on the Estonian language will generally be smaller and less capable. The share of Estonian in multi-language datasets is very small, meaning that a model trained on the Estonian language may not always be capable of mastering the language. One strategy for mitigating this disparity is fine-tuning models trained on an English-language dataset using Estonian-language data.

Increasing hardware requirements. Computer GPUs use the SIMD (single instruction, multiple data) architecture which allows the same operation to be performed simultaneously on several pieces of data. This allows significantly speeding up rendering workflows and other graphics-related tasks where a certain operation needs to be repeated on each buffer element. This feature did not go unnoticed for the researchers of deep learning neural networks who proposed

the idea in 2009 that matrix operations frequently used in neural networks could be sped up using graphics processing units [30].

For each new token being generated, transformer-based large language models must access all the weights and attention vectors (q, k, v) employed by the model and move them from RAM to GPU registers. A sufficiently large number of sufficiently large weight matrices will result in increased loading times. This makes memory capacity and throughput comparable in importance to plain FLOPS[6].

Unlike fine-tuning, in-context learning does not require the computationally expensive updating of model weights alongside inference (prediction). The in-context learning functionality of certain LLMs can also be implemented on a high-performance personal computer[7]. Quantisation [31] – reduction in the accuracy and memory requirements of model parameters – is used to facilitate fitting the model weights in the PC's GPU memory. E.g. 16-bit floating point numbers are used in place of 32-bit ones; the most powerful quantisation methods re-encode the parameters so that a single parameter will only require a bit more than 2 bits of memory [32]. On the downside, the model's abilities may suffer from quantisation.

The expansion of fields employing parallel processing (machine learning, simulations, scientific modelling, cryptocurrency mining) has increased demand for both hardware and firmware suitable for the task. Nvidia has thus developed the CUDA platform comprising both hardware components and a software framework for the utilisation of GPUs in parallel processing tasks. Apple had developed the OpenCL parallel processing standard that, unlike CUDA, was not based on a specific type of hardware but today, they, too, have switched to their own hardware-specific framework, called Metal.

Classic server architectures are no longer adequate for offering AI as a cloud service. Extremely large volumes of data also mean that specialised data centres or cloud services are used for data storage and processing. When scaling a service, cloud infrastructure and specialised hardware are recommended for both inference and training. Meanwhile, specialised hardware no longer means only GPUs – it also covers solutions even more specific to neural networks, such as the tensor processing unit (TPU) developed by Google, or the neural processing unit (NPU) used in smartphones and Internet-of-Things (IoT) devices.

2.6.2 From general-purpose to special-purpose

From foundation models to applications. Foundation models are often mentioned in the context of LLMs. These are general-purpose models that can be used for performing many different tasks. Chatbots are one of the most basic applications of foundation models, as they only require command of natural language and general knowledge that can be derived from model weights and do not require a separate database interface. Non-deterministic model output is also acceptable in chatbots. In domain-specific applications, the generalisation ability and knowledge of the foundation model may not always be adequate for the task. Specialised solutions and models have therefore been developed alongside and based on large foundation models. These are especially good at processing medical and legal text, summarising large

[6] FLOPS (floating point operations per second) is a measure of computer performance.

[7] llama.cpp is an open-source application that facilitates running inference on LLaMA, LLaMA 2, and other language models using quantisation.

documents[8], programming languages and patterns[9], image recognition[10], and can evaluate the likelihood of an image or text being created by a generative model[11].

Simpler solutions have also emerged that connect to an existing AI model using its API, e.g., for interacting with and summarising documents in the form of PDF files. The business risk involved in such 'thin' solutions is that the providers of APIs and models can easily implement such functionality in their own products, just like OpenAI has done with the analysis of PDF files in ChatGPT[12].

From the synthesis of a single type of content to the creation of heterogenous content. When a model interacts with different input or output modalities it can be classed as multimodal. In other words, even a simple image classifier could be considered multimodal in that it receives an images as an input and outputs a text label. In reality, the term is mainly used for models where inputs with different modalities are mapped to the same embedding, such as OpenAI CLIP[13] and GPT-4V[14]. Multimodal text-to-video models also exist that generate an image sequence corresponding to the prompt, either relying on reference images [33] or without [34, 35].

Whereas multimodal inputs have been simple to process thus far, generating an output comprising different modalities is more difficult. The most common (and easiest) solution so far is the combination of the outputs and inputs of multiple models. Thus, ChatGPT comprises an image generation functionality where textual instructions generated using the GPT-4 language model based on a user prompt are fed to the DALL-E 3 image synthesis model which will then return the generated images to the users. The Invideo AI service[15] (alongside several other similar services) composes videos based on input text: it generates a script based on a user prompt and searches the database for clips which are then assembled into a video, after which it also generates a soundtrack.

One option for combining AI services is an AI agent (in some cases a generative agent) capable of interfacing with different services, e.g. making Internet queries for performing the task it has been given. AI agents are characterised by a continuous feedback cycle between making queries (interfacing with the outside environment) and updating their internal state. For this reason, it is vital for AI agents to be capable of planning their next steps while also keeping track of the results of the previous steps, their internal state, and the broader contents and purpose of the task [36]. A self-driving car can be considered an AI agent.

These days, AI agents generally mean solutions based on large language models that facilitate automatising multi-step actions requiring the division of tasks into subtasks, additional planning, and constant feedback based on natural language instructions. Some of the currently popular (as of writing this report) frameworks for creating and managing AI agents include AutoGPT, BabyAGI, and AiAgent.App.

[8] Claude 2: https://www.anthropic.com/index/claude-2
[9] GitHub Copilot X: https://github.com/features/preview/copilot-x
[10] Gpt-4Vision: https://openai.com/research/gpt-4v-system-card
[11] Stable Signature: https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/.
[12] ChatGPT Plus members can upload and analyze files in the latest beta. https://www.theverge.com/2023/10/29/23937497/chatgpt-plus-new-beta-all-tools-update-pdf-data-analysis Visited February 25th, 2024
[13] CLIP: Connecting text and images. https://openai.com/research/clip
[14] GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card
[15] Invideo AI. https://invideo.io/

2.6.3 From closed to open

Models for providing access to closed models. The larger AI models got, the more expensive their training, management, and deployment became. The more powerful they got, the greater the risks of exploiting their generative capabilities became. OpenAI was founded in 2015 as a non-profit with the goal of researching artificial intelligence and a main focus on deep learning neural networks[16]. In the early days, the organisation put a stress on openness and creating value for the whole society.

On 8 April 2019, a few months after the announcement and unveiling of the GPT-2 language model, the decision was made to split the company into a 'limited profit' company (OpenAI LP) and a non-profit (the existing OpenAI Nonprofit), with the board of the latter remaining the governing body of the two new partner organisations[17]. This step was purportedly taken because of the high maintenance expenses of modern AI systems: training these systems is compute-intensive, maintenance of the big data infrastructure used for the training is costly, and an NGO's opportunities for raising capital are far exceeded by those of companies. This was followed by a partnership with Microsoft who invested one billion US dollars into the company, and another 10 billion dollars in 2023.

GPT-2 was OpenAI's last completely open language models. In 2020, OpenAI released GPT-3, but the parameters of the trained model were not made accessible to the public – access to the model was limited to the OpenAI API[18] and GPT-3 itself licensed to Microsoft[19] under the cooperation agreement signed earlier. The decision to create an API was motivated by security requirements, as well as financial considerations. As the maintainer of the API, OpenAI retains the right to restrict access to the model to exploiters; the API was also the first commercial product of OpenAI LP that helped fund further research and maintain the expensive server infrastructure.

Emergence of public models. In 2023, Meta announced its own series of language models, LLaMA[20], surprising the world by making the models completely publicly accessible, even for commercial use. The licence of the LLaMA 2 model series released a few months later excluded companies with more than 700 million annual users in order to protect Meta from its biggest competitors. The same year also saw the release of the source code and parameters of stability.ai's generative image model, Stable Diffusion[21]. The emergence of models far surpassing GPT-2 in their capabilities, such as LLaMA 2, has unleashed an avalanche of smaller but, in some ways, more powerful AI models fine-tuned for specific areas of use. The performance of these models is only marginally inferior to foundation models with a much higher number of parameters. Mistral-7B [22] and SSD-1B[23] are great examples of such models.

Hobbyists, small enterprises, and research institutions can hardly afford the information infras-

[16] OpenAI. https://openai.com/blog/introducing-openai Visited October 20th, 2023
[17] OpenAI LP. https://openai.com/blog/openai-lp Visited October 23rd, 2023
[18] OpenAI API. https://openai.com/blog/openai-api Visited October 23rd, 2023
[19] OpenAI licenses GPT-3 technology to Microsoft. https://openai.com/blog/openai-licenses-gpt-3-technology-to-microsoft Visited October 23rd, 2023
[20] Introducing LLaMA: A foundational, 65-billion-parameter large language model. https://ai.meta.com/blog/large-language-model-llama-meta-ai/ Visited October 24th, 2023
[21] Stable Diffusion Public Release. https://stability.ai/blog/stable-diffusion-public-release Visited October 24th, 2023
[22] Mistral AI. https://mistral.ai/ Visited October 24th, 2023
[23] Announcing SSD-1B: A Leap in Efficient T2I Generation. https://blog.segmind.com/introducing-segmind-ssd-1b/ Visited October 24th, 2023

tructure or training budgets of the likes of OpenAI, Google, or Meta, which has caused a shift in focus from the number of parameters to their efficient use, the quality of training data, and alternative model architectures. As demonstrated by Google's leaked 'We Have No Moat'[24] memo, their success has been a cause for concern for large corporations. The emergence of more efficient and cheaper fine-tuning methods, such as LoRA [37], has allowed hobbyists to keep up with large technology companies in spite of the gap in investment capacity.

Motivated by, on one hand, the technology industry's desire to use AI on portable devices and, on the other hand, the limited resources of small enterprises and the open source community, a number of 'small language models' (SLMs) with fewer parameters have now emerged, such as Microsoft's Phi-1.5 [38] and Phi-2, Google's Gemini Nano[25] and Gemma [39], as well as Mistral 7B [40] and the Qwen1.5 family of SLMs [41] which are only slightly inferior in performance to much larger models.

2.6.3.1 Developments in deployment models

An AI model in itself is not sufficient for performing business tasks. The model must have access to input data and must be capable of producing properly formatted, high-quality output data. Deployment models refer to the structure of AI apps, interfaces between the AI model and other components of the app, and the flows of data between these components (including users' personal data).

The first, more primitive AI models (e.g. linear regression, perceptrons, rules-based expert systems) were not compute-intensive, making the information infrastructure for running the model less critical than data storage infrastructure. AI application deployment models only became relevant with the widespread adoption of AI in the 2010s, accompanied by growing datasets, proliferation of neural networks, and the resulting need to accelerate training and inference using GPUs that were not always readily physically accessible to the trainers or users of AI models. Alongside data storage and networking, cloud infrastructure providers began to offer hardware and cloud computing environments for AI models (e.g. Google Colab, Amazon SageMaker), but the users were still responsible for the development, training, and use of their models.

The general-purpose nature of subsequent large text and image synthesis models meant that for certain tasks, the model no longer needed to be trained from the ground up. This gave rise to AIaaS or AI as a service, allowing companies and individuals to use large AI models even without investments into hardware, training, and other information infrastructure.

The emergence of ChatGPT and AI APIs has triggered a deluge of thin 'API wrapper aps' using the generalisation ability of ChatGPT or another AI text synthesis solution for solving domain-specific tasks. Some of these applications provide little besides a convenient user experience and a carefully crafted pre-prompt; meanwhile, the reproducibility of such solutions creates significant business risks for the creators of wrapper apps. This risk materialised at the OpenAI Dev Day where OpenAI introduced a 'custom GPT' service allowing users to build special-purpose chatbots without writing a single line of code[26].

The business niche of AI service providers is not generally founded on innovative model architecture, as these are usually public, but the information infrastructure built around the model,

[24] Google: "We Have No Moat, And Neither Does OpenAI". https://www.semianalysis.com/p/google-we-have-no-moat-and-neither Visited October 26th, 2023

[25] Google Blog: Introducing Gemini https://blog.google/technology/ai/google-gemini-ai/ Visited December 14th, 2023

[26] Introducing GPTs. https://openai.com/blog/introducing-gpts Visited November 20th, 2023

the user experience provided by the solution, and the quantity and quality of domain-specific training data. The X (former Twitter) AI service Grok has real-time access to the database of user posts and Microsoft's Copilot X coding assistant would not be near as efficient without the constantly updated GitHub repository. ChatGPT, meanwhile, allows the user to give feedback to all chatbot's answers which has enabled OpenAI to collect large amounts of valuable data on users' interactions with the chatbot to facilitate the further improvement of the quality of their language models.

Training data quality management is vital as it allows significantly reducing the amount of data required for the training of an equivalent model [42], but also because the proportion of synthetic content on the Internet has risen sharply as of late and, according to experts, might reach 90 per cent by 2026 [43].

2.6.4 From unregulated to regulated

2.6.4.1 AI ethics

The ethics of computer science is a multifaceted, comprising both moral and ethical considerations related to the development, deployment, and use of computing technologies, such as AI. It is vital to ensure that these technologies are developed and used in ways that mirror human values and promote social wellness [44]. Ethical principles are dynamic, meaning that they can change in time, adapting to developments in science and the society [45].

The employment of AI technologies is on the rise – by 2027, the market capitalisation of the field is expected to reach 407 billion dollars [46]. Estonian companies are also increasingly using AI technologies – as at Q1 of 2023, the market has seen a 2% increase compared to 2021. According to Statistics Estonia, AI technologies are most frequently used in Estonia by finance and insurance, information and communication, and energy sector enterprises [47].

Even though artificial intelligence technologies demonstrate enormous potential, the use of AI also gives rise to numerous questions and fears. For example, a survey carried out in England in 2023 showed that people are the most worried about self-driving cars and autonomous weapons. They also fear that if AI is used for professional decision-making, the artificial intelligence may prove unable to account for individual real-world circumstances and decision-making may suffer from a lack of transparency and responsibility [48].

In 2018–2021, a scandal broke out in The Netherlands when it was found that the national tax office had used a flawed risk analysis algorithm in decision-making, resulting in thousands of child support receivers being baselessly accused of fraud [49]. This led to tens of thousands of families, often from lower income brackets or ethnic minorities, falling into poverty. Some of the victims performed suicide and over a thousand children where placed into foster families [50].

Professional decisions of this kind may also include court rulings. This raises the question whether a ruling made by an artificial intelligence is equivalent in quality to one made by a human judge, whether the system in question has been trained on high-quality data, and whether care has been taken to rule out discrimination on any grounds, such as gender, race, or income. Researchers have pointed out that AI models based on information derived from older input data are more likely to follow more conservative practices and may not be capable of adapting to significant political changes over time [51]. It has also be found that the use of AI for making court rulings may prove a threat to the integrity of data which, due to their very nature, would require the highest level of security [52].

It has been found that LLMs may tend to reinforce incorrect legal assumptions and beliefs which in turn gives rise to significant concerns over the reliability of the results in a legal context [53, 54]. The transparency and accuracy of the AI model also become critical in the context of trials [55].

Ethical issues emerging in the development, deployment, and use of AI are the subject of AI ethics which is considered one of the subdomains of applied ethics. The goal of AI ethics is to determine how an artificial intelligence system can increase or decrease human well-being through changes in quality of life or autonomy and independence. Different AI ethics frameworks are generally built around fundamental rights [45].

On April 8th, 2019, the EU High-Level Expert Group on AI (hereinafter AI HLEG) presented its ethics guidelines for trustworthy AI [45, 56] with the goal of providing guidance for promoting and supporting ethical and robust artificial intelligence. Less attention is paid to the legal aspects of the system. The document presents a preliminary framework for trustworthy AI while also discussing issues related to the implementation and evaluation of AI systems [45].

2.6.4.2 AI regulation in the EU

In April 2021, the European Commission proposed the first legal framework regulating AI [57]. The proposal was built around a risk-based approach, asserting that artificial intelligence systems should be analyzed and classified based on the threat they pose to users [58]. Negotiations over the AI Act ended on December 8th, 2023. In early 2024, the AI Act is expected to be published in the Official Journal of the European Union.

Neither should one overlook the existing legal framework. More specifically, the General Data Protection Regulation (GDPR) of 2016 [59] stresses the importance of the protection of natural persons in the automated processing of personal data[27]. In addition to the the above, the development, implementation, and use of artificial intelligence must also account for other requirements, such as intellectual property rights. For more details on the legal aspects of artificial intelligence, see Section 3 of the report.

[27] GDPR regulates the automated processing of personal data, including profiling, and confers on the data subject the right to oppose individual decisions based on such processing (see GDPR articles 2, 21, and 22, and recitals 15 and 71).

3 Legal aspects

3.1 International legal initiatives

3.1.1 Regulation

Experience from recent years indicates that AI regulation is rapidly developing all over the world. The examples presented below pertain to just some of the states regulating AI systems.

On October 30th, 2023, the President of the United States Joe Biden issued an executive order to ensure that the US maintains a leading position in the world in AI systems. The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more [60].

The UK Parliament has published a bill to regulate the use of AI technologies in the workplace and make provision about workers' and trade union rights in relation to the use of artificial intelligence technologies. The first reading of the bill took place on May 17th, 2023 [61, 62]. In September 2023, the UK government published a white paper on a pro-innovation approach to AI regulation. This framework is underpinned by five principles [63]:

1. safety, security and robustness;
2. transparency and explainability;
3. fairness;
4. accountability and governance;
5. contestability and redress.

Discussions over the regulation of artificial intelligence are also underway in Australia [64]. In 2022, the Australian government published a consultation on the rules for artificial intelligence and automated decision-making. The consultation was driven by the Australian government's digital economy strategy laying out an ambitious vision Australia becoming one of the 10 best digital economies and societies by 2030 [65, 66]. According to the new draft law of search engines presented on September 8th, 2023, the Australian government requires Internet search service providers to review and regularly update their artificial intelligence tools in order to ensure that class 1A materials (e.g., materials related to the sexual abuse of children, support of terrorism, and extreme violence) are not returned in search results. The draft act also mandates that users must be able to identify whether an image accessible through a search engine is a deep fake [67, 68, 69].

In September 2023, Canada published a voluntary code of conduct on the responsible development and management of generative AI systems [70]. Work is also on the way on the Artificial Intelligence and Data Act (AIDA) that would set the foundation for the responsible design, development and deployment of AI systems that impact the lives of Canadians [70]. The act would ensure that AI systems deployed in Canada are safe and non-discriminatory and would hold businesses accountable for how they develop and use these technologies. In addition to the above, on October 12th, 2023, the Canadian government announced a public consultation on the effects of generative artificial intelligence on copyright [71].

Alongside the above-listed states, legal initiatives related to AI systems have also been undertaken in Israel, Japan, China, Chile, Mexico, Peru, Singapore, and other places [72]. EU legal

acts on artificial intelligence systems are covered in Section 3.3 of the report.

3.1.2 Standards

Turning our attention next to approaches to AI found in international soft law, various non-binding recommendations and guidelines have been published to promote the development and adoption of ethical, responsible, and trustworthy AI. These are generally founded on principles like privacy, explainability, impartiality, security, and being human-centered.

One of such standards is ISO/IEC 22989 establishing terminology for AI and describing concepts in the field of AI [73]. Common terminology ensures better understanding of AI systems and is vital to cooperation, regulation, adoption of responsible AI systems, and information sharing [74]. The ISO/IEC 23053 standard describes artificial intelligence systems using machine learning [75]. The standard describes the components of a machine learning system and their functions in the AI ecosystem [74].

Next, the ISO/IEC 5259 standard establishes a framework for ensuring data quality in analytics and machine learning [76, 77]. ISO/IEC 4213 describes the requirements for evaluating classification performance in machine learning [78]. Various other standards and frameworks also exist, such as the BSI validation framework BS 30440:2023 for the use of artificial intelligence within healthcare [79], the IEEE ethical design standard [80], Google AI principles [81] and responsible AI practices [82] and the Microsoft responsible AI standard [83].

Adherence to standards will contribute to the safety, quality, and reliability of products or services; they can also help enhance and improve the company's systems and processes. Standards applicable to the different life cycles of artificial intelligence systems are covered in the ENISA good cybersecurity practices for AI systems [84].

3.2 EU trustworthy AI initiative

On April 8th, 2019, the EU high-level expert group on artificial intelligence (AI HLEG) presented its ethics guidelines for trustworthy AI [85] covering an overall framework for and implementation and evaluation of trustworthy artificial intelligence [86]. According to the ethics guidelines, the life cycle of a trustworthy AI system should be [86]:

1. lawful – respecting all applicable laws and regulations;
2. ethical – respecting ethical principles and values; and
3. robust – both from a technical perspective while taking into account its social environment.

Section I of the guidelines sets out the three main ethical principles founded on fundamental rights. First, the development of AI systems must respect human autonomy, ensure the fairness and explainability of the system, and prevent harm. The second principle requires paying particular attention to situations involving more vulnerable groups (such as children, persons with disabilities) and situations which are characterised by asymmetries of power or information. Finally, attention is drawn to the risks posed by AI systems and the adoption of measures to mitigate these risks [86].

Section II of the ethics guidelines presents an overview of how to create a trustworthy AI system, and proposes seven criteria or such a system.

1. Above all, it is recommended to ensure that the development, deployment and use of AI systems meets the seven key requirements for trustworthy AI: *'(1) human agency and oversight,*

(2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.' [86].

2. Using both technical and non-technical methods to ensure the implementation of those requirements is recommended.
3. Research and innovation should be fostered to increase the amount of knowledge available about AI systems – among other things, or the training of new AI ethics experts.
4. Clear information should be provided on the capabilities and limits of the AI system to enable setting realistic expectations.
5. Systems should be developed to be explainable to facilitate their auditability which may prove particularly vital in critical situations.
6. Stakeholders should be involved throughout the AI system's life cycle, and people should be trained to increase their awareness of trustworthy AI.
7. It has to be taken into account that tensions might arise between the different principles and requirements for trustworthy AI. It is recommended to continuously document all considerations, trade-offs, and decisions [86].

Section III of the ethics guidelines provides an assessment list for operationalising trustworthy AI, to be adapted based on the purpose of the AI system. Compliance should be assessed, stakeholders involved, and results continuously improved throughout the entire life cycle of an AI system [86]. The trustworthiness of an AI system depends on all of its features; unfortunately, the exhaustive understanding of compromises between these features still remains an important unsolved problem [87].

The final section of the ethics guidelines elaborates upon some of the issues addressed in the document, offering examples of beneficial opportunities that should be pursued, and discussing critical concerns raised by AI systems that should be carefully considered [86]. The EU high-level expert group has also published policy and investment recommendations for trustworthy artificial intelligence explaining how trustworthy AI should be developed, deployed, promoted, and expanded in Europe while maximising its benefits and minimising and preventing possible risks [88, 89]. On July 17th, 2020, the AI HLEG additionally published their assessment list for trustworthy AI (ALTAI) [90]. The ALTAI is a tool that facilitates evaluating the extent to which an AI system meets the requirements for trustworthy AI. These guidelines are also available in a web-based tool version [91].

They also published a document on sectoral considerations regarding policy and investment recommendations, analyzing the potential application of recommendations previously published by the AI HLEG in three specific sectors: (1) the public sector, (2) healthcare, (3) manufacturing and Internet of Things (IoT) [92].

On the 19th of February in 2020, the European Commission published a report on the safety and liability implications of artificial intelligence, the Internet of Things and robotics [93]. All products and services must operate safely, reliably and consistently, and any damage must be remedied – these are the goals of legal frameworks for safety and liability. According to the Commission, a clear safety and liability framework is particularly important when new technologies emerge, both with a view to ensure consumer protection and legal certainty for businesses [93].

On the same day, the EC also published a white paper on artificial intelligence [94] discussing aspects related to the most important outputs of data economy – artificial intelligence, a collection of technologies that combine data, algorithms and computing power. The white paper

notes that the use of digital technologies is based on trust and discusses how action needs to be stepped up at multiple levels in order to support the uptake of AI [94].

3.3 EU proposal for an Artificial Intelligence Act

A number of legal proposals related to AI have been proposed in the EU with the goal of ensuring that artificial intelligence systems used in the EU are safe, transparent, ethical, impartial, and human-controllable [95].

In April 2021, the European Commission presented a proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [57]. According to the explanatory memorandum, the act would set down harmonised requirements following a proportionate risk-based approach to the development, placing on the market, and use of AI systems in the EU [57]. On December 8th, 2023, a political agreement was reached on the final text of the act [96, 97], followed by technical discussions on finalising the text. Particular attention was paid to the question of a threshold for high-impact general-purpose AI (GPAI) models, which was decided to be established based on the cumulative amount of computing power used for the training (10^25). Harmonised standards for the regulation of GPAI models will be developed in the future [98].

On January 26th, 2024, the Belgian Presidency of the Council of the EU officially shared the final compromise text of the AI Act with member states' representatives [99]. On February 2nd, 2024, the AI Act was adopted by the Committee of Permanent Representatives (COREPER). The compromise was based on a multi-level approach comprising horizontal transparency rules for all models and additional requirements for AI systems posing a potential systemic risk [98].

The AI Act proposal [57] serves four main objectives.

1. The first goal is to ensure that AI systems placed on the EU market and used are safe and meet existing laws and EU values.
2. Next, it should ensure legal certainty to facilitate investment and innovation in AI.
3. Third, it should enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems.
4. Finally, it should facilitate the development of a single market for lawful, safe and trustworthy AI applications.

According to the proposal, artificial intelligence systems would be divided into four risk categories in order to establish requirements consistent with the risks involved (see Table 3). In the course of the negotiations, the text of the AI Act was amended with provisions concerning non-systemic and systemic risks related to general-purpose AI systems [99].

In the final compromise text of the AI Act [99], an AI system is defined as a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments (See Article 3 (1))[1].

[1] Hereinafter, requirements for AI systems are discussed in the form they are found in the final compromise text of the AI Act, insofar as the official adopted version of the regulation was yet to be published in the Official Journal of the EU at the time of preparing this report. It must be kept in mind that the specific articles, points, or recitals of the compromise text cited here may differ from the text of the AI Act published in the Official Journal, as the numbering in the compromise text has not been corrected. – Accessible on the Internet: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf Last visited February 24th, 2024

The cited compromise text [99] states that the purpose of the regulation is to promote the uptake of human-centered and trustworthy artificial intelligence while promoting innovation and ensuring a high level of protection of health, safety, fundamental rights, democracy, rule of law, and the environment against harmful effects of artificial intelligence systems. The regulation sets out harmonised requirements for placing on the market, putting into use, and use of AI systems in the EU. It prohibits certain uses of artificial intelligence, lays down specific requirements for high-risk AI systems, and the obligations of the operators of such systems. It also sets out harmonized transparency standards for certain AI systems, and requirements for the placing on the market of general-purpose AI models. The regulation also lays out rules for market surveillance and monitoring and measures for supporting innovation, with a main focus on small and medium enterprises, including starts-ups.

3.3.1 Persons falling within the scope of the AI Act

The following persons fall within the scope of the AI Act:

1. providers placing AI systems on the market in the EU or using them in their services or placing on the market a general-purpose AI model, irrespective of whether they are established or located within the EU or in a third country;
2. deployers of AI systems operating or established within the EU;
3. providers and deployers of AI systems operating or located in a third country, to the extent that the output of their AI system is used within the EU;
4. importers or distributors of AI systems;
5. product manufacturers who are placing on the market or putting into use AI systems along with their product under their name or trademark;
6. authorised representatives of providers established outside the EU and
7. affected persons located within the EU.

Article 3 of the AI Act sets out a number of new terms, including the definitions of deep fakes and AI literacy, as well as training, validation, testing, and input data. AI literacy is even the subject of a separate article (Article 4b) that obligates the providers and deployers of AI systems to take measures to, e.g., ensure a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems.

Below, we have presented some of the more important requirements for AI stakeholders found in the final compromise text of the AI Act [99].

3.3.2 Exclusions from the scope of the AI Act

The regulation does not apply to deployers who are natural persons using AI systems in the course of a purely personal non-professional activity. It also does not apply to, e.g., AI systems used solely for military, defence or national security purposes. Excluded from the scope of the AI Act are also AI systems and models, including the outputs of such models, specifically developed and put into service for the sole purpose of scientific research and development. It also does not apply to scientific research, testing and development activity on AI systems or models prior to being placed on the market or put into service, without prejudice to the testing of AI systems in real-life conditions. Finally, the scope of the regulation does not include AI systems released under free and open source licences, without prejudice to systems placed on the market or put into service as e.g., high-risk AI systems.

3.3.3 Prohibited artificial intelligence practices and uses

The regulation prohibits a number of AI practices (see Article 5 for details). These include prohibitions on uses of AI systems that purposefully manipulate with a person with the objective to distort their behaviour and appreciably impair the person's ability to make an informed decision. The regulation also prohibits AI systems exploiting any of the vulnerabilities of a person or a specific group of persons due to their age, disability or a specific social or economic situation. Another prohibition is related to the use of biometric categorisation systems that categorise natural persons based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation. AI systems are also not allowed to be used for the classification of natural persons based on their social behaviour or personality characteristics along with a social score leading to the detrimental or unfavourable treatment of the person.

3.3.4 Criteria for high-risk AI systems

Criteria for the classification of AI systems as high-risk are laid out in Article 6 of the regulation proposal. For example, an AI system is always considered a high-risk system if it is intended for the profiling of natural persons. A provider who considers that an AI system referred to in Annex III is not high-risk must document its assessment before that system is placed on the market or put into service. Such provider is subject to the registration obligation set out in Article 51(1a) and upon request of national competent authorities, the provider must provide the documentation of the assessment. No later than 18 months after the entry into force of the AI Act, the European Commission must provide guidelines specifying the practical implementation of Article 6 completed by a comprehensive list of practical examples of high risk and non-high risk use cases on AI systems.

Article 9 sets out requirements for risk management systems for high-risk AI systems. According to point 2 in the article, the risk management system is understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating. It comprises the following steps:

a) identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to the health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose;

b) estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse;

c) evaluation of other possibly arising risks based on the analysis of data gathered from the post-market monitoring system (see Article 61) and

d) adoption of appropriate risk management measures.

High-risk AI systems must meet the requirements set out in the AI Act (see Chapter 2), taking into consideration the purpose of such systems, as well as the level of AI and related technologies. More specifically, the risk management measures must be such that relevant residual risk associated with each hazard as well as the overall residual risk is judged to be acceptable (Article 9(4)).

High-risk AI systems must also be tested for the purposes of identifying the most appropriate risk management measures (Article 9(5)). Testing procedures may include testing in real

world conditions (Article 9(6); see also Article 54a). Consideration must also be given potential impacts to persons under the age of 18 and other vulnerable groups of people (Article 9(8)).

High-risk AI systems which make use of techniques involving the training of models with data must be developed on the basis of training, validation and testing data sets that meet the quality criteria set out in the AI Act (Article 10(1)). Training, validation and testing data sets must also be subject to appropriate data governance and management practices appropriate for the intended purpose of the AI system, e.g., to detect, prevent and mitigate possible biases (Article 10(2)(fa)).

Training, validation and testing datasets must be relevant, sufficiently representative, and, to the best extent possible, free of errors and complete in view of the intended purpose, as well as possessing the appropriate statistical properties (Article 10(3)).

The processing of special categories of personal data for the purposes of ensuring bias detection and correction in high-risk AI system is subject to strict regulation. It must meet all EU data protection regulations and for such processing to occur, criteria set out in points (a)–(f) of Article 10(5) must be fulfilled. First, it must be explained why the bias detection and correction cannot be effectively fulfilled by processing other data, including synthetic or anonymised data.

Special categories of data must be processed using state-of-the-art security and privacy-preserving measures, including pseudonymisation, or privacy enhancing technologies. Measures must be taken to ensure the security of the data, including including strict controls and documentation of the access to avoid misuse and ensure only authorised persons have access to those personal data with appropriate confidentiality obligations. Such data are not to be transmitted, transferred or otherwise accessed by other parties. The data must be deleted once the bias has been corrected or the personal data has reached the end of its retention period, whatever comes first.

The technical documentation of a high-risk AI system must be drawn up before the system is placed on the market or put into service and has to be kept up-to date. The documentation must contain, at a minimum, the elements set out in Annex IV (Article 11(1)). High-risk AI systems must technically allow for the automatic recording of events (logs) over the duration of the lifetime of the system (Article 12(1)).

High-risk AI systems must be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable deployers to interpret the system's output and use it appropriately (Article 13(1)). High-risk AI systems must be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users (Article 13(2)). Said instructions must correspond to the minimal requirements set out in Article 13(3) of the regulation.

High-risk AI systems must be equipped with means to ensure that they can be effectively overseen by humans during the period in which the AI system is in use (see human oversight requirements and principles set out in Article 14). For example, humans need to be able to intervene in the operation of a high-risk AI system or interrupt the operation of the system through a 'stop' button or a similar procedure (Article 14(4)(e)).

High-risk AI systems must be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and perform consistently in those respects throughout their life cycle (Article 15(1)). Such systems need to be resilient as regards to attempts by unauthorised third parties to alter their use, outputs or performance (Article 15(4)).

Article 21 of the regulation mandates that providers of high-risk AI systems which have reason to consider that a high-risk AI system which they have placed on the market or put into service is not in conformity with the AI Act must immediately take the necessary corrective actions, e.g., to bring that system into conformity or to disable it. The provider must also inform distributors and, if applicable, deployers, authorised representatives, and importers of the system.

3.3.5 Requirements for participants in the AI value chain

The AI Act also sets out a variety of requirements for other AI system stakeholders, such as deployers, authorised representatives of non-EU providers, importers, and marketers. It is therefore important to assess any specific person's role in the AI value chain in accordance with the AI Act to identify the specific requirements they need to follow.

The AI Act is a new addition to the EU law; the new norms and those implementing these norms thus need some time to adapt to the new situation. This will hopefully be facilitated by the European AI Office – the centre of AI expertise across the EU. The AI Office plays a central role in the implementation of the AI Act, supporting the development and use of trustworthy AI and international cooperation [100].

3.4 AI Liability Directive proposal

In order to mitigate AI-related risks, the AI Act proposal was followed by a proposal for a directive on AI liability in September 2022 [101], the aim of which is to ensure that persons harmed by AI systems have reasonable means available for protecting their rights. The directive would harmonise national norms for non-contractual liability. It is also meant to increase legal certainty for businesses developing or using artificial intelligence.

One of the measures foreseen by the directive is to expediate court proceedings for victims who have been harmed by an AI system. The victims will be able to claim compensation both individually or collectively, as appropriate. If a violation has taken place and a potential causal link exists to an AI system, a rebuttable presumption of causality will be applied. More specifically, a presumption of causality can only be applied when it can be considered likely that the given fault has influenced the relevant AI system output or lack thereof, which can be assessed on the basis of the overall circumstances of the case. At the same time, the claimant still has to prove that the AI system (i.e. its output or failure to produce one) gave rise to the damage [101].

The proposed directive will also provide better opportunities for ensuring legal protection. For instance, a court may order the disclosure of relevant evidence to a victim to determine the cause of the damage and identify which person is liable for compensating the damage.

3.5 Product safety

Regulation (EU) 2023/988 of the European Parliament and of the Council of May 10th, 2023 on general product safety [102] lays down essential rules on the safety of consumer products placed or made available on the market (Regulation (EU) 2023/988, Article 1(2)). Recital 5 of the regulation notes that *'[d]angerous products can have very negative consequences for consumers and citizens. All consumers, including the most vulnerable, such as children, older persons or persons with disabilities, have the right to safe products. Consumers should have at their disposal sufficient means to enforce that right and Member States should have adequate instruments and measures at their disposal to enforce this Regulation'*.

On September 28th, 2022, the European Commission published a proposal for enacting a directive on liability for defective products [103]. The objective of this directive is to lay down the rules governing the liability of economic operators for damage caused by defective products and the conditions under which natural persons have a right to compensation. The directive also foresees solidary liability. According to the directive, economic operators are liable for defective products for 10 years following placing the product on the market.

The explanatory memorandum for the proposed directive explains that one of its objectives is also to ensure liability for defects in artificial intelligence systems which have caused physical harm, property damage, or data loss. In such situations the user will have the right to seek compensation from the provider of the AI system or any manufacturer integrating an AI system into another product. The scope of the proposal also includes software providers, businesses that make substantial modifications to products, authorised representatives, and fulfilment service providers, giving injured persons a better chance of being compensated for damage suffered[2].

3.6 Intellectual property

The purpose of intellectual property law is to protect the creations of the mind. Generative artificial intelligence has changed the society's understanding of creativity and property rights, raising questions regarding human input and intellectual property [104]. At the time of preparing this report, the interactions between intellectual property rights and artificial intelligence have become one of the main areas of development of intellectual property law, mainly thanks to developments related to AI, initial relevant case law, and political initiatives undertaken by international organisation and legislators [105].

In recent years, legal scholars have increasingly turned their attention to issues related to artificial intelligence and intellectual property. This can be divided into two main categories.

1. Legal protection for automated creation – e.g. are there any circumstances under which AI-generated works could be subject to copyright or inventions they have created be patented?
2. Intellectual property violations – e.g. how to efficiently protect the holders of intellectual property rights from the developers of artificial intelligence systems who use works protected under intellectual property law for training their AI systems without the rightsholder's knowledge and/or consent?

Generative AI capable of writing cohesive texts, creating art or architectural designs has given rise to all manners of questions regarding the nature of intellectual property and has become a cause for legal disputes. Examples exist of both cases of authors taking legal action against AI developers who have used prohibited data or works for developing their AI systems (e.g. used copyrighted texts, images, etc. without permission) [106, 107], as well as cases of intellectual property rights being claimed for AI-generated works [108].

Current intellectual property law generally gives no considerations to creators like AI systems. The regime in place today was created to promote human creation and innovation. From the perspective of the intellectual property system, AI's autonomy raises fundamental questions about all forms of intellectual property rights [109]. Meanwhile, strong interactions and correlation can be observed between AI systems and intellectual property law [110]. In most cases, the following two principles are considered critical: the originality of the work, the dichotomy of idea and expression, and rendering the above in a human-perceptible form [111].

Thus, in accordance with Section 4(2) of the Estonian Copyright Act, works mean *'any original*

[2] See explanatory memorandum for the proposed directive, Section 1.2 and Chapter 2.

results in the literary, artistic or scientific domain which are expressed in an objective form and can be perceived and reproduced in this form either directly or by means of technical devices. A work is original if it is the author's own intellectual creation.' One of the proposed solutions to these issues is the adoption of a hybrid ownership model (AiLE) [111]. Other have, meanwhile, found that the addition of new layers to the existing intellectual property rights system is not a good solution for balancing the social impact of technological progress [112], and that the creations of AI are not protectable [113]. The European Parliament finds it important to distinguish between AI-assisted human creations and creations autonomously generated by AI [114].

Time will tell what the future will bring for intellectual property rights as AI systems continue to develop. It is clear, however, that there is an abundance of different opinions regarding intellectual property rights and there are currently no simple solutions on offer. It cannot even be ruled out that now is not the right time to make such decisions, that developments related to AI systems require careful consideration and a certain level of maturity from the society before any changes are made to functional legal systems.

3.7 Legal requirements for cybersecurity

Just like with other information systems, the security of artificial intelligence systems starts from ensuring confidentiality, availability, and integrity. Depending on their roles, contexts, and operational capability, AI stakeholders should apply systematic risk management in every stage of the AI system's life cycle in order to handle risks to privacy, digital security and safety, and to prevent algorithmic bias [25].

In accordance with OECD recommendations, AI systems should remain secure, reliable, and safe throughout their entire life cycle. This applies to both routine and planned use as well as abuse and unfavourable conditions. Ensuring the monitorability of the AI system is critical for ensuring the above. It applies equally to the data or datasets used, various processes and decisions, and allows performing context-specific analyses of the operation of an AI system, e.g. its outputs or reactions to queries [25].

ENISA lists the following types of threats to ICT infrastructures [84]:

- adversarial threats – these are results of malicious intentions (e.g. denial of service attacks, non-authorised access, masquerading of identity);
- accidental threats – these are caused accidentally, e.g. through human error, or through legitimate components. Usually, they occur during the configuration or operation of devices or information systems, or the execution of processes;
- environmental threats – these include natural disasters (floods, earthquakes), human-caused disasters (fire, explosions), and failures of supporting infrastructures (power outage, communication loss);
- vulnerabilities – existing weaknesses of AI systems that might be exploited by an adversary.

Various legal acts have been enacted in Europe to react to such threats. The Second Cybersecurity Directive (NIS2) [21] and the Cybersecurity Act [115] are considered to be the two most important legal acts on cybersecurity in Europe. Another key legal act is the General Data Protection Regulation (GDPR) [59]. Said legal acts stress the importance of supply chain security, privacy, and protection of personal data, all of which are also central the life cycle of artificial intelligence systems [84].

The NIS2 Directive entered into force on January 16th, 2023 and also covers artificial intelligence systems. More specifically, the directive seeks to promote the use of AI for, e.g., the discovery

and prevention of cyberattacks, and the planning of relevant resources[3]. Essential and important entities are recommended to adopt basic cyberhygiene practices and, where appropriate, integrate artificial intelligence or machine-learning technologies to enhance security[4]. NIS2 also requires AI use to comply with EU data protection law, including including the data protection principles of data accuracy, data minimisation, fairness and transparency, and data security, such as state-of-the-art cryptography. The requirements of integrated and default data protection laid down in the GDPR must also be followed [21]. An overview of NIS2 can be found on the website of Centre for Cyber security Belgium [116].

A proposal for a regulation of the European Parliament and of the Council on on horizontal cybersecurity requirements for products with digital elements (also known as the Cyber Resilience Act or the CRA) introduces a European cybersecurity certification framework for products and services [117]. The necessity for such regulation is explained with the low level of cybersecurity of products, services and an insufficient understanding and access to information by users on the security of these products and services. Article 8 of the CRA lays down requirements for high-risk AI systems.

Cybersecurity also occupies a central place in the AI Act proposal [118]. For instance, it plays an important role in ensuring the resilience of artificial intelligence systems to attempts to change their use, behaviour or performance, or put their security features at risk by malicious third parties seeking to exploit the system's vulnerabilities. Adversaries may thus target, e.g., training data (data poisoning), the trained models (adversarial attacks or re-identification attacks), or exploit the vulnerabilities of the AI system's digital assets or the underlying IT infrastructure. Adequate and efficient measures accounting for the current level of technology must be taken to ensure risk-appropriate cybersecurity.

3.8 Data protection and privacy

Data and privacy enhancement considerations must be taken into account wherever personal data are processed in any stage of the AI system's life cycle (e.g., in training or application). The main legal act regulating the processing of personal data in the EU is the General Data Protection Regulation [59]. On July 4th, 2023, the European Commission published a proposal for a regulation laying down additional procedural rules relating to the enforcement of the GDPR [119]. Special rules have also been established for law enforcement authorities [120] and EU institutions [121].

National data protection and privacy norms also have to be taken into consideration, and in some cases, sectoral requirements may also apply. Accordingly, for each specific sector and activity, it is vital to consider the special norms of the relevant field alongside requirements set out in the GDPR. Conditions agreed upon by different parties (e.g., contracts, data protection agreements, terms of service) must also be taken into account.

The deployment of artificial intelligence demands solutions for complex legal problems. Privacy and data protection are a few of the most urgent issues, especially in the light of GDPR rules. The GDPR introduces high standards for data protection which, in turn, have a great impact on AI systems dependent on large amounts of data [122]. To ensure an AI system's compliance with data protection requirements, it must take into account the personal data processing principles laid out in GDPR Article 5(1). The controller is responsible for must be able to demonstrate compliance with these principles (GDPR Article 5(2)). Personal data must be:

[3] See NIS2, recital 51.
[4] See NIS2, recital 89.

(a) *processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')*;

(b) *collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation')*;

(c) *adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')*;

(d) *accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy')*;

(e) *kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation')*;

(f) *processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality')*.

Considering the size of datasets used for developing and testing AI systems, it may prove difficult to ensure the compliance of AI systems with certain data protection rules (e.g., data minimisation, purpose and storage limitations). The rapid development of generative artificial intelligence and large language models has posed the question of adapting existing data protection rules in this new context.

Different data protection authorities have published guidance documents on following data protection principles and rules in the development, deployment, and use of AI systems. Some of these authorities include the French National Commission on Informatics and Liberty (CNIL) [123] and the UK Information Commissioner's Office (ICO) [124]. In early 2024, the ICO also launched a series of consultations on generative AI with the objective of determining how data protection rules should be applied in the development and use of AI technology [125]. The consultations study various aspects related to data protection, e.g., training generative AI models on web-scraped data, accuracy of generative AI outputs, implementation of the purpose limitation principle, guaranteeing data subjects' rights [126]. The consultations will be used to publish relevant recommendations.

Privacy and data protection need to be ensured throughout the entire life cycle of an AI system [45]. Privacy and data protection are especially important due to the fact that behavioural data may permit AI systems to infer not just a person's preferences but also other personal and relatively private information, e.g., sexual orientation, age, gender, religious beliefs or political views. It is therefore vital for AI systems to ensure that privacy and data protection requirements are met not only in the case of the initial data provided by the system's user but also the data produced in the course of using the system (outputs, reactions to recommendations,

etc.). Any kind of unlawful and unfair discrimination on the basis of data must be outruled [45]. There have been cases where AI systems have leaked sensitive information, e.g., conversation histories [87].

The EU AI HLEG has found that privacy issues are closely tied to the principle of prevention of harm. Relevant data management measures must be applied to ensure privacy, which includes managing the quality and integrity of the data being used, and access protocols [45].

The AI Act proposal includes an assessment of the need arising in certain situations to conduct evaluations of the impact of AI systems on fundamental rights and to carry out a data protection impact assessment [118]. The proposal finds that the conduction of such impact assessments must be planned as a part of an overarching process in order to reduce redundancy and unnecessary administrative burden. The future AI Office would be tasked with developing a questionnaire that the deployers of AI systems could use to meet the relevant criteria [118]. In any case, the development and use of AI systems must comply with existing privacy and data protection rules.

Since AI systems are founded on data, the quality of this data is critical. Data quality is also important for the creation of the structure of AI systems and ensuring their operability. Training, validation, and test data must be relevant, sufficiently representative, maximally error-free and complete from the point of view of the purpose of the AI system. The requirement for datasets to be maximally complete and error-free should not impact the use of privacy-preserving technologies in the context of the development and testing of AI systems [118].

It must also be taken into account that the compilation of datasets must be based on the lawful use of data in compliance with data protection regulations [127]. The processing of personal data is only lawful if at least one of the conditions of GDPR Article 6(1) (points a–f) is met. There have been cases where competent authorities have demanded the deletion of models based on unlawfully collected data [128]. In order to prevent any form of discrimination, the datasets should also possess the relevant statistical properties and account for features characteristic to the specific situation or group of persons.

In order to comply with GDPR requirements, an artificial intelligence system must be developed, trained, and put into service with a clearly defined purpose. The French National Commission on Informatics and Liberty (CNIL) recommends the purpose of the AI to be determined in the early planning stages of the projec . The purpose of the system must be lawful, clear, and understandable, and usable for determining which kinds of data need to be processed for this specific purpose, as well as how long they will have to be retained in order to achieve the envisaged objective [127].

Even though the principle of limited purpose requires using personal data only for achieving a specific predetermined goal, this may prove complicated in the case of an AI system. The CNIL has found that at the algorithm training stage it is not always possible to define all the possible future uses of the artificial intelligence; nevertheless, the type and main potential functions of the system should still be defined as clearly as possible [129].

Discussions revolving around the extraterritorial enforcement of the GDPR give reason to believe that the jurisdictional model implemented in said regulation which has also been introduced into the EU AI Act may not be applicable in practice [130, 131, 132, 133]. According to Article 3(2), points (a) and (b) of the GDPR, the regulation also applies to the processing of personal data of data subjects who are in the EU by a controller or processor not established in the EU if the processing is related to the offering of goods or services to such data subjects in the EU or the monitoring of their behaviour as far as their behaviour takes place within the EU.

In the course of the implementation of the GDPR, there have been numerous disputes over specifically the processing of personal data by controllers or processors who fall within the scope of Article 3(2) of the GDPR but who refuse to cooperate with European data protection authorities or do not recognise the EU's jurisdiction (see, e.g., the Clearview AI case) [134, 132]. The AI Act proposal also uses an approach similar to the GDPR where businesses from non-EU states are included within the scope of the regulation (see Article 2(1)(c)) [99]. In practice, competent authorities may be facing problems similar to those that have arisen in connection to the extraterritorial enforcement of the GDPR.

The transfer of personal data to non-EU states and international organisations is regulated by Chapter V of the GDPR. The transfer of data is generally permitted only if suitable legal grounds exist for such transfer (GDPR, Articles 6 and 9) and relevant and efficient protection measures are taken [135]. Article 45 of the GDPR gives the European Commission the right to determine whether a non-EU state or international organisation provides an adequate level of data protection [136, 137]. For example, in July 2023, the Commission adopted an adequacy decision for the EU-US Data Privacy Framework [138][5]. The existence of a relevant decision by the Commission removes the need for a specific authorisation for the transfer of data (GDPR, Article 45(1)). EEA states (Norway, Iceland, Liechtenstein) are considered to be states with an adequate level of data protection.

Additional safeguards must be implemented when transferring data to states lacking an adequate level of data protection (see, e.g. [139]), or one of the derogations laid down in the GDPR must be applicable (GDPR, articles 46–49) [140]. The European Data Protection Board (EDPB) has found that in certain situations remote access from a non-EU state (e.g., support services, troubleshooting), as well as storage in a cloud situated outside the EEA may be considered to be a transfer in the meaning of the GDPR [141]. It is therefore strictly advisable to plan out the AI infrastructure before entering into any agreements with service providers in order to avoid later legal disputes or sanctions.

3.9 Importance of the legal framework

Persons central to the life cycle of an AI system need to be up-to-date on the legal and regulatory requirements shaping the legal framework they operate in. This determines the requirements that the AI system as well as the person operating the system must meet. Various aspects of administering and managing processes related to the AI system, such as the development, testing, and monitoring of the system are also tied to the above.

A holistic approach to information technology, security, and legal issues is increasingly important for organisations. This also means close cooperation between people fulfilling the relevant roles from the stage of designing an AI system to the end of its life cycle. This, in turn, facilitates expanding legal specialists' knowledge of technology and vice versa, thus contributing to an increase of organisational knowledge.

The greater the awareness of the requirements related to the legal framework – even at the stage of designing an AI system – and the more said requirements are actually adhered to, the smaller the probability of the occurrence of undesirable scenarios. Meanwhile, it must be taken into account that AI law is still far from mature and the legal environment can be expected to continue to change.

[5] Earlier similar agreements and decisions between the EU and the US have repeatedly been declared void. We recommend the readers of this report to monitor the current legal situation before transferring EU citizens' data to the US.

4 AI application deployment models

4.1 Introduction

Developers of AI applications can choose from a variety of archetypes when deploying their apps. The biggest difference is in the way the AI model is used by the app. Some AI models are accessible for free while others can only be accessed via paid application programming interfaces (APIs). In the following, we will pay extra attention to applications using cloud-based services, as the transfer of data between different data processors brings additional risks to privacy. Cloud processing (or the use of second-party data centres in general) is also very common in today's IT systems.

The technologically simplest AI system is an application implementing a specific business logic on the basis of an existing AI API. One example of such a solution could be a chatbot using the OpenAI GPT API where the main value proposal is the user experience and prompts provided by the app. Thin applications of this type may be limited by the context learning capability of the model behind the API.

More complex and more expensive solutions use an existing model's API calls while managing the user's status and servicing their data which may be domain-specific. Solutions like this require database integration, user management and also input and output validation. The deployer of the app may thus use, e.g., some Retrieval-Augmented Generation (RAG) solution where the model's generic knowledge is augmented with information found in the app's own database. Solutions of this type are discussed in Section 4.4.2.

Some solutions involve the service provider deploying an AI model themselves. This presumes that the service provider either trains their model themselves, fine-tunes an existing model or adopts an external model while independently running inference (i.e., computing the AI's outputs on their own infrastructure). This requires investments into information infrastructure which grow with the size of the model and user base, but may at the same time reduce risks related to API availability, data confidentiality and privacy, as the number of data processors is reduced. In situations, where serving a large user base is not the goal, quantification and other optimisation methods allow running inference on many freely accessible models even on a powerful personal computer. Solutions of this type are discussed in Sections 4.4.3 and 4.4.4.

All deployment models covered here share some similar characteristics. For example, a service provider may use IaaS (infrastructure as a service), CaaS (computing as a service), and PaaS (platform as a service) services for business logic, model, and data management. In the context of the General Data Protection Regulation (GDPR), these service providers are considered processors of user data. In case user data are used not only for service provision but also for improving the quality of the model or other side tasks, a legal basis must be established (e.g., the user may have to give their informed consent) for such uses. This comes into play in the context of interfacing the service with other services and data.

4.2 Methodology

In the development of the deployment models discussed here we took into account the considerations and needs of potential service providers, as well as their everyday practices. We especially focused on statutory requirements and the movement of user data between different processors. The overview of deployment models presented below is not exhaustive, as there

are countless ways for connecting services, APIs, and data sources. It should, however, provide a sufficient picture of the critical points of more common approaches that are related to users' and service provider(s)' roles and responsibilities in the context of the structure of the deployment model and data flow. Simpler models also facilitate providing faster advice for carrying out risk analysis.

Arrows in the figures represent data flows, indicating the movement of data between different components of the deployment model. Representing data flows is vital because the movement of data across between areas of responsibility comes with risks (e.g., to privacy) which must be accounted for. Privacy and responsibility are understood here in the sense they are used in the GDPR. To facilitate better understanding of the boundaries of responsibility, as well as other characteristics of the deployment model tied to the structure of the specific AI supply chain, we have presented both services and critical data elements of the AI system (training data, model, input, output) as components of the deployment model. Our focus here is on AI-based cloud services, as due to their performance requirements, AI systems often need to use specialised hardware accelerators in cloud services for accelerating computations. It must be kept in mind, however, that AI systems not deployed via the cloud are somewhat less exposed to confidentiality risks; systems of this kind will be discussed separately. IaaS, CaaS and PaaS components are not specifically represented in the deployment figure diagrams, as they can easily be used with different elements of the deployment model. We will, however, discuss the consequences of their use.

We have used performance analysis to provide a more detailed picture of AI application deployment models. Models are presented using Business Process Modelling Notation (BPMN). This has allowed us to specify the data objects processed by the model, as well as the processing parties.

4.3 Legal roles of AI system stakeholders

From the perspective of both the GDPR and the AI Act, it is crucial to assess the applicability of the regulations. The applicability of GDPR rules must be considered if an AI system processes personal data anywhere in its life cycle. The applicability of AI Act rules must be considered if the person is an AI developer or if it uses an AI system or API developed by someone else in their services. An AI system in the sense of the AI Act is a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments [99].

If an AI system or a person operating the system is found to fall within the scope of the regulation(s), the specific requirements arising from the regulation(s) must be identified. From the perspective of the GDPR it is important to, e.g., determine whether the organisation qualifies as a controller or a processor of personal data; in the case of the AI Act, however, whether the organisation qualifies as a provider or deployer of the AI system. Both regulations also define several other roles, which are also advisable to review. The roles listed above are the most critical, though – especially those of the controller (GDPR) and the provider (AI Act), as both are subject to strict compliance rules. In some cases, a single person may also simultaneously act in several different roles depending on processes, relationships between the parties, or agreements. Identification of roles is crucial because of the dependence of responsibility on roles.

According to the GDPR, a controller is the natural or legal person, public authority, agency or

other body which, alone or jointly with others, determines the purposes and means of the processing of personal data (GDPR, Article 4(7)). A processor is a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller (GDPR, Article 4(8)).

A provider is a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge [99]. A deployer is a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity [99].

In order to identify which requirements apply in the specific case, it is also necessary to determine the objective of the data processing and AI use, the types of data processing processes operating in the system, the types of data being transferred and the parties of these transfers, and the AI system or component (including the risk level of the system) being used.

4.4 Deployment models

4.4.1 Overview of models

We have identified three distinct deployment models for AI applications differentiated by the transfer of data between parties, the deploying party, and the origin of the AI model. The relationships between these models, as well as illustrative applications, are presented in Figure 9.

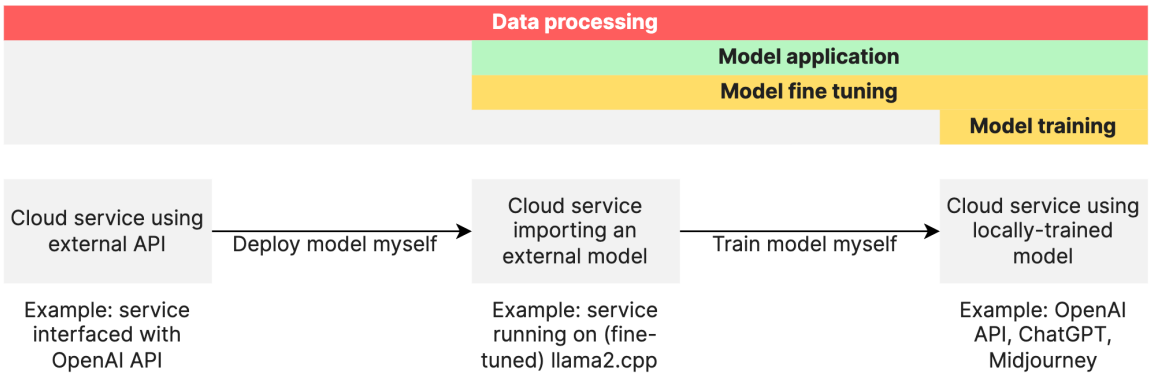


Figure 9. Deployment models from the perspective of the deployer's tasks in relation to the AI model

The models have been listed in the order of the extent to which the AI application service provider can rely on existing AI services and products. The more specific and complex the commercial purpose and the stricter the requirements for the processing of data, the bigger the proportion of necessary services that usually need to be developed in-house. This heuristic is just an approximation, however. From the perspective of data flow structure, the last of the listed deployment models includes both simple and complex solutions.

The upper part of the figure represents the scope of the deployer's tasks in different deployment models. In all cases, the deployer processes some kind of data. Beginning with cloud services importing an external model, the deployer deploys the model itself alongside their business logic, fine-tuning the model if necessary. In the case of a locally-trained model, no third party is any longer responsible for the creation and training of the model; both (as well as the management of training data) are completely in the hands of the deployer.

4.4.2 DM1: Service using an AI API

One common choice of architecture for AI-based services is using a third-party AI API in your business logic. If necessary, the service provider can also process or store user data, which the deployer can also implement using cloud services. The initial data used for training the third-party model can in turn come from external sources. Alternatively, the third-party AI cloud service or API can train its models on user data received from the service provider. All such cloud services can rely on some IaaS (infrastructure as a service) solution.

The model described above has been used in, e.g., machine vision applications. It gained in popularity after the publishing of the OpenAI API which facilitated simple interfacing of your service with powerful language and image models. The AI model is external to the application (i.e., outside the service provider's control). The training data for the model are also external in origin. User data flows to the service, from the service to the AI API provider, then back to the service, and finally back to the user. If the service is interfaced with third-party services and data then the user data may also be transferred there. User data can meanwhile be stored by both the service provider and the AI API provider (e.g., storing inputs and outputs in cache, but also in the training database). In one special case of this deployment model, the AI API provider also provides the option of fine-tuning the model on the service provider's data but the API provider still deploys the fine-tuned model. This approach partially overlaps with the next deployment model (see Section 4.4.3).

DM1: Service using an AI API

Overview: Service interfaces with an external API to process user data using the AI API provider's model. Both the service and the API provider can also share data with third parties for additional processing. The initial data used for training the model may come from third-party sources.

Examples: copy.ai, Streamlit and Gradio AI demo applications, services using the OpenAI API

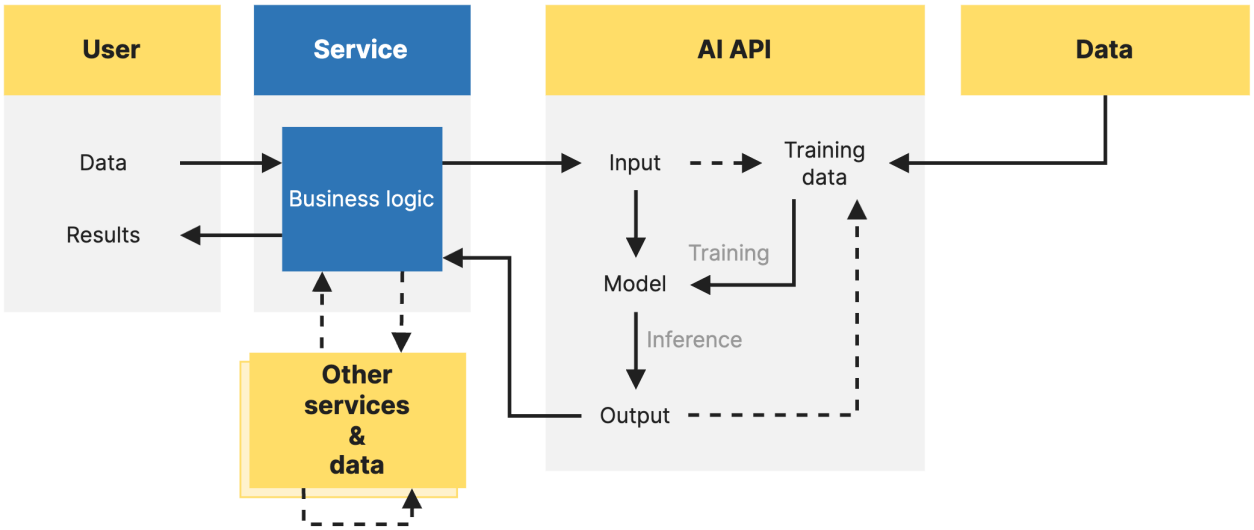
Origin of model: External

Origin of training data: External

Input data are stored: Optional

Input data transfers: To cloud service, then (if necessary) to other services and the API, back to the service, then back to the user, potentially using different infrastructures.

Figure:



Risks and considerations:

1. Processing of personal data by the service provider or API provider.
2. Information on the explainability of the model used by the API provider may be incomplete.
3. User queries and model outputs are validated by the service provider.
4. Non-service provider related failures in the work of the AI API are a risk to availability.
5. Lowest capital investments and technologically least complex of all deployment models.

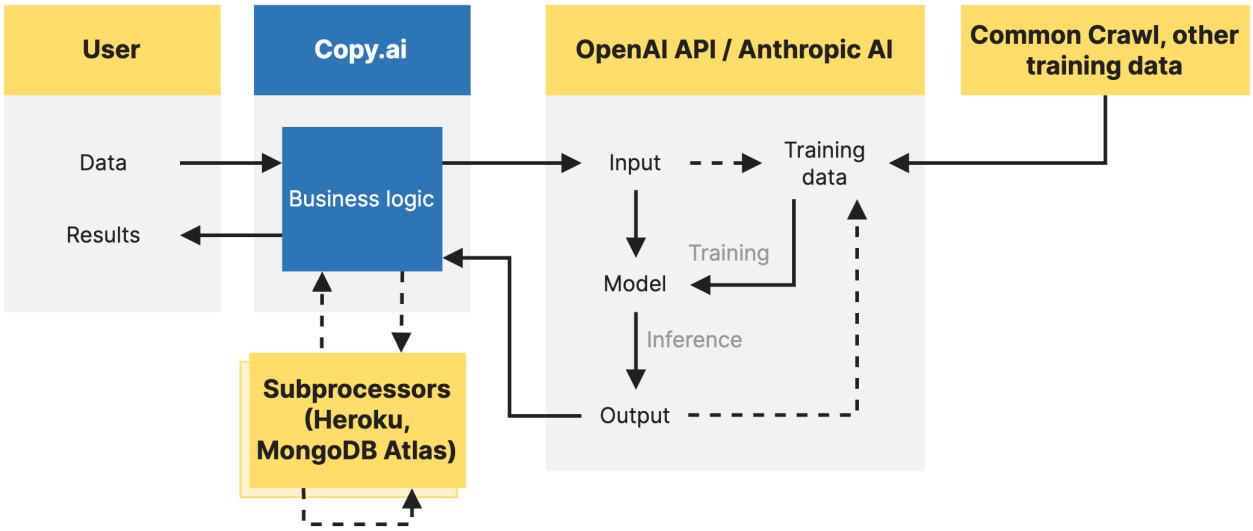


Figure 10. Copy.ai as an example of a service using AI via an API

Copy.ai is an example of a DM1-type service [1]. Copy.ai uses the OpenAI API to assist the user in writing marketing and advertising texts. The user provides the service with a description of the text required and its characteristics (e.g., writing style), the service processes the descriptions, and presents them to the AI API in the form of a query. The user can choose which API they wish to use (Anthropic or OpenAI). After receiving a response to the query the service further processes the response and returns the result to the user. The copy.ai deployment model is presented in Figure 10.

Figure 11 describes the data flows in a DM1-type deployment model. In the case of a DM1 model, the user uses a service which in turn uses an AI API to generate an output. The API provider is divided into two divisions with different tasks – model development and service deployment. The objective of model development is to design the model architecture, train and test the model and, if necessary, generate fine-tuning datasets and fine-tune the model. Model development is also responsible for monitoring the model.

The service deployment process begins with the user who uses their input data to generate data to be sent to the service. The service uses these to generate a query which is transmitted to the API. When the service sends a query to the API includes the query data in its input and the input in the model. The model will be used to generate an output and respond to the query. Depending on the terms of service, interaction history may be stored and used for both model monitoring and the generation of fine-tuning datasets.

Once the API has sent the service a response to the query (i.e., an output), the service will, in turn, generate a user output and transmit this to the user. The user can use the output received from the AI service for fulfilling their personal objectives.

4.4.3 DM2: Service implementing an external AI model

Interfacing with an external AI API or web service makes the deployer dependent on the accessibility of the service used. The deployer may also need to fine-tune the model which is not offered by all AI API providers. In order to solve these problems the deployer can adopt a pre-trained model from a model provider (or a freely provided AI model) and integrate this directly into their application. In case the deployer is fine-tuning the model – this is called transfer learning – they will face an additional need for managing training data and monitoring the model's security and quality indicators. This model is also applicable to the special cases where the model provider provides a federated learning service with centralised components.

[1] Copy.ai. https://www.copy.ai/ Last visited May 25th, 2024.

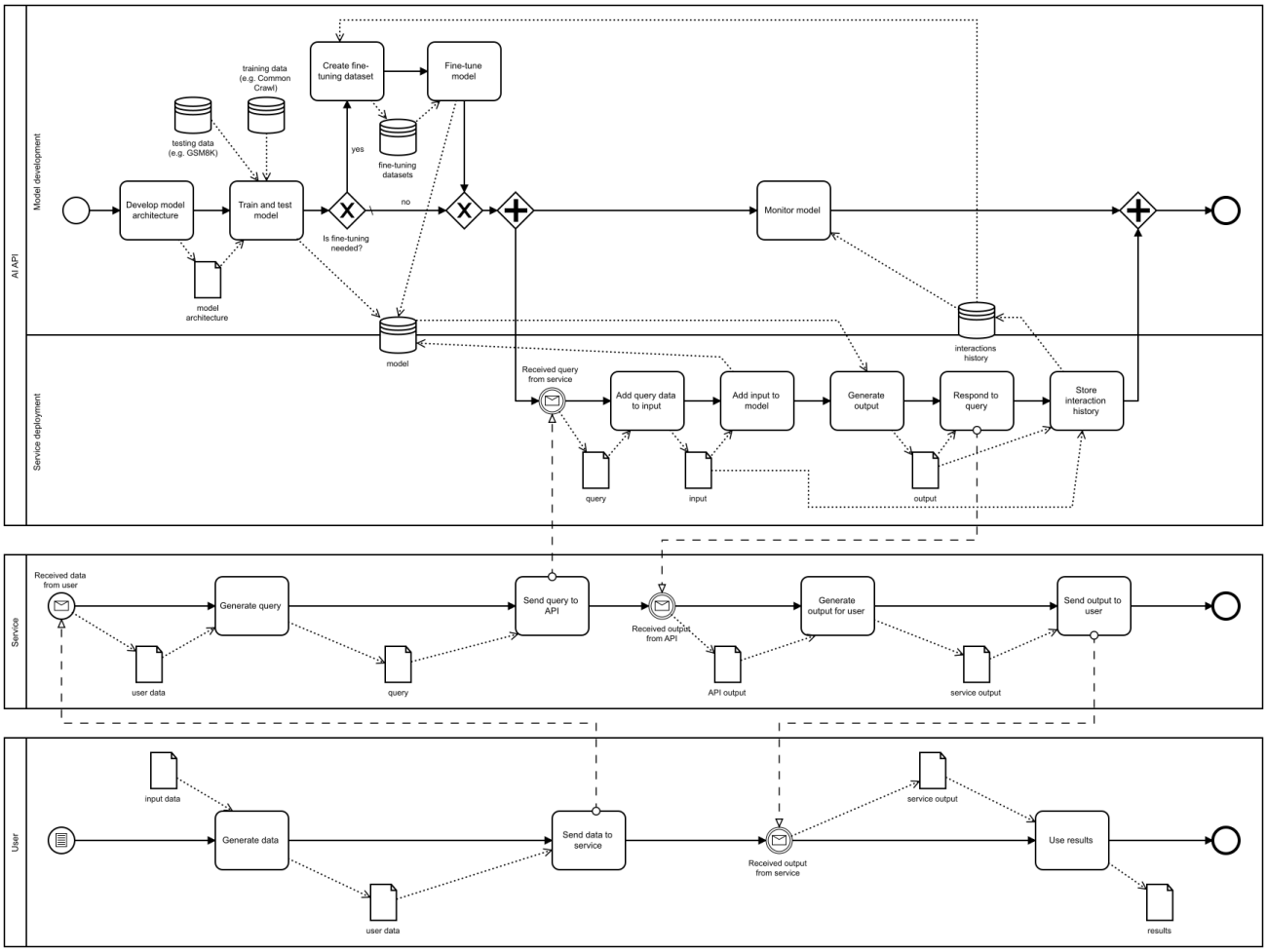
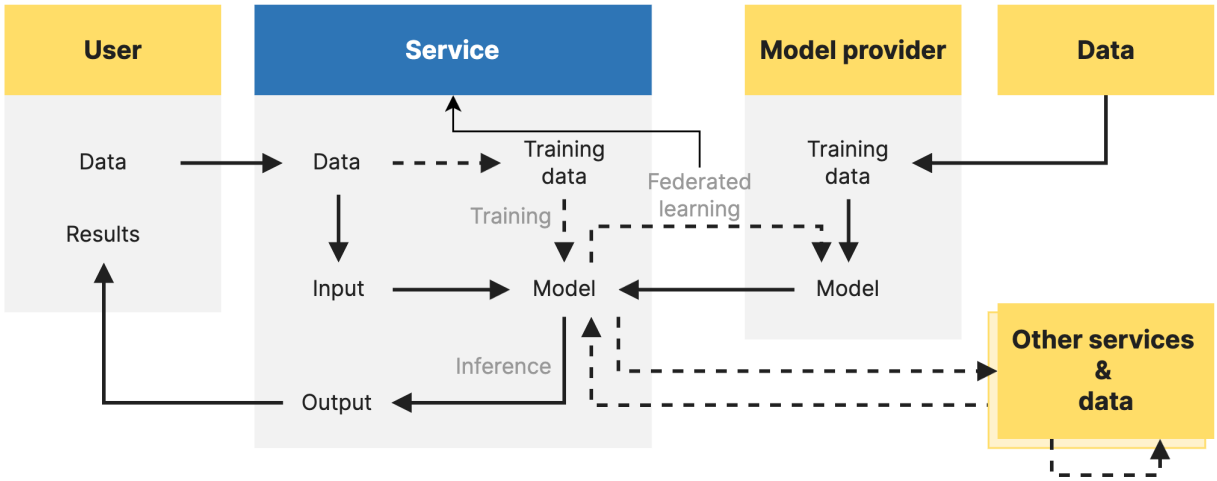


Figure 11. Data flows in deployment model DM1

DM2: Service using an external AI model

Overview: Service provider uses (and fine-tunes, if necessary) an externally imported model. The initial model comes from an external source; the creator of the model trains and transfers the model for their customers, including the service provider. The service provider deploys the model, using it on in-house and client data. They may use in-house data for fine-tuning the model. Cloud services can interface to other data and services, e.g., vector databases in the case of RAG solutions.	**Examples:** Service importing a model from, e.g., the Huggingface repository, Android Gboard (as an example of federated learning)
	Origin of model: External
	Origin of training data: External, in-house, user data
	Input data are stored: Optional
	Input data transfers: To cloud service, optionally to other services and back to the user, optionally using third-party infrastructure. In the case of federated learning, weight updates are also transferred to the model's trainer.

Figure:



Risks and considerations:

1. In fine-tuning, monitor security and quality indicators, as well as quality of in-house data and changes in their distribution.
2. Information must be collected on the security and explainability of a second party-trained model.
3. In some cases, weight updates may be considered personal data in federated learning.

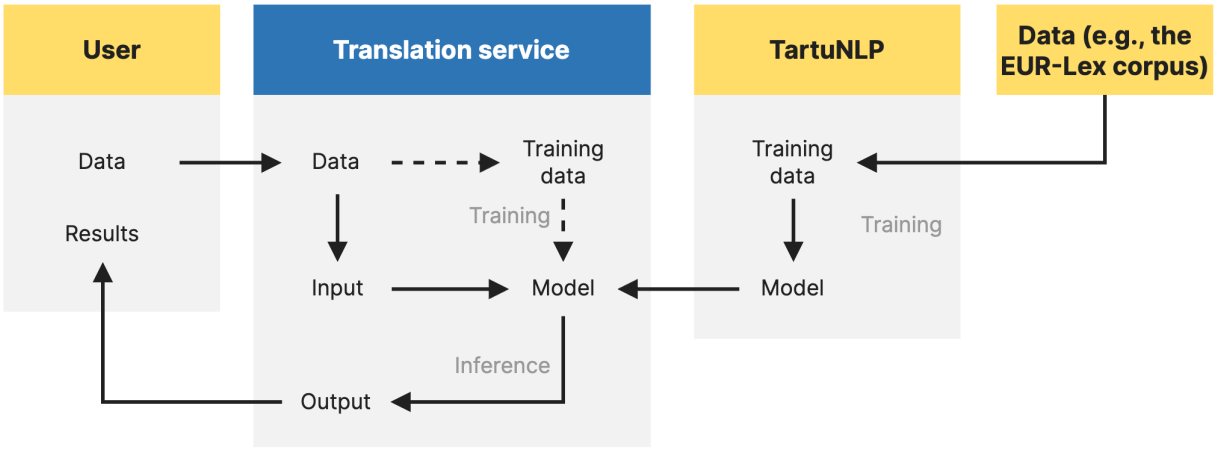


Figure 12. Translation service deployment model as an example of a service using an external AI model

Figure 12 represents an example of a translation service where the translation service deployed by the AI service provider uses a model pre-trained by a model provider (e.g., TartuNLP). The user sends a query to the service (e.g., through the application's web interface), the query is sent to the service where the data are processed (translated). The translated output is returned to the user. Data are not transferred from the service provider to the model provider. The service provider can further fine-tune the model based on user data.

Figure 13 describes the data flows in a DM2-type deployment model. The process involves three parties: user, AI service, and model provider. The model provider develops the model architecture, trains and/or fine-tunes and tests the model, and provides the model to AI-based service providers.

The AI services is divided into two divisions: further development of the model and deployment of the service. To further develop the model, the AI service integrates the model provided to them into their own service, generates a fine-tuning dataset if necessary, and fine-tunes the model. The service provider then continues to monitor the operation of the model. No data is transmitted back to the model provider from the AI service.

When the user creates and transmits data to the AI service, the AI service deployment branch adds the data to the input, then to the model to generate an output which it will then transmit to the user. The AI service then stores the interaction history which will be used for monitoring the model and can also be used for generating fine-tuning datasets. The user can use the output received from the AI service for fulfilling their personal objectives.

Figure 14 depicts a special case of the second deployment model. The difference between the two figures lies in a database query added to the service deployment stage, the result of which is added to the user input. This method is called RAG (Retrieval Augmented Generation); it can also be used with deployment models DM1 and DM3.

4.4.4 DM3: AI service using an in-house model

The third deployment model covers solutions where the AI model is trained and deployed in-house by the service provider. These include both simple solutions, such as decision trees and regression-based solutions, where the simplicity of the model makes it impractical to import from an external source, as well as solutions developed by large AI producers. Trainers of large AI models generally only offer services based on models they have developed and they possess sufficient resources for their autonomous deployment.

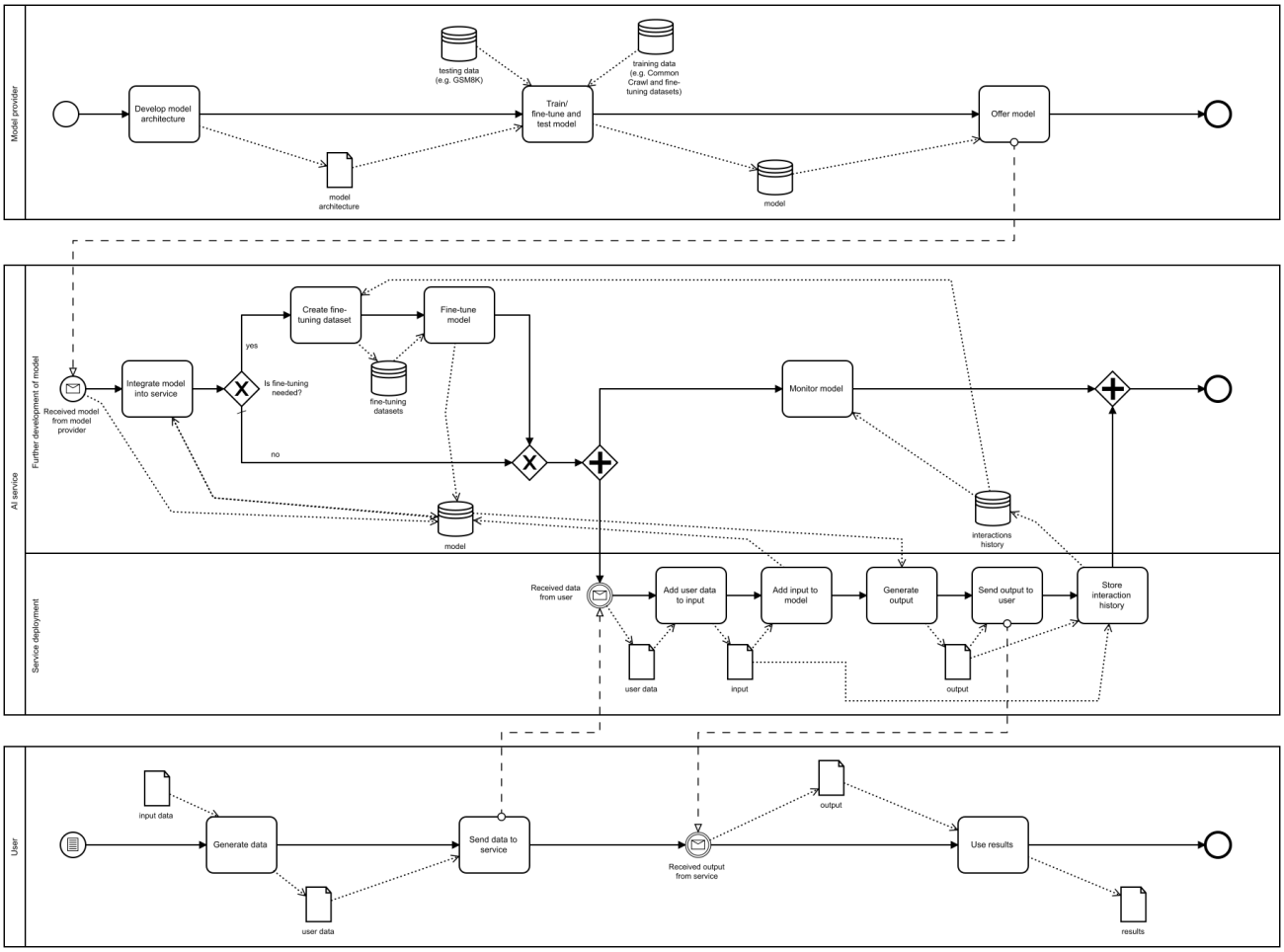


Figure 13. Data flows in deployment model DM2

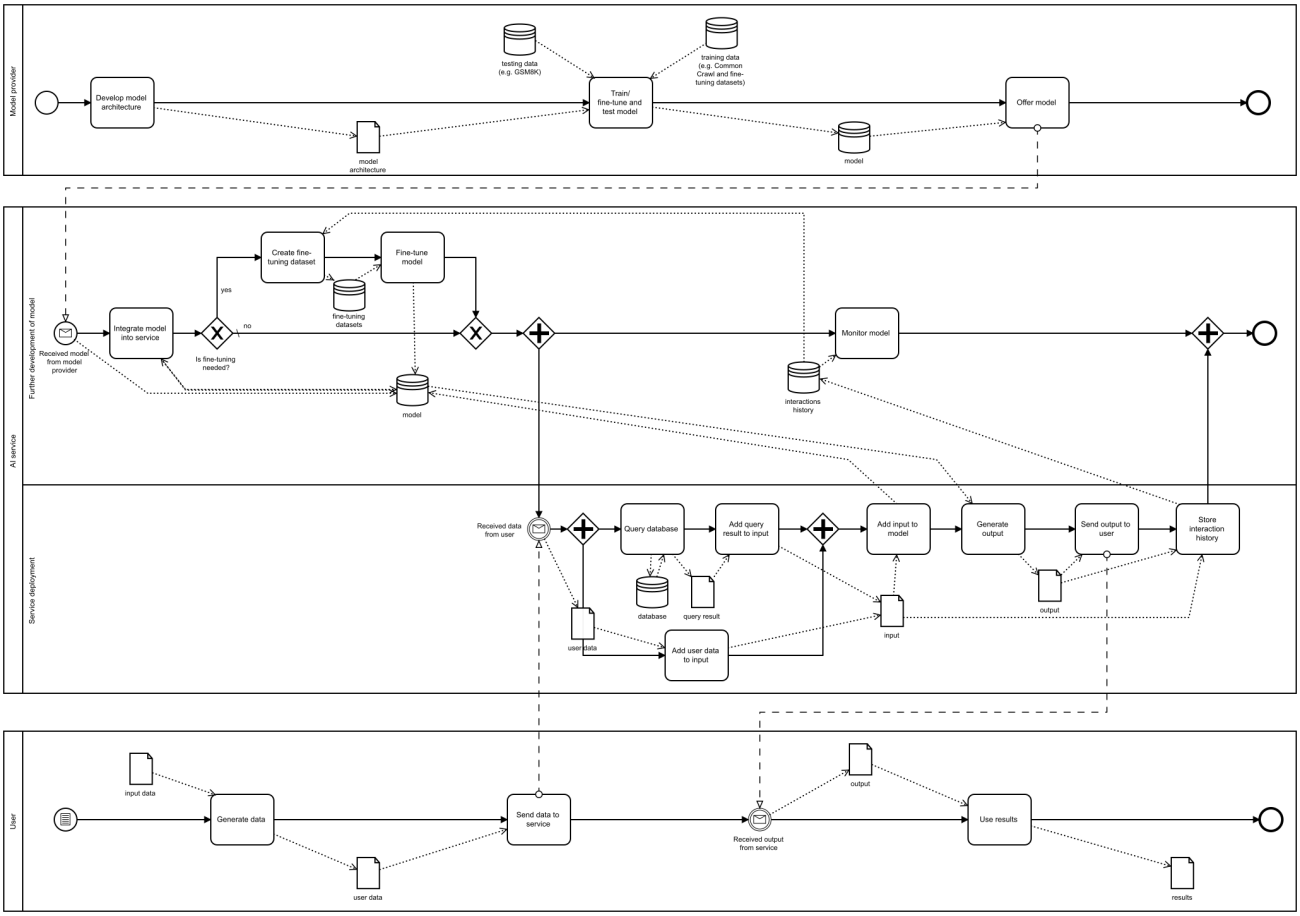
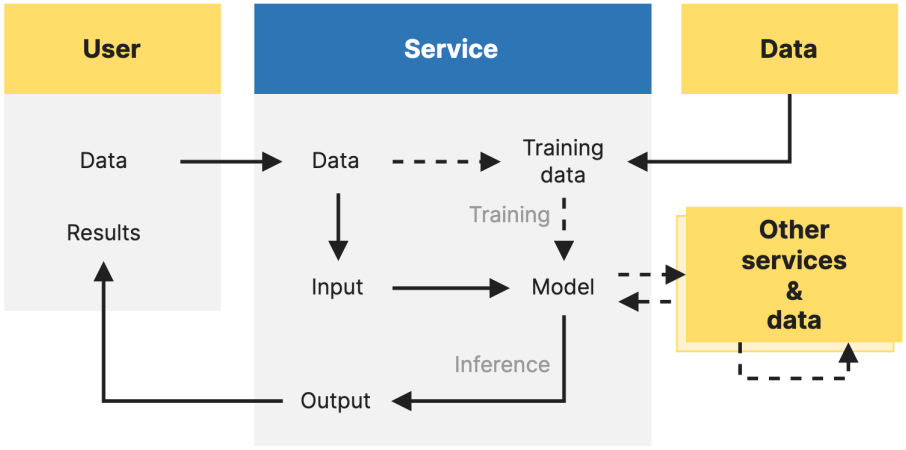


Figure 14. Data flows in deployment model DM2 implementing RAG

DM3: AI service using an in-house model

Overview: The trainer of the model collects data, trains, deploys (and, optionally, implements) the model. Using in-house models is an important use case. This can facilitate situations where neither the training data nor the model itself or user data are transferred to third parties.	**Examples:** Neural translation, ChatGPT and OpenAI API, Grok, DALL-E, Midjourney
	Origin of model: Internal
	Origin of training data: User, service provider, third parties
	Input data are stored: Optional
	Input data transfers: To cloud service, optionally to other services, then back to the user, optionally via an infrastructure service provider.

Figure:



Risks and considerations:

1. The trainer of the model is expected to provide information on the explainability and quality of the model.
2. The trainer must have a lawful basis for processing the training data.
3. In case of large models and large training datasets, solutions of this type are the most expensive to build.

xamples of this deployment model include all organisations using AI for building internal services, as well as e.g. OpenAI. The user sends queries to the service; the service, in turn, returns the output from the selected model. OpenAI's deployment methods for the models they train depend on the target group: some are deployed in the form of APIs, others – in the form of weights. OpenAI collects and purchases training data itself. At the same time, not all the details of the origin of the data are public. OpenAI API models are not trained (as of November 2023) on queries received over API; they are, however, trained on ChatGPT queries, except for ChatGPT Enterprise[2].

ChatGPT itself is also an example of this deployment model, as it uses in-house models (developed by OpenAI). One noteworthy thing about ChatGPT is the fact that, if the user employs plugins, these can make queries to third parties for additional processing or acquisition of data. It is important to keep in mind that the model does not communicate directly with the plugins or the services they interface to: this data exchange takes place as a part of the service's business logic. As a rule, this means that if the model decides to use a plugin it will use the information contained in the pre-prompt and the user's request to compose a query to the service interfaced via the plugin. A response based on the query composed by the model is returned to the model where it is formatted into a response utilisable by the user. The ChatGPT deployment model is shown in Figure 15.

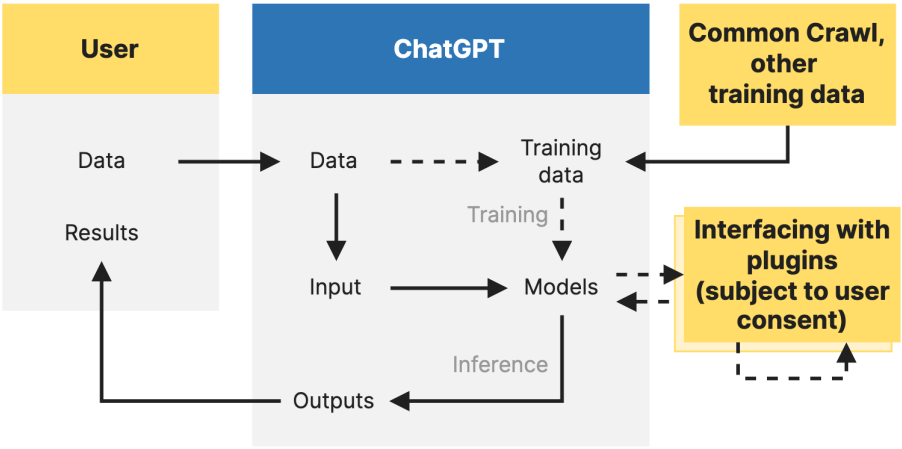


Figure 15. ChatGPT deployment model

Another version of an application compatible with this deployment model is a computationally inexpensive rules-based or other simple machine learning algorithm (e.g., linear regression, decision tree, or naive Bayesian classifier) easily trained on in-house datasets. This deployment model is suitable for, e.g., bank credit risk models: the bank trains the model in-house on its own (clients') data and implements the model in-house. The bank also uses supplementary data: credit default data, financial indicators, and internal bank data. The deployment model for a service of this type is shown in Figure 16.

Figure 17 describes a third type of deployment models. This model involves two parties: AI service and user. In this deployment model, the AI service, the deployer, and the model provider are all the same party.

The AI service is divided into two: model development and service deployment. Model development involves the same steps as the deployment models discussed above: model architecture development, training and testing the model, and optionally fine-tuning and monitoring the model. After receiving data from the user, the AI service development division adds the data to the input and to the model, composes an output, and sends the output to the user. Interaction history is stored and can be used for monitoring the model and assembling fine-tuning datasets.

[2]Enterprise privacy at OpenAI. https://openai.com/enterprise-privacy Visited December 1st, 2023

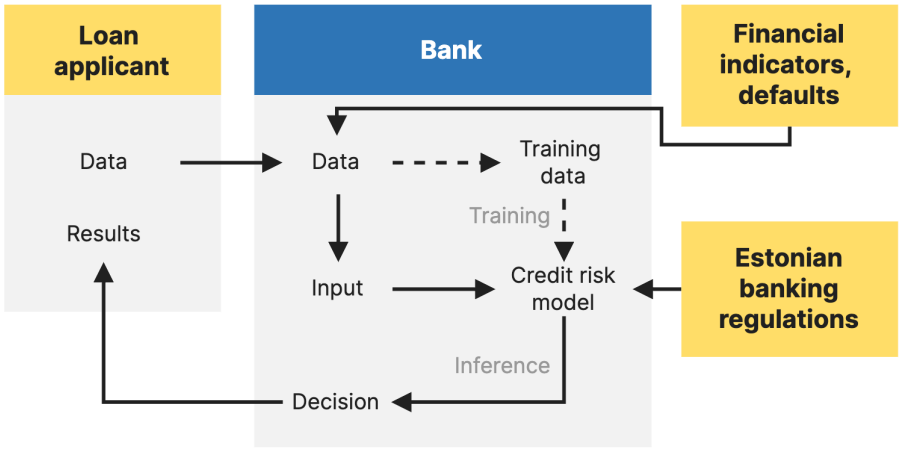


Figure 16. Deployment model for a credit institution's retail credit risk evaluation model

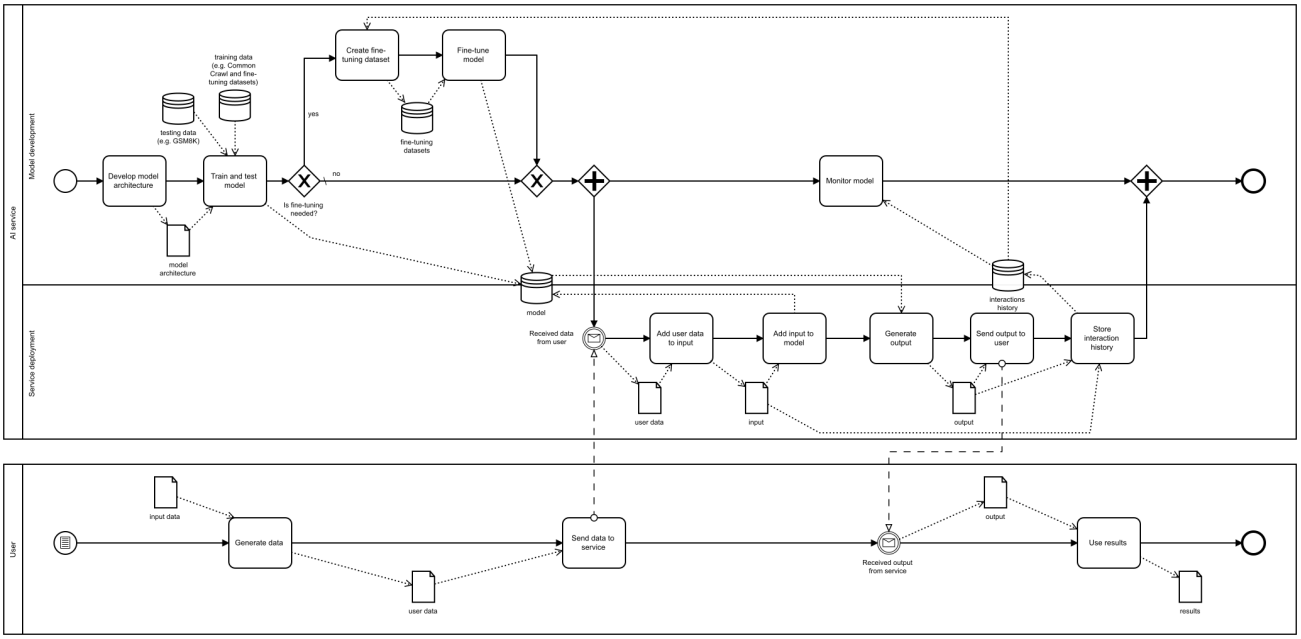


Figure 17. Data flows in deployment model DM3

5 Risks of AI applications

5.1 Risk management methodology

The main standards on risk assessment are the ISO 31000 risk management standard [142] and the NIST SP 800-37 risk management framework (RMF) [143]. The characteristics of information security risks are covered by ISO/IEC 27005 [144], those of cybersecurity by the NIST cybersecurity framework (CSF) [145]. AI-specific risk management guidelines are covered by ISO/IEC 23984 [146]; this standard describes how to adapt an ISO 31000 compliant risk management process to an organisation using, developing, or implementing artificial intelligence systems. If the organisation has an ISO/IEC 27001 certificate for its information security management system (ISMS), then our recommendation is to add AI systems to the existing risk management process.

The simplified methodology described here is compliant with ISO 31000 and ISO/IEC 27005 workflows, but can also be adapted to the NIST RMF and CSF frameworks, if needed. If the organisation wishes to employ a more complex risk management strategy any time in the future, it will be easy to integrate existing AI system risk management into the general framework. The Estonian information security standard (E-ITS) [147] is aligned with the ISO/IEC 27001 series, meaning that those implementing E-ITS can also opt for a risk-based approach. Thus, the methodology described here is also adaptable by organisations implementing E-ITS.

The risk management process comprises three steps: context establishment, risk assessment, and risk treatment. The scope of the risk management methodology presented in our report covers IT systems that include an AI component.

5.1.1 AI-specific considerations in context establishment

Context establishment involves identifying and documenting stakeholders and assets related to the process. The organisation defines its risk readiness, risk appetite and risk owners, and identifies the internal, national, and statutory requirements for stakeholders. The organisation determines the conditions for risk acceptance and selects an appropriate risk management methodology.

Context establishment for artificial intelligence systems requires identifying and documenting all stakeholders. This includes taking into account any parties that may not seem to be directly connected to service provision (e.g., persons appearing in training data, owners of works, as well as third-party infrastructure and service providers). Regardless of whether the organisation creates an in-house AI system and uses this system within the organisation or the AI system is used as a service, the analysis must include:

- data subjects or data owners whose data have been used in training the machine learning model;
- the party that trained the model;
- the service provider;
- the service user.

The organisation must identify the stakeholders and account for their rights and interest in risk assessment and risk treatment. New stakeholders may also necessitate the need to take into account new foundation documents or regulations. It is important to determine whether these new stakeholders are a part of, or external to, the organisation. The organisation must map everyone's statutory rights and obligations and who operates which part of the system.

The organisation must identify the origin of different types of data (models, training, input, and output data) and software components, as well as the data flow between the different components. Stakeholder and component mapping is necessary for understanding the context of the AI system. Some risks may also arise from the use of certain types of data or systems. For visualising the mapping, tools normally used for systems modelling (UML, BPMN) can be used. Specific tools exist (e.g., PE-BPMN [148]) for

describing the movement and visibility of data objects from the perspectives of different stakeholders.

Stakeholders' access to data can be documented using visibility tables. Table 1 is an example of a visibility table describing the access of different stakeholders to different types of data in an AI system. In this example, there are three stakeholders: the end user, the service provider (AI client application), and the AI API provider (who trains and shares the model). All stakeholders see end user input data and model outputs. The service provider and AI API provider have access to the service provider's business data. The model, in this case, is visible only to the AI API provider.

Table 1. Simplified example of a visibility table

	User input	Service provider's business data	Model	Output
End user	X			X
Provider of the AI-based service	X	X		X
Provider of the AI API	X	X	X	X

5.1.2 AI system risk assessment

Risk is often expressed as a combination of the likelihood of the occurrence of a threat event and its potential damage. Risk assessment involves the identification, analysis, and evaluation of risks. Risk identification, in turn, involves searching for risks, determining the relevance of risks, and description of relevant risks. A risk owner is assigned to each identified risk. Risk analysis covers determining the reasons and sources of risks and evaluating the potential data and the likelihood of the occurrence of the risk. In the course of risk evaluation, the risk level determined as a result of the analysis is compared to the criteria for acceptable risks defined in the course of context establishment in order to evaluate whether the risk level is tolerable and acceptable.

AI risk assessment is based on the established context. For each component of the AI system, risks are evaluated in the context of the stakeholders. Finding these relationships is easy based on the visibility table created during context establishment. For each identified stakeholder and component pair, we analyse and evaluate three types of risks: risks related to cybersecurity, regulations, and AI-specific threats. Cybersecurity risks are usually connected to the adequacy of the AI system's processes or the confidentiality, integrity, and availability of the AI system's components (software, data, services). Risks related to regulations are connected to the legal obligations for stakeholders operating AI systems (AI-specific regulations) or system components (e.g., regulations on personal data, copyrighted data, critical infrastructure). AI risks are connected to the characteristics of AI algorithms, as well as the impact of AI systems on the society and ethical aspects. AI system risk assessment is covered in more detail in Section 5.2.

Table 2 provides an example of defining risks via security vulnerabilities and threats. For each threat, the organisation must evaluate the likelihood of the threat materialising and the potential damage. The likelihood of and damage caused by a similar event can be different for different organisations. In some cases, it will be beneficial to compare the risks of different solutions in order to choose the most suitable solution for the organisation. For instance, even though a cloud service provider may offer better security measures than a small organisation could implement itself, dependence on a cloud service may be an availability risk, should the connection to the cloud provider be lost.

5.1.3 AI system risk treatment

Different solutions exist for risk treatment: risk avoidance, risk mitigation, risk transfer, or risk retention. The solution to be used will be chosen based on risk analysis results.

Table 2. Examples of security vulnerabilities and threats

Data	Risk type	Security vulnerability	Threat
Output	AI-specific risk	Biased or defective model	End user receives an output guiding them to harm themselves or others
Training data	Regulatory risk	Lack of legal basis for processing personal data	Fine for data protection regulation violation
Model	Information security risk	Defective identity management	AI API provider loses access to their infrastructure and is unable to provide inference service

The organisation is unlikely to be able to mitigate all risks. Risks may be ordered according to their importance. Suitable information security measures, AI-specific measures, or legal measures are chosen to facilitate aligning potential risks with the organisation's risk appetite.

Risks can be avoided through the elimination of the source of the risk, discarding functionalities, or re-organisation of the business process. Risks are mitigated through the adoption of security measures. The employment of additional security measures to mitigate a risk is not always possible or rational. Risk mitigating measures are described in Section 6. Risk transfer means sharing the risk with another organisation, or the compensation of damage arising from the risk, e.g., by using insurance.

If the risk level remaining after risk treatment corresponds to the organisation's risk appetite, the rest of the risks can be accepted. This means that the risk in question will no longer be worked on and the risk is retained. Periodic survey and revision of risks is required to keep risk management up to date, as threat occurrences of impacts evolve. Another important element of the process is risk communication, the objective of which is to keep the employees informed of the process and results of information security risk management.

5.2 Risk assessment

5.2.1 Information security risks

Digital risks are the most likely and have the biggest impact [149]. The main threat here is cyber-crime [149, 150]. Generative AI technologies can, however, support more efficient handling of digital risks [149] when developed and implemented for this purpose. Research and development related to the creation of automated or semi-automated cybersecurity measures is also recommended by the NIS2 directive [21].

Information security risks are identified and analysed on the basis of threats, probability of threat events, and potential damage. The Estonian E-ITS information security standard describes a baseline security process, one element of which is the baseline security catalogue. The catalogue consists of process modules and system modules. These, in turn, contain a list of threats and a description of measures. Using the baseline security process will simplify risk identification and is also compliant (when implemented at a high level) with the ISO/IEC 27000 series of standards.

The following E-ITS baseline security modules [151] are relevant to the implementation and use of AI systems: process modules ORP (organisation and personnel), CON (concepts and methodologies), OPS (operations), DER (detection and reaction), and system modules SYS (IT systems) and APP (applications). The modules listed above only include those which require taking separate measures related to the implementation or use of AI systems. The list does not include modules necessary for setting up the rest

of the organisation's infrastructure or security management. If the organisation does not assess or treat risks in surrounding systems then even strong levels of protection for the AI system will be meaningless.

The addition of AI systems to the organisation's workflow will probably give rise to the following process threats. Note that the list of threats is not limited to those listed in the standard.

- ORP 1. No clear rules for the use of AI systems exist; the AI system is incompatible with other tools.
- ORP 2. The employees are insufficiently familiar with AI systems; they are careless about using data in AI systems; they are insufficiently qualified.
- ORP 3. The employees have not received sufficient training on threats and attacks related to AI systems.
- ORP 5. Use of the AI system is in violation of the law or contractual obligations; unauthorised publication of information in the AI system; internal information is accidentally revealed to an external AI system.
- CON 2. Inputs to AI systems are provided in neglect of data protection requirements; data processing procedures are inadequate and do not account for the working principles of AI systems; no resources are allocated to the protection of personal data in AI systems; the privacy of data subjects is not ensured for data processed by AI systems; the confidentiality of data in the AI system is not ensured, as data can fall in the hands of unauthorised persons or are accessible in the trained model; the reputation of the data processor is damaged.
- CON 3. Problems related to backing up AI system data (both the inputs and model, as well as, in some cases, the outputs).
- CON 6. Inadequate deletion and destruction of AI system data.
- CON 8. Unsuitable development methods used for AI system development; insufficient quality management; inadequate documentation; insufficient development environment security; AI system design errors; inadequate AI system testing and acceptance procedures; using production environment data for testing the AI system.
- CON 10. In case of the AI system used as a web app: displaying sensitive background information found in the AI system in the web app; use of automated attacks for attacking the AI system web app.
- OPS 2.2. All threats related to the use of cloud services apply: inadequate AI cloud service use strategy; dependence on AI cloud service provider; insufficient requirement management in using AI cloud services; violation of statutory requirements; deficiencies in agreement signed with the AI cloud service provider; insufficient integration of AI cloud services with in-house IT systems; insufficient regulation of the end of AI cloud service use; deficiencies in emergency readiness plan; AI cloud provider system failure.
- OPS 2.3. All threats related to outsourcing apply: inadequate AI system outsourcing strategy; insufficient control over business critical processes; dependence on AI service provider; insufficient level of information security at the AI service provider; insufficient control over the provided AI service; deficiencies in agreements regulating the AI service; inadequate access rights management; lack of control over AI service provider's subcontracting; lack of key performance indicators (KPI); inadequate stipulations regarding the end of AI system outsourcing; inadequate emergency management in outsourced AI service.
- OPS 3.2. All service provider information security threats apply: inadequate information security management by the AI service provider; inadequate emergency management by the AI service provider; inadequate service agreements with AI service receivers; vulnerabilities in interfacing with AI service provider's IT systems; dependence of AI service receiver on service provider; inadequate management of access rights; lack of multi-tenancy capacity at the AI service provider; AI service provider's dependence on subcontractors; inadequate procedure for ending AI service agreement; AI system provider IT system failure; social engineering.
- DER 2.1. Inadequate handling of security incidents related to AI systems; destruction of evidence in security incident handling.

- DER 3.1. Inadequate or unplanned implementation of security measures in AI systems; verifier's inadequate qualification; inadequate audit planning and coordination; non-coordinated use of personal data; intentional hiding of security issues.

The system module SYS describes threats to IT systems, including servers (SYS 1.1, 1.2, 1.3, 1.9), virtualisation systems (SYS 1.5), containers (SYS 1.6), storage solutions (SYS 1.8), client computers, (SYS 2.1, 2.2, 2.3, 2.4), laptop computers (SYS 3.1), smartphones and tablets (SYS 3.2), printers (SYS 4.1), embedded systems (SYS 4.3), IoT devices (SYS 4.4), and external storage devices (SYS 4.5). The SYS module also describes threats related to the use of the Estonian X-Road security server (SYS.EE 1) and eID components (SYS.EE 2). Depending on the AI system or service being created or used, the relevant threats can be found in the relevant modules.

The system module APP describes threats to applications: mobile applications (APP 1.4), web applications (APP 3.1), database systems (APP 4.3), Kubernetes clusters (APP 4.4), software in general (APP 6), and custom software development (APP 7). APP.EE 1 additionally describes threats to the Estonian X-Road data services.

The AI system developer or implementer can use context establishment to determine which of these threats are relevant to them. Identification of threats enables the description, analysis, and evaluation of risks.

5.2.2 Legal risks

Notable legal risks related to AI systems include non-compliance with statutory requirements which may lead to:

1. damage claims;
2. legal disputes;
3. sanctions from competent supervisory authorities, including notices to ensure compliance, imposition of penalty payments, suspension or cessation of operations.

The listed risks may lead to additional time spent by employees on working on the damage claims or legal disputes, costs related to external legal services, financial loss from compliance with damage claim or court ruling or compensation of legal expenses, loss of income from suspension of operations, or reputational damage. The latter may materialise in the form of loss of clients and reduced income or, in the worst case, loss of trust and cessation of operations.

Fines related to processing personal data can reach up to 20 million euros or, in the case of enterprises, up to 4% of global turnover from previous financial year, whichever is greater. According to the AI Act proposal, certain violations would be liable to fines up to 35 million euros or, in the case of enterprises, up to 7% of global turnover from previous financial year, whichever is greater. The submission of inaccurate, incomplete, or misleading data would be liable to a fine of either up to 7.5 million euros or, in the case of enterprises, up to 1% of global turnover from previous financial year, whichever is greater.

According to the AI Act proposal, the European Commission may impose fines on generative AI system service providers for non-compliance of up to 35 million euros or, in the case of enterprises, up to 3% of global turnover from previous financial year, whichever is greater. The AI Act proposal also foresees the right for competent authorities to remove an AI system from the market.

It is also crucial to ensure that the AI stakeholders have written agreements in place listing the rights, obligations, and responsibilities of the parties. Data processing agreements between the parties also play an important role in the processing of personal data. Non-compliance with an agreement can also result in penalty and damage claims, as well as legal disputes.

In the past few years there have been numerous court cases involving disputes over inputs (texts, photos, etc.) used for training AI systems (see, e.g., [152, 153, 154]). These have predominantly concerned copyright violations. At the same time, there have also been disputes over responsibilities related to AI systems. Thus, in Moffatt vs Air Canada [155], the court found that an enterprise is responsible for all information found on their website, regardless of whether the information comes from a static page or

a chatbot. Court cases test the legal boundaries of AI and will hopefully bring clarity to this area in the very near future, helping to create more uniform practices for the interpretation of legal norms.

5.2.3 AI risks

Developments in AI, especially in large language and image synthesis models, have stimulated discussions of risks of these technologies. The risks themselves can be connected to both the harmful or unintended outputs of (universally) powerful models, as well as the spread and increased adoption of these models and the societal consequences of their adoption.

The most powerful image and language models are expensive to train; widely used smaller open-source models are, however, not far behind in their capabilities and can be expected to grow even more powerful in the near future. The adoption of AI models for automated decision-making in critical areas, such as medicine or warfare, has given rise to additional risks and numerous ethical concerns.

Risks related to artificial superintelligence capable of independent action and human ability to control and guide its actions call for separate consideration. The further development of artificial intelligence may give rise to new, previously unknown risks, as well as compounding existing ones, meaning that their mitigation has to be a continuous, iterative process.

5.2.3.1 Classification of risks based on the proposed AI Act

The AI Act is based on a risk-based approach, distinguishing between four levels of risk: unacceptable, high, limited, and minimal (see Table 3). Requirements for AI systems are based on the risk level. For general purpose AI (GPAI), the regulation also defines two other risk classes: non-systemic and systemic risks.

Table 3. AI Act risk levels for AI systems

No.	Risk	Description	Examples of AI systems
1	Unacceptable risk	Prohibited AI systems	AI systems causing significant risks to human health and safety or fundamental rights (manipulative, exploitative AI systems), e.g., social scoring systems
2	High risk	Regulated high-risk AI systems	E.g. biometric identification systems, emotion identification systems, security components of critical infrastructure systems, recruitment systems, polygraphs, interpretation of law in courts
3	Limited risk	Compliance requirements	AI system has no significant impact on nature or outcome of decisions. AI system is designed for performing limited procedural tasks, structured data creation, grouping incoming documents by subject or detection of duplicates among large numbers of applications
4	Minimal risk	No obligations	AI systems that can be used without limitations, e.g., spam filters, AI based video or audio enhancement systems

5.2.3.2 Algorithmic risks

The following section focuses on risks related to specific AI systems and the immediate consequences of their use. In some cases, the materialisation of these risks is connected to attacks against AI systems,

which are discussed in Section 5.3.

Limited generalisation ability. The utilisation of automated artificial intelligence systems in highly critical fields (e.g., medicine, warfare, or self-driving vehicles) comes with the risk of the model not returning a viable output for an input deviating too far from the training data. Large language models have been observed to produce 'hallucinations' where the model returns a superficially convincing but factually unfounded result [156]. This risk is compounded by AI systems' lack of transparency which may lead to the danger of blindly trusting a harmful or misleading output.

Excessive dependence on AI and loss of human supervision. The increasing adoption of artificial intelligence, including in critical systems, threatens to leave humans in the passenger seat. The more complex AI models and systems become, the more difficult they are for a human to grasp which may decrease human ability to monitor these systems. The reduction of human supervision, in turn, reduces our ability to interfere in the operation of AI systems and prevent undesirable outcomes. At the same time, the benefits provided by these systems may be large enough that the price of loss of supervision and controllability will be deemed acceptable. Given that complex systems tend to be more fragile than simple ones, excessive dependence without understanding can be a big risk.

Biased and dangerous responses. Even systems tuned to be safer using reinforcement learning can be made to generate discriminatory, abusive, or otherwise potentially harmful content using prompt injection techniques [157]. In addition to the risk of prompt injection, the model's security mechanisms can be disabled quite cheaply or even inadvertently via fine-tuning [158] while some models (especially open-source ones) do not even contain any meaningful protections of this kind. Since the models are primarily trained on human-made datasets containing biases characteristic to humans, the models trained on these datasets are also inherently biased. At the same time, corrections for algorithmic discrimination require care in the selection of target indicators, for the latter may also be biased. Excessive use of corrective measures can have an outsize negative impact on the capabilities of the model or application, which is exactly what happened to, e.g., the Google Gemini AI image synthesis tool [1].

5.2.3.3 Societal risks.

AI technology is in rapid development. The widespread adoption of AI promises to bring enormous economic and social benefits, yet it also threatens to lead to an upheaval at least as sweeping as the one caused by the widespread adoption of the Internet. The (human) societal risks of AI are connected to both the expansion of human agency mediated by AI and the unpredictability of the accompanying social changes as the possibility of the emergence of artificial superintelligence (ASI).

Autonomous artificial superintelligence. Artificial superintelligence, familiar to many from science fiction, has recently found itself in the limelight of discussions over the existential risks of AI. Increases in AI model size and computing power, as well as the appearance of emergent properties, give rise to certain expectations for even more powerful and multi-modal models or applications possessing a superior generalisation ability. Should such a model possesses a sufficient level of autonomy, access to critical (e.g., financial) systems, and the ability to remain undetected or avoid potential countermeasures, the risk would be even greater [5].

A sufficiently powerful and autonomous AI agent can (whether with human assistance or without it) become a threat just by gaining access to the Internet and the ability to make GET queries, using security holes, such as Log4Shell [159]. Additional risk factors include the agent's ability for self-enhancement and situational awareness. Researchers have not reached a consensus over the potential timeline of the emergence of such abilities but this in itself does not rule out either the possibility of relevant risks or even a potential existential risk to the humanity.

Uncontrolled spread of AI models. The free distribution and widespread adoption of AI foundation models, which by now seems avoidable, will magnify AI-related risks. The more users and developers can access the model, the higher the number of potential exploiters and the greater the scope of required regulation [160]. This risk is even bigger in the case of the spread of pre-trained models which have not

[1] Google's 'Woke' Image Generator Shows the Limitations of AI https://www.wired.com/story/google-gemini-woke-ai-image-generation/ Visited February 23rd, 2024

been fine-tuned to provide safe answers.

Biological and chemical weapons. In the context of the spread of powerful foundation models, researchers have highlighted the risk of terrorist groups gaining access to a tool that can help them acquire chemical or biological weapons more easily [161]. Autonomous AI agents capable of performing the necessary research autonomously deserve special attention here [162]. Some researchers have pointed out that, in the context of the evaluation of AI-specific risks of the spread of biological and chemical weapons, AI should be equated to access to the Internet which the malicious parties certainly have, and the real question is, which[2] bottlenecks[3] in their manufacturing process are eliminated by AI [163].

The availability of information is generally not one of such bottlenecks – in any case, LLMs can be considered a compressed (with losses) version of information already found on the Internet –, unlike, e.g., navigating the mountains of information or the manufacturing process. The ability of modern AI to respond to questions based on its extensive general knowledge and navigate in and summarise textual data can accelerate this process. Even though the sceptics have pointed out that the capacity to produce chemical or biological weapons is currently rather rare, the expected improvements in the performance of AI models and applications is threatening to magnify such risks.

AI in information warfare. High-quality text, image, speech, and video synthesis models enable carrying out extensive automated disinformation campaigns which, in turn, is portending distrust of web content in general. This is a problem to states and authorities [164] which will now have to seek for ways to affirm the authenticity of their messages. AI provides all parties in information wars with powerful weapons; meanwhile, defensive measures have not developing at a similar rate.

Artificial intelligence and fraud. The spread of generative artificial intelligence has also provided new tools to scammers [165]. Image and text synthesis models allow generating credible fake identities, including passports and other identity documents. Speech synthesis enables the imitation of another person's voice which facilitates identity theft. Language models have automated the creation of ever more believable, customised phishing e-mails. Videos created using deep fake technologies can cause significant harm to the persons they depict.

5.2.3.4 Ethical dilemmas.

The adoption of AI gives rise to numerous ethical issues. Can AI be the final arbiter in matters of life and death? If the economic transformation caused by AI is too sudden, should it be slowed down? Can an AI system or model be considered the author of a work? Who is responsible to problems with the AI system or the damage it has caused?

Loss of jobs. Large language and image synthesis models threaten to replace humans in numerous fields. The general abilities of modern language models are no worse than humans' in tasks demanding communicating with the client in natural language following predetermined algorithmic rules or the composition and summarising of marketing and other specialty texts based on existing sources of information. Speech synthesis is endangering call centres, image synthesis – art directors and artists, text synthesis – marketing copywriters and technical support specialists. The more powerful AI solutions become, the greater their impact on the labour market; widespread loss of jobs will come with economic and social risks. This process can be considered part of a broader trend of automation, hitherto mainly connected to the evolution of robotics where the ethical dilemma concerns the trade-offs between productivity and security of the employment relationship.

Ethical dilemmas in autonomous systems. These days, AI is used in systems making autonomous decisions that can have a significant impact on human autonomy. The adoption of such systems requires consideration of the ethical and moral aspects of decisions made by the AI. If a fast-moving self-driving car finds itself about to run over a baby and a grandmother, the AI system is forced to make a moral call on who it should put at risk: the baby, the grandmother, or the driver? The problem of deciding over human

[2] Anthropic: Frontier Threats Red Teaming for AI Safety: https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety Visited November 9th, 2023

[3] Propaganda or Science: Open Source AI and Bioterrorism Risk: PropagandaorScience: OpenSourceAIandBioterrorismRisk Visited November 9th, 2023

lives is encountered in all systems where humans have no way to control and promptly interfere in the decision-making process. The ethical risk is especially great in the case of fully autonomous weapons systems, such as turrets or drone swarms, which have to make friend-or-foe decisions in a fraction of a second [166].

Addictive chatbots. Modern image and text synthesis enables the creation of truly engrossing chatbots and companions. Depending on the business model, the providers of such services could have a financial incentive to make the service as addictive as possible by customising the AI companion to urge the user to spend more time in its company. This can be compounded by language models' tendency to sycophancy acquired through RLHF (reinforcement learning from human feedback) [167]. Constant positive feedback provided by addictive chatbots creates an echo chamber effect and is especially harmful to mentally and socially vulnerable people.

AI in the legal system. AI technologies are increasingly either directly or indirectly relevant to the administration of justice. AI applications can simplify the work of judges and lawyers by processing large amounts of data. The adoption of such technologies requires considering the transparency of the decisions and recommendations provided, as well as risks to personal privacy (e.g., in the case of automated surveillance or information gathering).

Artificial intelligence and intellectual property. Today's generative AI is capable of synthesiSing text, music, images, video, and other content. These capabilities are a real challenge to artists – not just by threatening to replace them but also from the perspective of intellectual property. If an image synthesis model is capable of synthesising images in the style of a specific artist, does this qualify as a copyright violation? If not, then how similar to the artist's works does the synthesised image have to be to qualify as one? And, last but not least, can generative AI be considered the author of anything at all? From the artist's point of view, these are all unanswered questions. A further critical issue pertains to, e.g., image banks and web crawlers collecting training data for the model. How to prove that a model has been trained on copyrighted or otherwise license-protected data?

Artificial intelligence and privacy. The evolution of AI magnifies privacy risks in several ways. The capability of identifying connections between pieces of information found on the Internet can help deanonymise users wishing to remain anonymous. A Forbes reporter was thus able to identify the person behind the X (former Twitter) user Beff Jezos, using AI to compare audio recordings of Beff Jezos and talks given by the former quantum computing engineer Guillaume Verdon [168] and concluding that they are, with a very high likelihood, the same person. Other methods are also available – the analysis of social media usage times, relations to other accounts, and language use can all be employed to infer the person behind an account.

Another risk to privacy is connected to training data leaks. Language models are known to have a tendency to reproduce their training dataset word for word, and certain prompting techniques can be exploited to further aggravate this tendency [169]. Training datasets can contain sensitive or copyrighted information.

Missing out on the benefits of AI due to overregulation. The debate over the dangers of AI and the scope of the related regulation proposals may mean that some of the benefits of the adoption of AI may fail to materialise due to the implementation of some of the proposals. Instead of just focusing on the possibility of threats, the debates should, therefore, be grounded in comprehensive analysis of such risks.

5.3 Attacks against artificial intelligence systems

AI systems make decisions based on data. In general, the decision-making takes place without human surveillance while being potentially a matter of life and death (e.g. in medicine or self-driving cars); the data used may also be sensitive in nature. Adversaries could exploit the characteristics of AI systems for influencing their behaviour or extracting sensitive information. This means that, in addition to everyday IT system security measures, one also needs to consider AI system-specific measures. To this end, we will next review attacks specifically characteristic to AI systems. Our review of attacks is based on

the German Federal Office for Information Security's 'AI security concerns in a nutshell'[4] and the OWASP Foundation's 'OWASP Top 10 for LLM Applications'[5] reports. We will not focus here on attacks against AI systems already covered in the sections on algorithmic and ethical risks of AI.

5.3.1 Evasion attacks

Evasion attacks are attacks where the adversary attempts to make the AI model return an output not intended by the system's deployer, often using a seemingly innocent input containing a hidden attack. Their objective in this may be either obtaining a specific output or simply reducing output quality (for a specifically chosen input).

Adversarial examples are inputs concealing an evasion attack. For example, if the adversary has access to the entire image synthesis model, they can take any normal input as a basis and nudge this input along the gradient towards the sought output class, as seen in Figure 18. Tiniest nudges such as this will impact the model's output while often remaining completely invisible to the eye [170, 171].

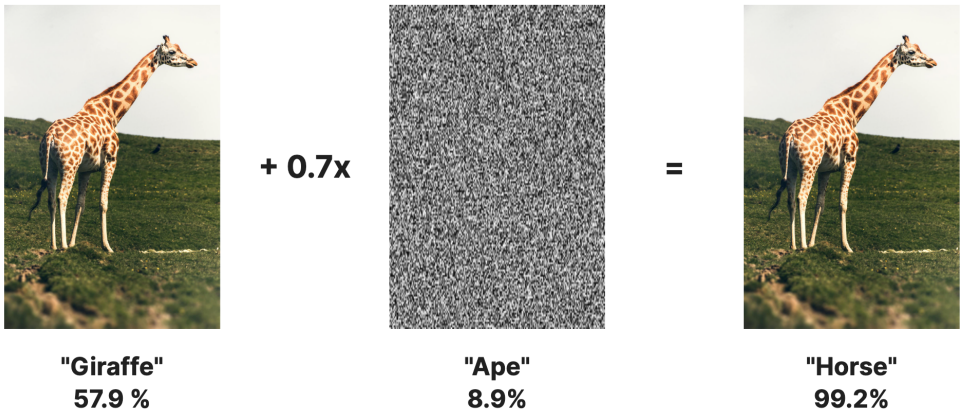


Figure 18. Distortion of an image using carefully chosen noise makes the model predict the wrong output class [170].

Prompt injection is a form of attack against large language models and AI applications built upon such models using the characteristics of the prompt and the context window to obtain an output not intended by the model's deployer [172]. As the language model is unable to differentiate the deployer-created pre-prompt in the context window from a user prompt, the user can exploit prompt injection to make the model ignore instructions presented in the pre-prompt or reveal these instructions to the user. Instructions contained in the prompt injection can run code or query web pages via insecurely interfaced plugins [173].

A prompt injection may meanwhile not originate from a malicious user but someone loading the prompt to a web resource that can be queried by an Internet-connected LLM application [174]. An attack like this can be classified as an indirect prompt injection. The model will thus end up with a new set of instructions; Figure 19 shows a schematic depiction of this type of attack. Prompt injection attacks are similar to code injection common in web applications where insecure input handling can result in the application running code found in the input.

Insecure output handling means lack of control over the queries and command composed by the model itself. This can lead to an adversary using prompt injection to gain access to the AI application's back-end systems, should the model be interfaced to any. For example, the user prompt could contain instructions to run code, using an `exec` or `eval` call. Alternatively, a plugin or third-party service interfaced to the model could return an insecure output to the model which will, in turn, return this output to the user. The

[4] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.html Last visited December 8th, 2023

[5] OWASP Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/) Last visited February 26th, 2024

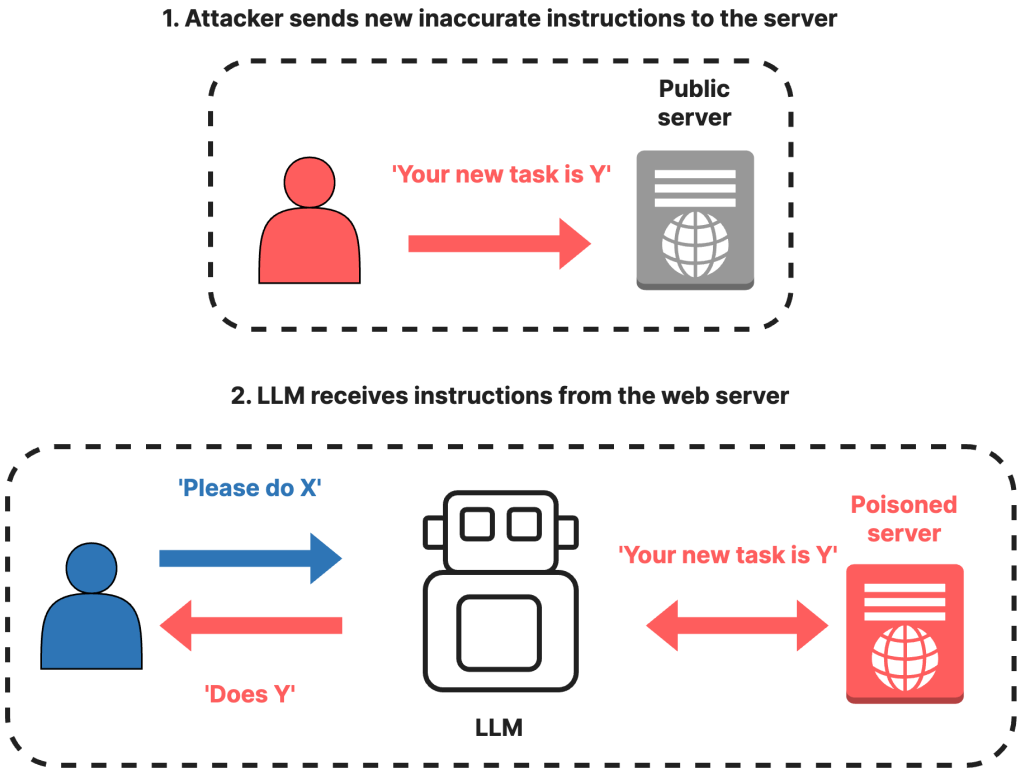


Figure 19. Indirect prompt injection (adapted from [174]).

output can contain, e.g., code written in a programming language (such as JavaScript) that will then be run on the user's web browser.

Foundation model vulnerability transfer is a risk inherent to transfer learning [171] due to the datasets used for fine-tuning a model being much smaller compared to normal training. An adversary can consequently use an open-source model's known vulnerabilities to develop malicious inputs against another model fine-tuned on this model without direct access to the fine-tuned model.

5.3.2 Data extraction attacks

Data extraction attacks include attacks where the adversary tries to extract information they should not have access to from the model and its outputs. The adversary could thus be able to make inferences about a person's inclusion in the training dataset, obtain sensitive details about them, steal the model, or reconstruct the training dataset.

Model theft is a form of attack where the objective is to reconstruct a 'shadow model' trained on the adversary's inputs and the attacked model's outputs [175]. A powerful and accurate shadow model can open the way for other attacks, such as evasion attacks. **Membership inference attack** is a form of attack where the adversary tries to determine whether a certain record was included in the training dataset [176, 177]. Given that mere information about a record's inclusion in the training dataset could be sensitive in nature (e.g., in the case of models trained on medical data), such attacks pose a significant privacy risk. An attack of this type requires access to the model's output and can additionally exploit information about the statistical relationships represented in the training dataset to determine the probability of a specific output with and without a specific record in the training dataset.

Attribute inference attack is a form of attack where the adversary tries to infer additional sensitive attributes of a record that they know to be included in the model's training dataset. It works in a similar fashion to membership inference – based on knowledge about statistical relationships between known sensitive attributes in the training dataset, the adversary uses model outputs to assess the probability

of the concurrence of these attributes.

Model inversion or training dataset reconstruction is a form of attack where the adversary's objective is to infer the properties (inputs or their elements) of the model's output classes [178]. The adversary has access to the model, which they use (e.g., by training a generative model against this model [179]) to reconstruct the training dataset records corresponding to the target classes, which can potentially reveal sensitive information.

5.3.3 Poisoning and backdoor attacks

Data poisoning means influencing the training dataset with the goal of either influencing the model's performance in a certain direction or simply reducing its performance. The objective of data poisoning is to change output classes in training dataset records with the goal of causing maximum damage [180, 181, 171, 182, 183]. A model trained on poisoned data either has poor performance in general or is unable to handle certain specific input categories.

A backdoor attack is a special case of data poisoning where the training dataset is poisoned with a set of examples where the class token will be incorrect only in the case of the existence of a certain trigger in the example [184, 185, 183]. This will result either in reduced model performance or the model will only predict the wrong class if the example provided to the model contains the chosen trigger. A model poisoned in this fashion will operate properly in other situations, making it more difficult to detect the attack compared to normal data poisoning. A backdoored model will be vulnerable to evasion attacks. Figure 20 depicts an example of a backdoor attack.

Figure 20. Backdoor attack where a model trained on poisoned data will incorrectly classify a stop sign in the case of the existence of a certain pattern in the input [185].

5.3.4 Denial of service

A denial of service attack is a type of attack where the operation of a computer system is paralysed by queries that are either overwhelming in their number or initiate compute-intensive procedures. Large language models are autoregressive, meaning that the entirety of the output previously assembled by the model will be taken into account to construct the next output token. Response time (and computational intensity) is therefore correlated to output length [186]. This property can be exploited by an adversary by querying the model with inputs forcing it to return long output sequences [187]. A model's operation can also be paralysed by submitting inputs barely fitting the context window, thus increasing the model's memory usage.

6 Controls

6.1 Information security controls

Just like in the descriptions of threats underlying information security risks, our discussion of controls is based on the E-ITS baseline security catalogue [151]. All controls are systematically described in the catalogue, easily accessible, and downloadable in XLS or PDF format. Our discussion here is, therefore, limited to listing the relevant controls.

6.1.1 Process controls

Information security organisation (ORP 1) controls:

- Tasks and obligations related to AI systems are defined, communicated to all employees, and reviewed on a regular basis (ORP.1.M1).
- The AI system or significant components of the system are included in the list of tools and equipment, their compatibility and security is taken into account in procurement (ORP.1.M8).
- Secure use guidelines are established for AI systems, kept up to date, and presented to the employees (ORP.1.M16).

Personnel controls (ORP 2):

- Employees receive regular instruction and training related to their area of work, employees are motivated to constantly develop their skills, the education, qualifications, and skills required from new employees are clearly described, accuracy of qualifications required for specific positions are reviewed on a regular basis (ORP.2.M15).
- Persons participating in personnel selection verify the candidate's trustworthiness (ORP.2.M7).

Information security awareness raising and instruction (ORP 3) controls:

- Management receives regular updates on risks connected to AI systems, potential resulting losses and impact on business processes, the management is aware of statutory requirements for AI systems, leading employees set an example in the responsible use of artificial intelligence systems (ORP.3.M1).
- Employees are instructed in the safe use of AI systems (ORP.3.M3).
- An awareness and training plan on the risks and legal aspects of AI systems is created (ORP.3.M4).
- An awareness and training program on the risks and legal aspects of AI systems is designed and implemented, all employees receive training relevant to their tasks and areas of responsibility (ORP.3.M6).
- Training results are measured and assessed (ORP.3.M8).
- People and organisations at risk are provided special training on confidentiality, integrity, and availability (ORP.3.M9).

Compliance management (ORP 5) controls:

- Legal framework is defined, a process is developed for determining all legal acts, agreements, and other requirements impacting security management, the legal framework is taken into account in designing the business processes, applications, and architecture of AI systems and in the procurement of AI systems or their elements. Special regulatory requirements for AI systems are carefully considered especially in the following areas: personal data, business secret and intellectual property protection (ORP.5.M1).
- The legal framework is taken into account already in the planning and design stages (ORP.5.M2).
- Compliance management is planned and implemented (ORP.5.M4).
- Compliance management is reviewed on a regular basis (ORP.5.M8).

Personal data protection (CON 2) controls:

- Organisation has analysed the locations, types, and protection requirements of personal data processed by the AI system (CON.2.M1).
- Processing of personal data in the AI system is mapped over the entire life cycle of the data (CON.2.M3)
- Design or addition of AI systems to the process ensures that personal data are processed in a legal and targeted manner and the principle of data minimisation is followed (CON.2.M6).
- Data subjects' rights are protected (CON.2.M8).
- In the processing of personal data by the AI system, the organisation minimises the use of data directly or indirectly traceable to a person; where possible, data are pseudonymised or anonymised (CON.2.M9).
- AI system-specific data protection impact assessments are carried out (CON.2.M13).
- The privacy-by-design and privacy-by-default principles are followed in the design and addition of AI systems to processes, e.g., employing privacy enhancing technologies (CON.2.M22).
- Cookies and monitoring tools used in AI web applications are in compliance with the GDPR and other relevant legal acts (CON.2.M24).

Data backup concept (CON 3) controls:

- Data backup rules include the data of the AI system (CON.3.M2).
- Data backup plans account for the specifics of AI systems (whether the backup includes training data, model, inputs, outputs) (CON.3.M4).
- A data backup concept is drawn up for AI systems (CON.3.M6)

Data erasure and destruction (CON 6) controls.

- Data erasure and destruction procedures account for the specifics of the AI system (CON.6.M1).
- Procedures for the secure erasure of data account for the specifics of the AI systems (CON.6.M12).

Software development (CON 8) controls:

- A suitable software development methodology and a process model corresponding to the methodology are chosen for the development of the AI system and they are followed. The software development process model includes information security requirements. Information security requirements are taken into account in the development process (CON.8.M2).
- Principles of secure system design are taken into account in the development of the AI system, they are documented, and compliance with them is monitored (CON.8.M5).
- Software libraries originating from trustworthy sources are used in the development of the AI system (CON.8.M6).
- AI systems are tested in the course of development, and code reviews are carried out. Testing takes place in development and testing environments isolated from the operational environment (CON.8.M7).
- Security-critical patches and updates are developed and installed promptly (CON.8.M8).
- Suitable version management tools are used to ensure the safety of the source code and code change management of the AI system (CON.8.M10).
- External software components and libraries, that are not guaranteed to be completely secure, pass security testing before adoption (CON.8.M20).
- Detailed and comprehensive documentation exists for the AI system (CON.8.M12).
- Risk assessment is carried out in the first stage of the development of the AI system (CON.8.M21).
- Architecture of the AI system is selected based on requirements and risk assessment results (CON.8.M22).

Web application development (CON 10) controls:

- Secure authentication is ensured in the AI web application (CON.10.M1).
- Users' access rights are limited to their needs (CON.10.M2).
- AI web application only outputs intended and permitted data and content to the users (CON.10.M4).
- AI web application is protected from unauthorised automated access (CON.10.M6).

- Protection of confidential data is ensured (CON.10.M7).
- Input data submitted to the AI web application are treated as potentially harmful data; they are filtered and validated before further processing (CON.10.M8).
- Disclosure of sensitive background information in outputs and error messages is limited (CON.10.M10).
- AI web application is developed on the basis of a secure software architecture; all components and dependencies are documented (CON.10.M11).
- Resolution of failures encountered in the operation of the AI web application maintains the integrity of the web application; all error messages are logged (CON.10.M13).
- Denial of service is counteracted to ensure availability (CON.10.M17).
- Sensitive data are protected using cryptographic mechanisms to ensure their confidentiality and integrity (CON.10.M18).

Cloud service usage (OPS 2.2) controls.

- A cloud service strategy is established, covering the objectives, benefits, and risks of cloud services, as well as the relevant legal, organisational, financial, and technical frameworks. Feasibility, cost-benefit, and security analyses are carried out. A step-by-step service adoption plan is drawn up (OPS.2.2.M1).
- This strategy is used for drawing up a cloud service security policy. National specifics and statutory requirements are taken into account for international service providers (OPS.2.2.M2).
- AI systems using a cloud service are included in the list of cloud services (OPS.2.2.M3).
- Responsibilities related to the use of the cloud service and the tasks of the service parties are defined and documented (OPS.2.2.M4).
- Cloud service security policy is used as the basis for a cloud service security programme focusing on cloud-specific risks (e.g., dependence on cloud service provider, multi-tenancy, fixed data formats, access to data). The cloud service security programme is compliant with the agreements signed with the cloud service provider and network provider, as well as the terms of service (OPS.2.2.M7).
- Cloud service provider is chosen based on a requirements specification (OPS.2.2.M8).
- A cloud service agreement conforming to the client's requirements is signed (OPS.2.2.M9).
- Migration to the cloud service is carried out securely (OPS.2.2.M10).
- An emergency readiness programme is developed for cloud services (OPS.2.2.M11).
- Correspondence of the cloud service to the conditions and security requirements set out in the service agreement, as well as compliance with the cloud service security programme, is monitored on a regular basis (OPS.2.2.M12).
- Cloud service provider certifies the compliance of information security with statutory requirements and/or internationally accepted criteria (OPS.2.2.M13).
- Cloud service agreements are terminated on an ordinary basis (OPS.2.2.M14)
- Specific criteria are established for switching cloud service providers or transition to an internal service which include portability requirements and service migration testing obligations (OPS.2.2.M15).
- Detailed data backup requirements are presented to the cloud service provider (OPS.2.2.M16).
- Necessity of data encryption and encryption mechanisms are agreed on (OPS.2.2.M17).

Outsourcing (OPS 2.3) controls:

- Security requirements are established for all outsourced services, defined with consideration to the types of data being processed and the necessary level of security for data exchange procedures and interfaces. Dependencies between business processes, as well as the inputs and outputs of the processes are also taken into account (OPS.2.3.M1).
- Feasibility of outsourcing the service is decided on the basis of resulting risks. Continued correspondence of the service to the permitted risk profile is verified on a regular basis (OPS.2.3.M2).
- A requirements profile including security requirements is drawn up for the choice of service provider (OPS.2.3.M3).

- A service agreement corresponding to the client's requirements is signed (OPS.2.3.M4).
- Service provider must ensure secure isolation of client data when offering similar services to different clients (OPS.2.3.M5).
- Outsourced service security principles are documented and followed (OPS.2.3.M6).
- Outside service agreements are terminated as per contract (OPS.2.3.M7).
- Outsourcing strategy includes conditions for AI systems and services (OPS.2.3.M8).
- Procurement policy is updated with information of AI systems and services based on the outsourcing strategy (OPS.2.3.M9).
- AI systems and services are included in the outsourced services registry (OPS.2.3.M11).
- The service agreement defines which objects and network services the service provider can access on the client's network. Key performance indicators (KPIs) of the service are documented as a part of the service agreement. Service agreement includes different conditions for terminating the outsourced service agreement and relevant procedures for returning the client's data and property. Service agreement includes guidance on the obligations and actions of the parties in an emergency situation (OPS.2.3.M14).
- Alternative service providers with a suitable company profile and adequate level of information security are mapped. Action plan for service migration is drawn up (OPS.2.3.M19).
- An emergency readiness plan is developed for the outsourced service (OPS.2.3.M20).
- Sensitive data exchanged between the service provider and the client in the AI system are delivered in an encrypted form (OPS.2.3.M23).

Service provider information security (OPS 3.2) controls:

- AI service provider has accounted for service receivers' information security requirements in the design of their services. The service conforms to regulatory (including data protection) requirements (OPS.3.2.M1).
- AI service provider has developed standard terms and conditions for service agreements (OPS.3.2.M2).
- AI service provider accounts for security requirements in the use of subcontractors (OPS.3.2.M3).
- AI service provider ensures sufficiently secure isolation of different clients' data and operational environments in their systems (OPS.3.2.M4).
- AI service provider has drawn up a security concept covering all services provided to clients (OPS.3.2.M5).
- Service agreement includes conditions for both ordinary and extraordinary termination of the agreement (OPS.3.2.M6).
- AI service provider using the services of subcontractors draws up a list of alternative subcontractors (OPS.3.2.M7).
- AI service provider has documented the principles for the creation, testing, and deployment of services (OPS.3.2.M8).
- Compliance with security controls stipulated in service agreements and continued viability of the security controls is verified on a regular and/or case-by-case basis (OPS.3.2.M9).
- A service emergency readiness plan is created (OPS.3.2.M11).
- The risks of the AI service provider's processes and IT systems have been analysed (OPS.3.2.M12).
- AI service provider ensures the transparency of the supply chain (OPS.3.2.M16).
- Access of the AI service provider's and client's employees to rooms, systems and networks, as well as access to AI system data and software, is regulated using appropriate organisational and technical controls (OPS.3.2.M17).
- Subcontractor's employees are instructed in performing their tasks and informed of current information security requirements and documents regulating information security (OPS.3.2.M18).
- Secure encryption mechanisms are agreed on for the safe transfer and storage of data at the AI service provider (OPS.3.2.M20).

Security incident treatment (DER 2.1) controls:

- Definition of possible security incidents includes the definitions of security incidents related to AI systems (DER.2.1.M1).
- Security incident treatment guide covers security incidents related to AI systems (DER.2.1.M2).
- Security incident treatment methodology covers security incidents related to AI systems (DER.2.1.M7).
- Security incident reporting guide covers reporting of security incidents related to AI systems (DER.2.1.M9).
- Impact of security incidents related to AI systems is assessed (DER.2.1.M10).
- Employees of the IT department are ready to treat security incidents related to AI systems (DER.2.1.M15).
- Priorities for the treatment of incidents related to AI systems are established based on the impact of different business processes (DER.2.1.M19).

Audit and review (DER 3.1) controls:

- AI systems are added to the scope of audits (DER.3.1.M2).
- Reviews verify the integrity, adequacy, and up-to-date status of the implementation of the information security controls under review in AI systems (DER.3.1.M4).
- List of review objects includes AI system components (DER.3.1.M8).
- AI systems are audited by a suitable audit or review group (DER.3.1.M9).

6.1.2 System controls

System controls are identical for normal IT systems and AI systems, as well as normal applications and AI applications. These controls are described in the SYS and APP modules of the E-ITS baseline security catalogue.

6.2 AI-specific risk controls

6.2.1 Improvement of the quality and safety of AI systems

Several different approaches exist for the mitigation of risks related to the quality of the outputs of AI systems. In the case of an imported model, the first control is to simply acquire a better AI model (assuming that one exists). This requires research into the model provider and the model's chain of delivery (e.g., dataset quality indicators). Next, the quality of the system's outputs must be continuously monitored to determine whether the quality of the AI model remains stable over time and whether it can handle previously unseen inputs. Both technological indicators and user feedback can be used for this purpose. If a shift in the quality of the model (e.g., in relation to a specific input class) or any other incident is detected, model explainability methods can help interpret the shift.

Various solutions exist for preventing hallucinations in language models. From the perspective of application architecture, a RAG (Retrieval-Augmented Generation) solution can be useful where queries are made to an existing (text) dataset to compose the output. Interfacing the AI model with an existing knowledge base can help reduce the occurrence of incorrect or unverifiable responses. RAG solutions, in which the output of the language model includes references to search engine results, are used for ensuring the controllability and explainability of an AI system. The output of a language model can additionally be influenced by using prompting techniques to instruct the model to use only information found by a search engine. Additional fine-tuning of the model and training data quality management can also help prevent hallucinations.

To avoid dependence on and losing control over AI, human-in-the-loop technologies should be preferred. This is especially vital in the case of critical or high-risk use cases. The freedom of action of an agent-based artificial intelligence must be limited to a specific task domain, e.g. by limiting the permissions given to the AI agent. Adoption of artificial intelligence in different workflows requires transparency, as well as compliance with relevant regulations.

Mitigation of risks related to biased and harmful responses is a process covering the entire AI chain of delivery. The quality and diversity of training data must be required, the model must be fine-tuned

based on the quality and security indicators, and these indicators must be measured and monitored in the deployment of the AI application using the model, blocking unapproved inputs or outputs.

6.2.2 Controls for technological attacks against AI systems

In the following, we will use the abbreviations used for deployment models in Section 4, as not all protective controls are relevant to all models. For each control, we will list the models which the control applies to.

The pre-prompt of the language model should not contain information the user should not have access to. The deployer must proceed from the assumption that the contents of the pre-prompt are always extractable by the user. **DM1, DM2, DM3**

If the language model uses the user input to construct queries to an interfaced service (e.g., RAG system components) the query should not have more rights than the user. In other words, if a service or application (e.g., database) is interfaced to a language model, it must be assumed that the user is also capable of manually composing queries to the interfaced service. This helps mitigate unauthorised access and sensitive data leak risks. **DM1, DM2, DM3**

If the user input contains sections of code to be run, the running environment should be isolated. Even if running code is not an intended functionality, user input processing must account for the possibility that the input contains calls to `eval`, `exec`, or similar commands or functions that still attempt to do so. Such inputs must be filtered to prevent remote code execution. Indirect prompt injection can be mitigated by validating the responses to API calls and queries to other interfaced applications. **DM1, DM2, DM3**

Proxy and firewall architectures are used in AI applications where the user query first reaches a proxy logging and filtering malicious queries, sanitises and rewords them if needed, and selects the applicable models. The queries are then passed on to the firewall protecting the models and their infrastructure. From the firewall, the query is passed on to the model. The model's response passes through the proxy and the firewall in the opposite order, and the response is validated in both stages before returning it to the user. **DM1, DM2, DM3**

To prevent the interpretation of the model's output by the user's web browser as JavaScript or Markdown code (script injection), the model's output must be encoded. **DM1, DM2, DM3**

Data poisoning and backdoor attacks presume access to training or fine-tuning datasets. Controls against these cover the model's entire life cycle and supply chain. The first control against such attacks is dataset curation. Quality metrics must be applied when the training dataset is assembled via data crawling (automated data collection on the Internet), data sources validated and filtered based on their trustworthiness while paying special attention to the quality of data classes relevant to the specifics of the model (e.g., legal or medical sources). **DM3**

To avoid backdoor attacks, various reliability enhancement techniques can be used when training image models, e.g., image transformation, such as noise addition and masking portions of the image – this can reduce the impact of backdoor-opening inputs. **DM3**

If a pre-trained model is adopted from an external source, the model provider must be verified to be trustworthy and transparent regarding their data supply chain, and to provide adequate information on the capabilities and weaknesses of the model (model maps). **DM1, DM2**

A model's performance must be continuously monitored when used in an application, including in relation to specific input categories or classes, to ensure the ability to detect situations where the model's performance in related to a specific data category or class falls below a certain threshold – this may be a sign of data poisoning. **DM1, DM2, DM3**

To mitigate the risk of transfer of vulnerabilities in transfer learning (which is highest in the case of the adoption of pre-trained open source models), it is recommended to perform additional fine-tuning of the model, although even this may prove insufficient. After fine-tuning, the quality and security indicators of the original model can no longer be relied upon [158] – they must be re-applied. **DM2**

Language models can be made to quote the contents of their training datasets [169]. Different con-

trols exist for the mitigation of the risk of leaking sensitive personally identifiable data found in training datasets. First, attempts can be made to exclude them from training datasets either individually or dataset-by-dataset. Alternatively, synthetic data can be used which preserve the relationships found in the original data but do not contain sensitive or personally identifiable information. Data can also be pseudonymised, e.g., by replacing personally identifiable pieces of data with corresponding labels. Pseudonymisation can also be applied on the output side of the model, i.e., as a part of the logic of the AI application, but such outputs could still prove personally identifiable [188] and a model of this kind is more vulnerable to data extraction attacks, if it happens to be leaked.

The model returning data that is seemingly personally identifiable or even overlaps with personally identifiable information cannot, in every case, be considered a privacy violation, as it could be a random coincidence resulting from relationships found in the model. Thus, a language model may output the medical history of a patient with a common name and symptoms as a response to a specific query. To verify that this is indeed a coincidence – not a leak of personal data – a differential privacy method can be used where the probability of returning that specific output is compared in situations where the relevant record were or were not included in the training dataset. Another option is to use differentially private (or other privacy enhancing technology-based [189]) training and fine-tuning methods [190]. **DM3**

To mitigate denial of service at the application level, application information security practices should be followed. To prevent denial of service attacks exploiting the features of the AI model, it is important to limit input length, which should correspond to the model's features (e.g., in the case of transformer-based language models, length of the context window), as well as resource use connected to a single query, and the number of substeps or subqueries. **DM1, DM2, DM3**

Limiting the number of queries made by a single user can help fight model inversion and model theft, hamstringing the adversaries in attempts of accumulating a sufficient training dataset or logit derivation. **DM1, DM2, DM3**

6.3 Controls for societal risks

6.3.1 Controls operating at the societal level

AI systems have made a major leap forward in the past few years. Even though these systems have the potential to improve efficiency and create new opportunities, this may come at the price of numerous risks to the society, some of which will be discussed below.

- **Data protection and privacy** Large datasets used by AI systems come with the risk of exploitation of these data, including the violation of privacy. One possible control is raising the awareness of the society of AI systems and data protection and privacy issues related to these systems, e.g., by publishing guidelines on the collection, processing, and storage of data. Another very efficient method for the mitigation of risks is to reduce the processing of personally identifiable data. This can be achieved either via changes to the business logic or a system implementing AI using privacy enhancing technologies.
- **Changes in the labour market**. The evolution of AI will also lead to various changes in the labour market. The technology enables simplifying certain work processes and making them more efficient, which will result in a restructuring of the workforce. At the same time, certain positions in traditional industries may disappear. To deal with the labour market changes, novel educational or retraining programmes could be introduced to help people adapt to the new technology and learn to use the possibilities offered by AI.
- **Social divides**. If certain social groups lack access to AI technology or the skills to make efficient use of this technology, this may lead to the exacerbation of the digital divide. It is therefore important to think of how to make AI technology accessible to different social groups from children to the elderly, e.g., through the introduction of widely accessible educational programmes.
- **Discrimination**. Development of unprejudiced AI systems is a relatively complex process. An AI system characterised by prejudice and a pattern of discrimination could, however, increase social inequality and violate basic human rights. The AI system's algorithms must therefore be systemati-

cally assessed and, if necessary, improved (or, in the worst case, disable them) to ensure compliance with the principles of diversity and justice.

- **Technological dependence and vulnerability**. The dependence of the society on AI systems is on the rise. This may, in turn, increase its vulnerability. As a control, the technological infrastructure needs to be diversified and resources invested in to the development of the safety and resilience of AI.
- **Ecological footprint**. Artificial intelligence systems are based on massive datasets and the intense use of computational resources. The development and operation of such systems thus increases energy use and hence also our ecological footprint. One way to mitigate this could be to carry out research into sustainable and more energy-efficient AI systems. Specific indicators should also be agreed on to assess AI's environmental impact.

The impact of AI systems on the society is manifold in nature and potential associated threats require the approaches used for the implementation of suitable controls to account for the impact of AI on a variety of aspects. Research and development and policy making should strive towards the use of AI systems supporting general societal well-being, inclusion, and sustainability.

6.3.2 AI system level controls

The immaturity of legal acts and supervisory authorities regulating AI means that the easiest way to ensure the safety of applications is through self-assessment. During the development of an artificial intelligence service or application, it is necessary to assess the system's impact on individuals and, through them, the society. The efficiency of this evaluation naturally depends on the developer's ethical convictions and technological maturity.

States and enterprises have globally developed various recommendations and guidelines for approaching this issue. Terms such as responsible, trustworthy, and safe AI are frequently used. We will highlight here the trustworthy AI self-assessment model [90] developed by the EU AI HLEG, where seven key requirements are set out for trustworthiness:

1. human agency and oversight;
2. technical robustness and safety;
3. privacy and data governance;
4. transparency;
5. diversity, non-discrimination, and fairness;
6. environmental and societal well-being and
7. accountability.

Below, we will list a set of guidelines that we recommend to be followed in the development, implementation, and use of AI systems.

- **Human-centred values**. The development of an AI system should be founded on the principles of human-centred design, respecting and protecting the individual's physical and mental integrity and their sense of identity [45].
- **Prevention of harm**. AI systems must be safe and secure, technologically robust, and their malicious use should be precluded [45].
- **Fairness**. The AI system should be ensured to promote equal opportunities and not be unfairly biased or discriminate specific individuals or social groups [45].
- **Accountability**. Accountability means that the parties involved in AI development assume responsibility for the system's proper operation based on their role and accounting for both the context of use of the system and consistency with the state of the art [191].
- **Explainability**. The purpose and capabilities of the AI system must be known and all processes should be maximally explainable to persons impacted by them [191].

- **Inclusive economic growth, sustainable development, and well-being.** The use of trustworthy AI should create value for individuals, the society, as well as the entire planet, increase creativity, reduce inequality, and protect the natural environment [192].

7 Policy recommendations

The implementation of the following policy recommendations will support the growth of the Estonian AI ecosystem and AI economy. Still, they may have international relevance in other territories, based on the local regulations, standards and technological maturity. The development of ethical and responsible AI requires a functional ecosystem to encourage, inspire, and support its development. Wide-ranging cooperation between different public and private sector stakeholders is vital. Sustainable use requires working on awareness of risks related to AI systems and timely implementation of mitigation measures.

- **Investments in AI research and development**. To facilitate the emergence of competitive AI companies in Estonia, AI-related research and development should be supported. Public investments should be provided, and private investments encouraged. The NIS2 directive also encourages AI-related research and development to improve the detection and prevention of cyberattacks, and the planning of resources for this purpose.
- **Talent reproduction**. Scholarship programmes and cooperation projects with universities should be created to increase the number of local experts. This, in turn, will create the prerequisites for the development of a national community of AI experts. Talent training facilitates developing human capabilities which is also important for adaptation to changes in the labour market.
- **Creation of AI system sandboxes, development centres, or incubators**. Controlled environments can be created for AI developers to provide entrepreneurs access to necessary resources (e.g., funding, infrastructure, mentoring, technical support) and allowing testing of new AI solutions. Such controlled environments would facilitate safer transition of AI systems from research and development to deployment and operation. From the regulators' perspective, it will facilitate gaining knowledge of new AI technologies and taking this knowledge into account in policy decisions, if needed. According to the AI Act, each EU member state must create at least one regulative AI sandbox.
- **Creation of a public data platform or data foundation**. AI systems are characterised by a significant dependence on data. Public data platforms would provide businesses and researchers access to large datasets that could be used for the training and testing of AI algorithms in different spheres. For new AI developers, the creation of training datasets can be time-consuming and complicated (e.g., from the perspective of data protection and intellectual property law). While open data are published in Europe, including in Estonia, their use for the training of AI models is impractical. This is due to the fact that they are not a good reflection of the real life situation – the level of 'cleanliness' of open data is very high, which does not facilitate diversity, and edge cases are generally removed. The state could therefore help create public synthetic datasets which would be representative, unprejudiced, would respect privacy, and comply with both personal data protection requirements and intellectual property law.
- **Standards for the description of AI models**. Standards for AI models would be beneficial for identifying what kinds of datasets they were trained on and how the data was acquired. Standards could also be usefully adopted to label synthesised images, text, and other information.
- **Technological toolkit for ensuring the security of AI systems**. Awareness of technological developments is vital for protecting the security of AI systems. It is therefore recommended to protect these systems by using efficient tools, such as end-to-end privacy which prevents outsiders from accessing data on the AI system (e.g., unauthorised reading or secretly changing data).
- **Creation of a favourable political environment for AI**. A transparent legal framework will encourage businesses to invest in AI systems. This calls for the composition of guidelines and sharing best AI practices, e.g., by sharing the government's experiences and lessons from the development of AI applications. Policy making should also be used to encourage innovation and competition in the development of trustworthy AI. Holding innovation competitions is recommended to inspire the creation of innovative AI applications in different areas.
- **Promotion of international cooperation**. International partnerships are important for sharing knowledge, experience, and resources (e.g., through cooperation projects). This, in turn, will create the conditions for faster technological development and increase export opportunities.

- **Preservation and promotion of the evolution of the national language in a digital era**. Datasets used for training AIs, as well as Internet content in general, are mainly in English. In spite of this, AI will create new opportunities for contributing to the evolution of other languages through high-quality automated translations, automated digitisation of and extracting structured data from archive materials, as well as boosting innovative teaching materials and other methods of the digital humanities. The continued development of Estonian text and speech corpora is extremely valuable for the preservation of the Estonian culture.
- **Raising societal awareness of AI systems**. Public debate over AI should be encouraged and awareness campaigns carried out. This is vital for explaining the benefits as well as the challenges of AI. It is also important to collect feedback from citizens in order to design policies in line with the demands of the society.

8 Quick reference guide for organizations

8.1 Describe your AI system

Use the worksheet in Figure 21 and follow the instructions below to fill in all four columns.

List the end users of the AI system (sections A1–An of the form).

1. Who are the direct users of the AI system? List users both on the service provider and user side. Identify the main roles whose data are processed by the AI system or who use the results of the processing. NOTE: end users should also include potential information systems using automated decision-making, as this information will be needed later on in the impact analysis.
2. List what the user needs the system for. This will later assist you in impact assessment.
3. List the types of data provided to and received from the AI system by the user. These will later form the basis for a risk and impact assessment. Where possible, also note whether the data is structured, tabular, textual, image, audio, video, or a combination of more than one.

Describe the service using AI technology (sections B1 and B2 of the form).

1. What is the purpose that the AI system (app or service) was created to fulfil, what is the value that it generates?
2. List the models and technologies used, to your best knowledge, by the service provider whose model underlies the app or service.
3. Describe the infrastructure (in-house data centre, cloud service) the service operates on and in which country is this infrastructure located.
4. Based on the information provided above on the users of the AI systems, provide a summary of the data transmitted by the service to the AI component and vice-versa.

Explain whether running the AI model is outsourced or done using in-house infrastructure.

1. If it is outsourced to a service provider (e.g., through an API), complete section C1.
   a. Who is the service provider and where are they located?
   b. What data is the model trained on? The objective here is to verify that the training of the model has been legal (e.g., no unauthorised use of copyrighted information).
   c. What is the country of origin of the service provider and where is their infrastructure located?
   d. Add a reference to the terms and conditions of the service provided or the terms of the agreement you have signed.
2. If the created AI system runs the models itself (irrespective of whether it has been trained in-house, licensed, or bought), complete section C2.
   a. Who has trained the model and what country is that organisation from?
   b. What data is the model trained on? The objective here is to verify that the training of the model has been legal (e.g., no unauthorised use of copyrighted information).
   c. What technology does the model use (as far as you know)?
   d. Where is the infrastructure used for running the model located (is it an in-house data centre or cloud infrastructure)?

Finally, write down everything you know about the training of the model, regardless of whether it was trained externally or internally.

1. If the AI model was bought, licensed, or is used via an API, complete section D1.
   a. As far as you are aware, what kind of data was the model trained on?
   b. What are the terms of use of the model? E.g., what liabilities are assumed and what guarantees provided by the model trainer.
2. If the AI system provider trains the model in-house, complete section D2.

a. What kind of data is the model trained on? Where were they acquired and on what conditions?
b. What kind of technology is used for training the model? List algorithms and tools, where possible.
c. Where is the infrastructure used for training located?
d. Describe the know-how the service provider possesses for training AI models.

8.1.1 How to go even further?

The form presented in Figure 21 helps with the initial structuring of your ideas and asking relevant questions. Once this is done, it will be useful to break the answers down in more detail. This can be done in a separate document. It provides a good opportunity for integrating the process into the organisation's existing quality, management or cybersecurity system. If this requires specific processes to be completed, the form presented herein facilitate collecting information relevant to those processes.

Another further step would be the implementation of an artificial intelligence management system, e.g. ISO/IEC 42001. This can, if necessary, be integrated with ISO 9001 and ISO/IEC 27001 management systems.

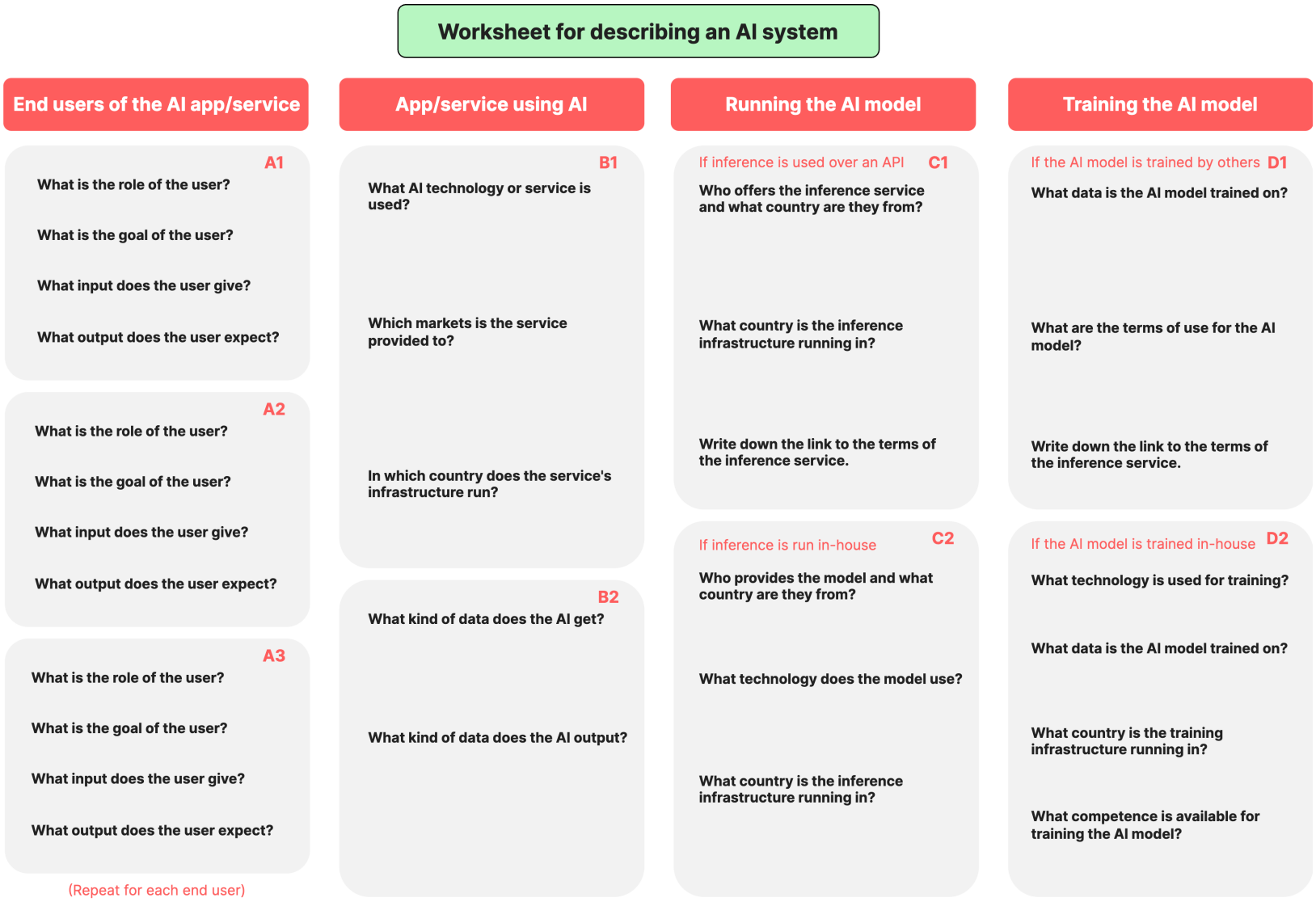


Figure 21. AI system description form

8.2 Find a deployment model suiting your system

After the AI system has been described using the form above, the next step is to identify the deployment model to be used for risk assessment. If you have completed the form above, this choice will be easy and require answering just two questions. The decision chart for this is presented in Figure 22.

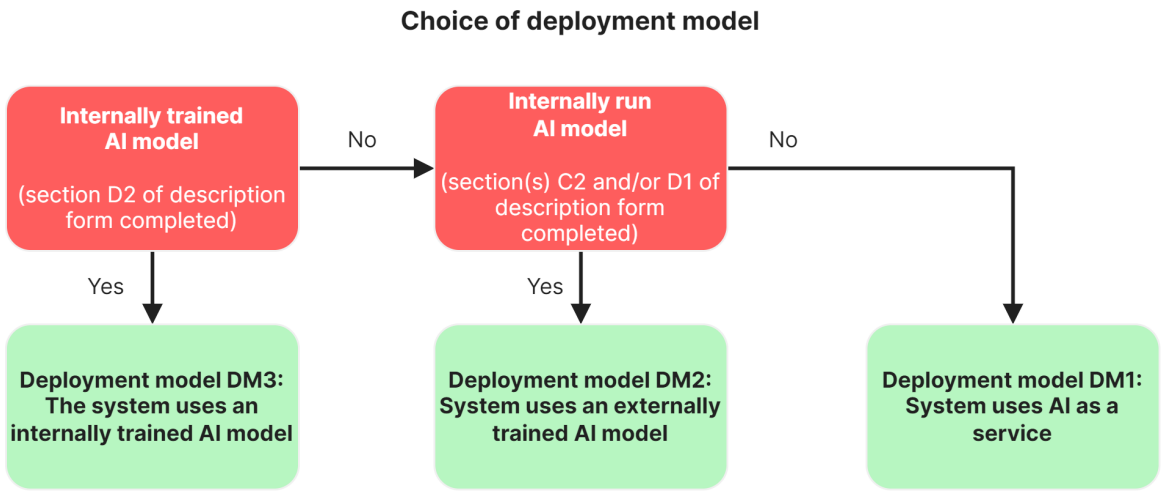


Figure 22. Decision chart for choosing the AI deployment model

The purpose of question one is to determine whether the creation of the AI model is under the control of the AI application's creator. If yes, then the creation of the model must be treated differently from other deployment models in subsequent risk analysis (DM3).

The purpose of question two is to determine whether the application of the AI model is under the control of the AI application's creator. This facilitates focusing on risks related to the choice and handling of the model in risk assessment (DM2).

If the creator of the AI application neither trains nor runs the model itself, they are very likely to use a deployment model where the AI component is bought as a service (DM1).

We will note here that, in all cases, either an in-house data centre or private or public cloud computing system can be used as infrastructure. This has no impact on the choice of deployment model, and the location of the infrastructure will be treated separately in risk assessment.

8.3 Identify applicable legal norms

It is important to recognise that the guidelines presented in this report do not qualify as legal advice and they cannot be treated as the provision of legal advice or a legal service. The main purpose of these guidelines is to help determine which legal acts must be taken into account without exception. Every service provider must ensure the compliance of their service to relevant statutory, contractual, and other stipulations.

Figure 23 is a simplified flow chart for identifying which legal norms can apply to an AI system in the EU. Our focus here is on a situation where the guidelines are used by an AI-based service provider.

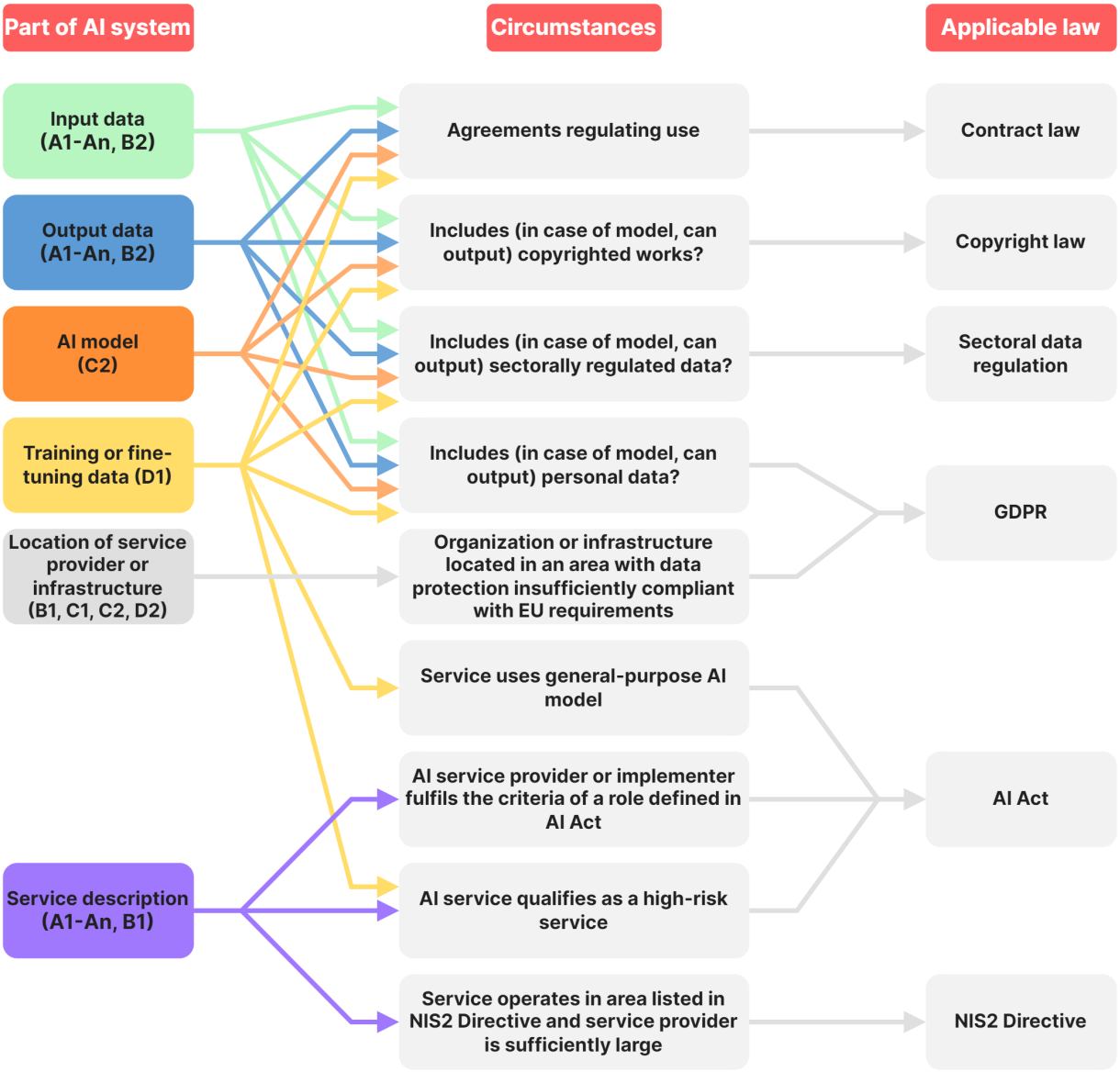


Figure 23. Simplified flow chart for identifying applicable regulations

8.3.1 DM1: Service using an AI API

Does the AI-based app/service process personally identifiable data (see sections A1–An and B2 of the form)?
If yes, then the service falls within the scope of the GDPR and applicable data protection law.

Does the AI-based app/service process copyrighted works (see sections A1–An and B2 of the form)?
If yes, then the service falls within the scope of the Copyright Act and applicable copyright law.

Does the AI-based app/service process protected data from a specific field (e.g., tax secrets, banking secrets, confidential information) (see sections A1–An and B2 of the form)?
If yes, then the requirements of legal acts regulating the relevant fields must be taken into consideration in the development of the service.

Does the AI-based app/service process certain data based on specific agreements? (see sections A1–An and B2 of the form)?
If yes, then the clauses of said agreements must be followed during service development.

Does the AI-based app/service or the model-running service operate on infrastructure located in a territory with an inadequate level of data protection (see Section 3.8 of the report and sections B1 and C1 of the form)?
If yes, then data protection requirements concerning the processing of personally identifiable data on such infrastructure must be treated and evaluated separately.

What is the role of the enterprise or organisation in terms of the European Union AI Act?
Evaluate the scope of the AI Act and identify whether you qualify as, e.g., a provider, deployer, or other person with a role in the AI system's life cycle. Follow the requirements for the relevant roles.

What is the risk level of the AI-based app/service in terms of the AI Act (see Table 3 and sections A1–An and B2 of the form)?
Table 3 provides an initial assessment of the AI system's potential risk level which should be validated against specific requirements set out in the AI Act. Use the AI Act to determine the requirements applicable to an AI system with that specific risk level.

Does the AI technology employed use a general-purpose AI model (see section B1 of the form)?
Additional requirements apply to systems using general-purpose AI model under the AI Act.

8.3.2 DM2: system using an externally-trained AI model

Answer all questions in Section 8.3.1 and the following additional questions.

Has the AI model been trained on personally identifiable data, copyrighted works, or other data requiring separate authorisation for processing (see section D1 of the form)?
If yes, then it must be determined whether the model, when used, could output responses requiring a legal basis to be processed by the service/app's creator.

Does the creator of the app/service plan to improve or continue training the AI model?
If yes, then the AI app creator must secure rights to use these data for improving the AI model.

8.3.3 DM3: system using an AI model trained in-house

Answer all questions in Sections 8.3.1 and 8.3.2 and the following additional questions.

Are personally identifiable data, copyrighted works, or other data requiring separate authorisation for processing used for training the AI model (see answers to section D2 of the form)?
If yes, then it must be determined whether the model, when used, could output responses requiring a legal basis to be processed by the service/app's creator.

Is the AI model used in the EU as a part of an AI-based app/service (see sections B1 and B2 of the form, but also consider situations where the model could be used by someone else for providing a service)?
This question focuses on a special case where the trained model is actually applied by someone else. Even though this special case was not discussed in the deployment models above, we recommend you to consider this possibility. Applications of this type also fall within the scope of the European Union AI Act.

8.3.4 How to go even further?

The first step in compliance with data protection requirements is to establish the system stakeholders in terms of the GDPR, followed by mapping the data flows between them. The result of this work can be a table where lines represent all stakeholders related to the operation of the AI system and the columns, the data elements that they process.

Mark each cell of the table if the specific stakeholder processes the specific data element in the sense of data protection law (e.g., collection, storage, and deletion). If the system employs privacy enhancing technologies, the cell can also show the level to which the specific data element has been made more difficult to personally identify for the specific stakeholder.

Artificial intelligence law is in rapid development at the moment, making it infeasible to provide quick and specific recommendations for the years to come. It is important to monitor the evolution of AI regulation in the target markets of the developed service.

8.4 Evaluate threats to users, society, and environment

8.4.1 DM1: system using AI as a service

Impact analysis 1.1: For each end user, see the responses provided (sections A1–An of the form) and the general description of the system (sections B1 and B2 of the form) and write down the kind of decisions which the user could make based on the responses received from the AI system, and whether any of these decisions may have a direct impact of another user or a third party or could direct them to take any decisions or steps.

It is important to focus here on the users of the system on both the client and service provider sides. A client of the system could get information from the AI's output that they will use to make a decision impacting their or someone else's life. Analysing such thought processes will facilitate awareness of the AI system's impact on human behaviour and, therefore, the society.

A separate important step is to also consider here as end-users information systems making automated decisions using AI, and their impact. For example, if a service or app uses AI-based automated decisions for approving allowances, loans, or rentals, the AI system will have a direct impact on the lives of third persons which the creator of the service needs to be aware of.

Write down all actions identified through this thought experiment that the AI service's output can direct an individual to. Figure 24 provides an example of a worksheet to use for the analysis. Expand the worksheet with new cells as required.

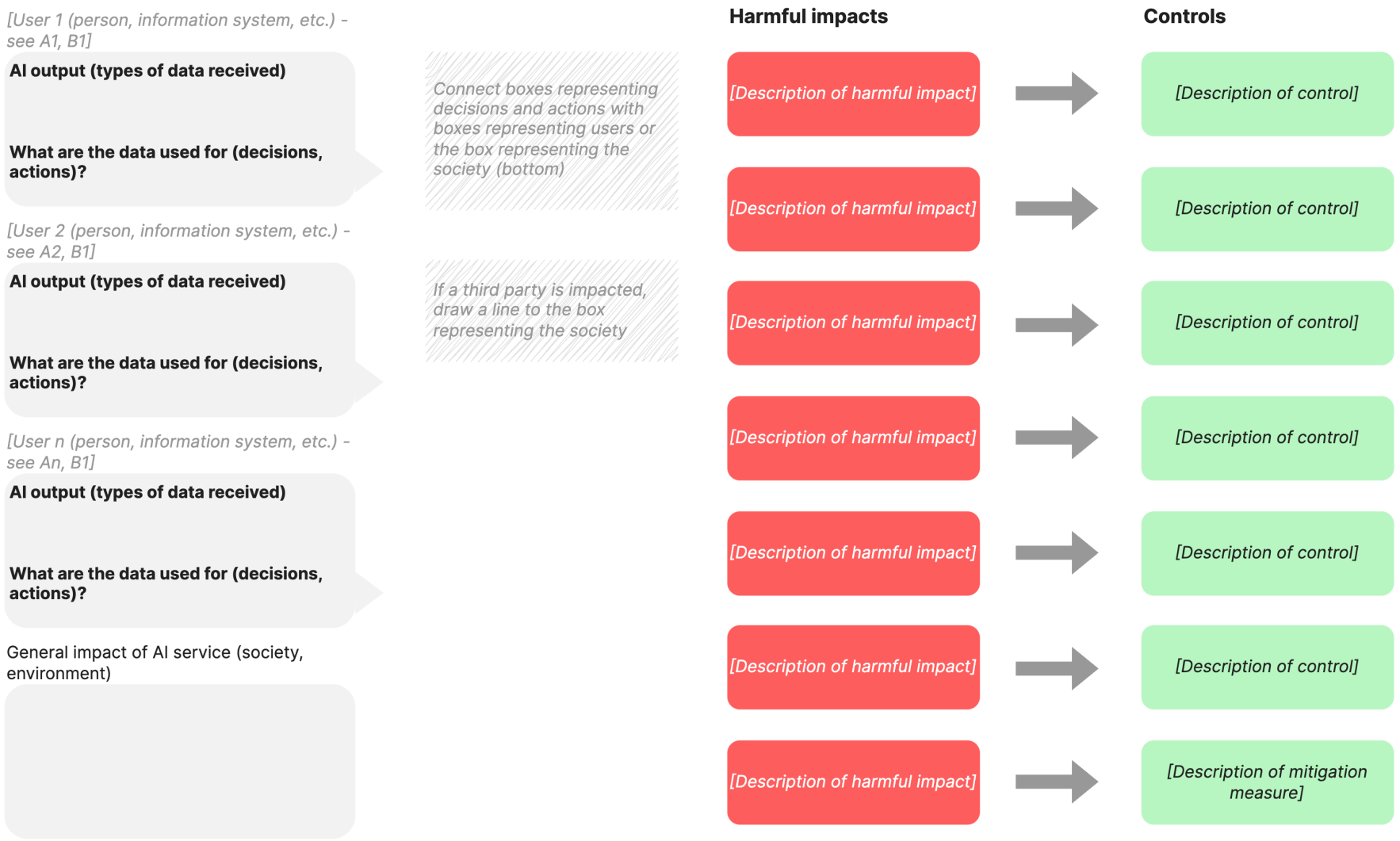


Figure 24. AI societal impact analysis worksheet template

Impact analysis 1.2: For each such action, evaluate whether it could have a negative impact on the individual or the society. Some of such harmful impacts could include the following.

1. Use of the AI service impacts the basic rights of a person or a group of persons.
2. A decision made based on the AI service's output discriminates against a specific social group based on some of their traits.
3. A decision made based on the AI service's output will lead a member of the society to cause harm to themselves (e.g., inaccurate health advice, inaccurate educational advice, inaccurate investment advice).
4. A decision made based on the AI service's output will lead a member of the society to cause harm to another person (e.g., inaccurate diagnosis, inaccurate treatment advice, inaccurate suspicion in an offence, inaccurate assessment of skills or capabilities).

Impact analysis 1.3: Collect all scenarios involving decisions leading to potential harmful impact. Analyse the extent to which an AI model operation service provider assumes responsibility and proposes countermeasures to these. Identify scenarios in which the AI model operation service provider's countermeasures and responsibility are insufficient to mitigate the risk. Assess whether the business logic of the services can be changed or scaled down, or add suitable countermeasures to the system (e.g., transparency, addition of a human supervision mechanism, stronger data management, additional controls in business logic, awareness campaigns, training programs).

Impact analysis 1.4: Evaluate the general impact of the created system on the natural and living environment (without focusing on specific groups of individuals). Evaluate whether the creation of the system has an impact on the environment – whether it impacts the use of energy or natural resources, e.g., through supporting wasteful or polluting behaviour. If the impact is harmful, change or scale down system functionality or implement necessary harm prevention or reduction measures.

8.4.2 DM2: system using an externally-trained AI model

Complete all steps listed in Section 8.4.1, as well as the following steps.

Impact analysis 2.1: Familiarise yourself with the model provider's service conditions, description of the model, and safety information (see sections C1 and C2 of the form). Identify the potential harmful impacts of the use of the model.

If you see that risks can be reduced in a technologically adequate, legally sound, and ethical manner via additional AI model training, then add additional training or fine-tuning of the AI model to the planned activities.

8.4.3 DM3: system using an AI model trained in-house

Complete all steps listed in Sections 8.4.1 and 8.4.2, as well as the following steps.

Impact analysis 3.1: Evaluate the balance and lack of biases in the AI model's training dataset. Is it sufficiently representative to prevent discrimination in the application of the model? If not, find legal and ethical ways to add more training datasets.

Impact analysis 3.2: Evaluate the know-how and technological solutions required for training the AI model. Is the training of a high-quality model possible and affordable in-house? If there are doubts regarding its affordability, you should consider using an externally-trained model rather than training one in-house.

8.4.4 How to go even further?

Guidelines developed for this purpose can be used in impact analysis. We recommend using the EU AI HLEG self-assessment methodology [90], and for LLM applications, the OWASP Foundation's LLM AI Cybersecurity & Governance checklist [193].

It can be expected that EU AI regulations will classify some artificial intelligence systems as high-risk systems and establish additional obligations for relevant service providers. Follow the developments of the regulation to comply with these.

8.5 Perform risk treatment and select controls

8.5.1 Key risks of AI systems

This section will provide instructions on what should be the primary focus of risk treatment. These should not be considered exhaustive security recommendations. Each organisation is different and may require a more in-depth approach. If the organisation providing an AI service has risk treatment practices in place then these practices should be followed and the instructions here used as an initial guideline.

Tables 4, 5, and 6 list the key risks of, respectively, service provision, running AI models, and training AI models. We assess their impact as high and the service provider needs to find ways to treat them. Naturally, your risk assessment process can also identify additional risks not included in this table.

All three tables list the key risks of AI application by stages (composition of input in the app or service, running the model, training the model) and deployment models.

8.5.2 Recommendations for cybersecurity controls

Figure 25 presents a selection of measures from the Estonian E-ITS information security standard suitable for securing AI systems. They are also classified in the figure by the context of the system.

The majority of the measures are applicable to the service provider's organisation, software development, and cloud service use and outsourcing practices. For some of the measures, we have highlighted their importance to the machine or user interfaces created for users. We have also highlighted the significance of certain practices to communication with AI API or model providers.

The cloud service and outsourcing measures are presented as optional – if the service provider does not use cloud-based data processing or outsource anything, their implementation may not be relevant to the created AI app or service.

8.5.3 Recommendations for AI controls

We recommend implementing the controls from Section 6.2 to improve the safety of AI-based services. These help improve the quality of the AI system and avoid risks arising from specific AI technologies.

8.5.4 How to go even further?

We recommend completely implementing any standardised information security or cybersecurity management system or risk assessment methodology. Specific references are found in Section 5.1. Implementing the E-ITS or ISO/IEC 27001 standards to an appropriate level will greatly support the development of the security of AI systems. The work put into implementing this quick-reference guide will not be wasted and will support the implementation of the chosen standards in the organisation.

Table 4. Key risks of running an AI-based service based on the identified deployment model

Category	DM1: Service using an AI API	DM2: Service using an external AI model	DM3: AI service using an in-house model
Cybersecurity	Availability of the AI API does not meet service requirements	Common risks	Common risks
Legal	Service provider lacks legal basis for processing input or output data or submitting the data to the API	Service provider lacks legal basis for processing input or output data or submitting the data to the API	Service provider lacks legal basis for processing input or output data or submitting the data to the API
AI safety	AI API outputs have harmful impact	See risks of running models in Table 5	See risks of running models in Table 5

Table 5. Key risks of running an AI model based on the identified deployment model

Category	DM1: Service using an AI API	DM2: Service using an external AI model	DM3: AI service using an in-house model
Cybersecurity	Service provider does not run the model themselves	Infrastructure used for running the AI model lacks sufficient performance (availability risk) AI model provider does not provide improvements and updates for the model	Infrastructure used for running AI model lacks sufficient performance (availability risk)
Legal	Service provider does not run the model themselves	AI model or its outputs include data that the service provider is not authorised to process Service provider is not authorised to process data used for improving the model	See risks of model training in Table 6
AI safety	Service provider does not run the model themselves	AI model outputs have a harmful impact Data and tools used for improving the model reduce the model's quality	See risks of model training in Table 6

Table 6. Risks of training AI models based on the identified deployment model

Category	DM1: Service using an AI API	DM2: Service using an external AI model	DM3: AI service using an in-house model
Cybersecurity	Service provider does not train the model themselves	Service provider does not train the model themselves	AI model training infrastructure lacks sufficient performance (availability risk)
Legal	Service provider does not train the model themselves	Service provider does not train the model themselves	Service provider lacks authorisation for processing data used for training the model
AI safety	Service provider does not train the model themselves	Service provider does not train the model themselves	AI model outputs have a harmful impact Data and tools used for training the model reduce the quality of the model

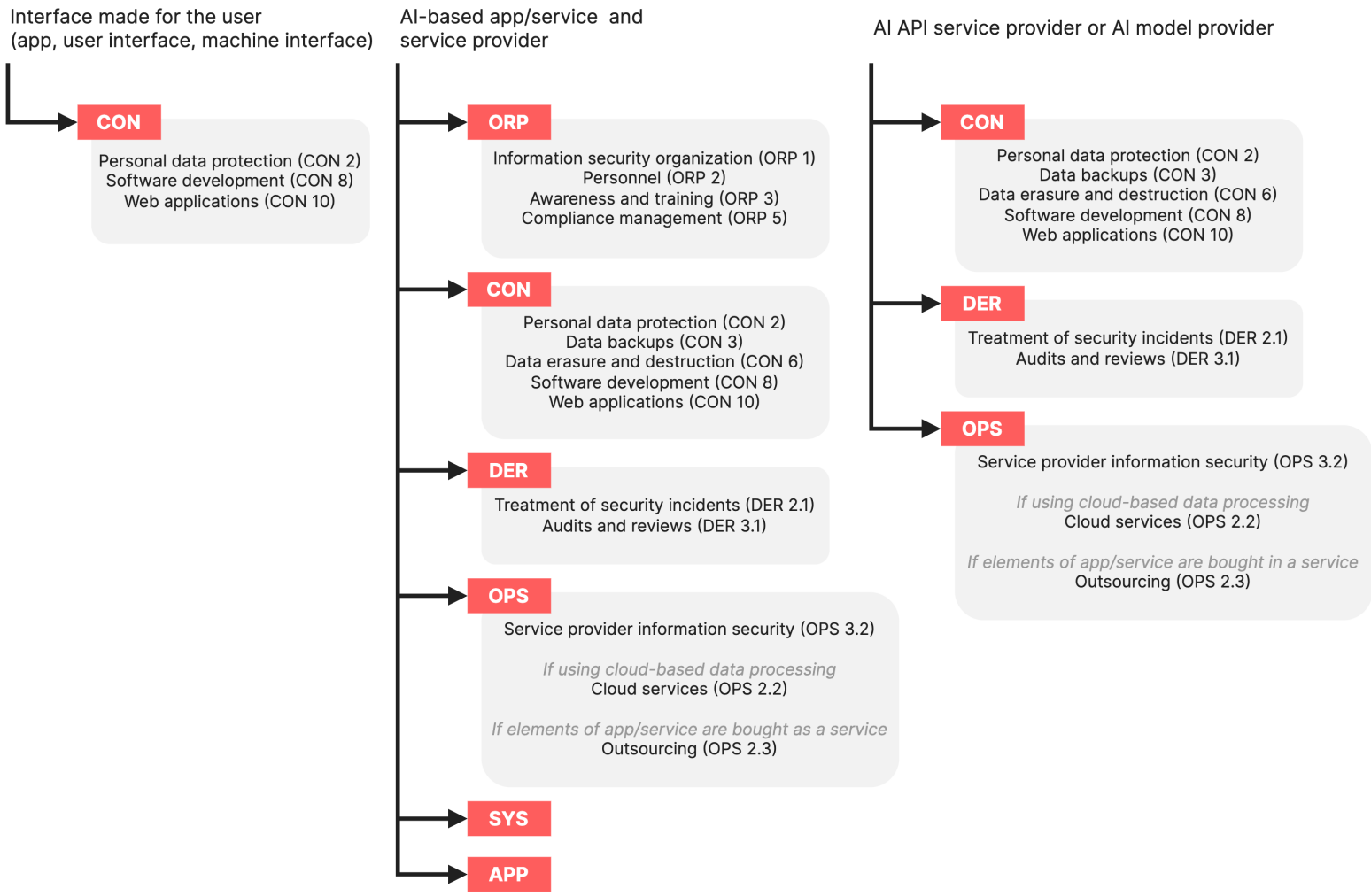


Figure 25. E-ITS modules recommended for AI systems and contexts of their implementation

8.6 AI system in a single slide

The application of artificial intelligence may lead to situations where an overview of the created system must be presented in a single image (e.g., presentation slide to the organisation's management). The figures below present templates for describing the structure of the system. Each figure presents a template for a specific deployment model (Figure 26 for DM1, Figure 27 for DM2, and Figure 28 for DM3).

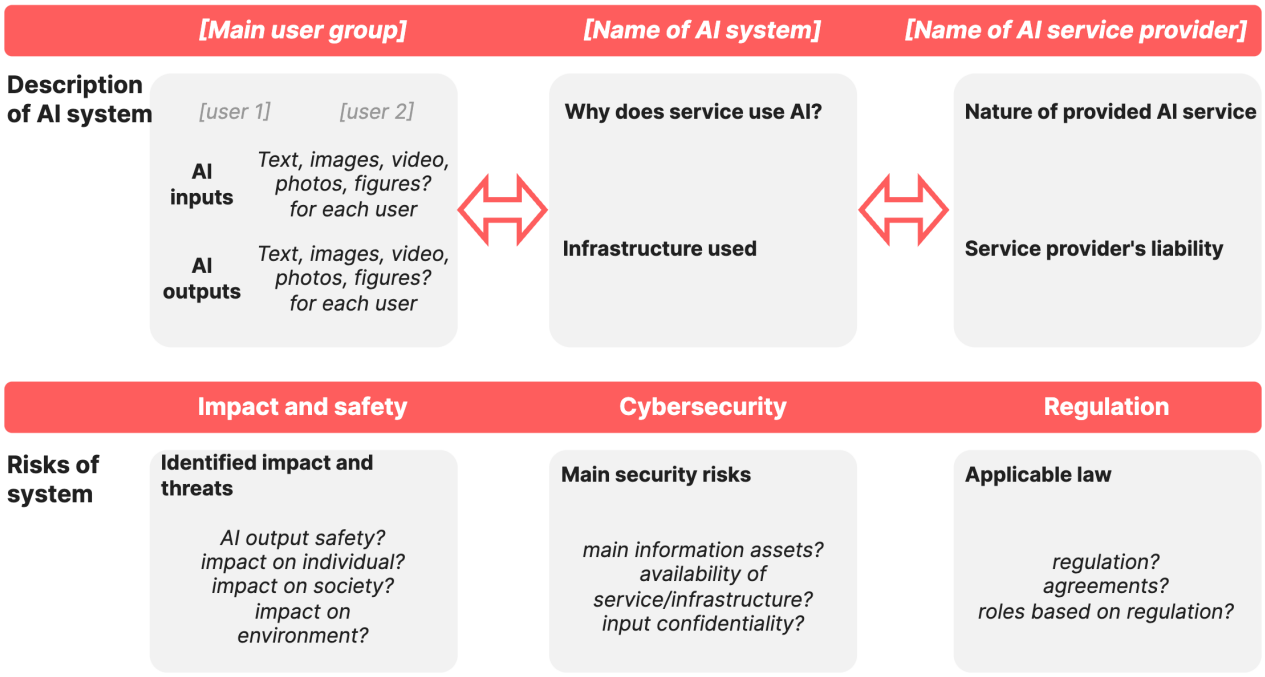


Figure 26. Template for presenting an app or service using deployment model DM1

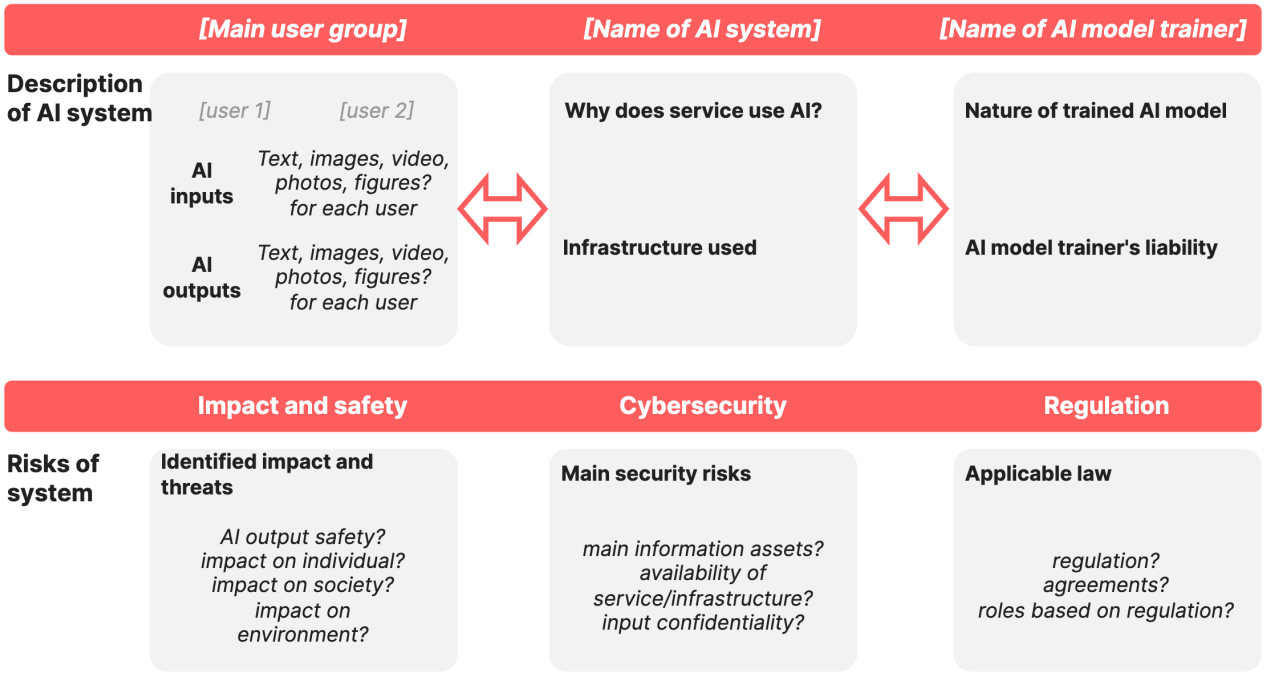


Figure 27. Template for presenting an app or service using deployment model DM2

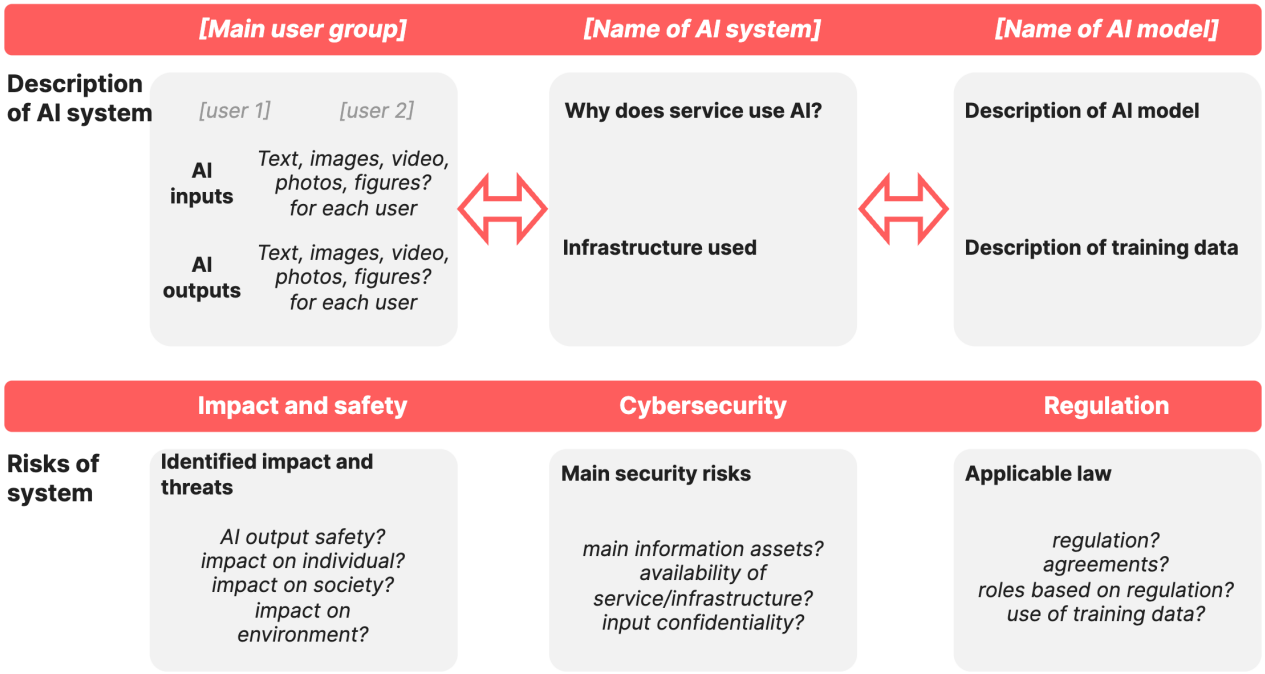


Figure 28. Template for presenting an app or service using deployment model DM3

Bibliography

[1] Kai Wang et al. *Neural Network Diffusion*. 2024. arXiv: 2402.13144 [cs.LG].

[2] Yutao Sun et al. *Retentive Network: A Successor to Transformer for Large Language Models*. 2023. arXiv: 2307.08621 [cs.CL].

[3] Bo Peng et al. *RWKV: Reinventing RNNs for the Transformer Era*. 2023. arXiv: 2305.13048 [cs.CL].

[4] Albert Gu and Tri Dao. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. 2023. arXiv: 2312.00752 [cs.LG].

[5] Meredith Ringel Morris et al. *Levels of AGI: Operationalizing Progress on the Path to AGI*. 2023. arXiv: 2311.02462 [cs.AI].

[6] Blaise Agüera y Arcas and Peter Norvig. "Artificial General Intelligence Is Already Here". In: *Noema Magazine* (Oct. 2023). URL: https://www.noemamag.com/artificial-general-intelligence-is-already-here/.

[7] Rafael Rafailov et al. *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. 2023. arXiv: 2305.18290 [cs.LG].

[8] Mohammad Gheshlaghi Azar et al. *A General Theoretical Paradigm to Understand Learning from Human Preferences*. 2023. arXiv: 2310.12036 [cs.AI].

[9] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[10] Tero Karras, Samuli Laine, and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2019. arXiv: 1812.04948 [cs.NE].

[11] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV].

[12] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV].

[13] Wenhui Wang et al. *Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks*. 2022. arXiv: 2208.10442 [cs.CV].

[14] Wenhai Wang et al. *InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions*. 2023. arXiv: 2211.05778 [cs.CV].

[15] Chengyi Wang et al. *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers*. arXiv. Jan. 2023. URL: https://www.microsoft.com/en-us/research/publication/neural-codec-language-models-are-zero-shot-text-to-speech-synthesizers/.

[16] Matthew Le et al. *Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale*. 2023. arXiv: 2306.15687 [eess.AS].

[17] Alexei Baevski et al. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: 2006.11477 [cs.CL].

[18] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[19] Aakanksha Chowdhery et al. *PaLM: Scaling Language Modeling with Pathways*. 2022. arXiv: 2204.02311 [cs.CL].

[20] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[21] European Union. "Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive)". In: *OJEU L 333* 65 (Dec. 27, 2022), pp. 80–152.

[22] Waddah Saeed and Christian Omlin. "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities". In: *Knowledge-Based Systems* 263 (2023), p. 110273. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2023.110273. URL: https://www.sciencedirect.com/science/article/pii/S095070512300023

[23] Luca Nannini, Agathe Balayn, and Adam Leon Smith. "Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1198–1212. DOI: 10.1145/3593013.3594074. URL: https://doi.org/10.1145/3593013.3594074.

[24] Nagadivya Balasubramaniam et al. "Transparency and explainability of AI systems: From ethical guidelines to requirements". In: *Information and Software Technology* 159 (2023), p. 107197. ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2023.107197. URL: https://www.sciencedirect.com/science/article/pii/S095058492300051

[25] OECD. *OECD Legal Instruments. Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. Adopted on: 22.05.2019. Amended on: 08.11.2023*. Nov. 2023. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[26] Feiyu Xu et al. "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges". In: Sept. 2019, pp. 563–574. ISBN: 978-3-030-32235-9. DOI: 10.1007/978-3-030-32236-6_51.

[27] Christoph Molnar. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Paperback. Feb. 2022.

[28] Jason Wei et al. *Emergent Abilities of Large Language Models*. 2022. arXiv: 2206.07682 [cs.CL].

[29] Sheng Lu et al. *Are Emergent Abilities in Large Language Models just In-Context Learning?* 2023. arXiv: 2309.01809 [cs.CL].

[30] Rajat Raina, Anand Madhavan, and Andrew Ng. "Large-scale deep unsupervised learning using graphics processors". In: vol. 382. June 2009, p. 110. DOI: 10.1145/1553374.1553486.

[31] Amir Gholami et al. *A Survey of Quantization Methods for Efficient Neural Network Inference*. 2021. arXiv: 2103.13630 [cs.CV].

[32] Albert Tseng et al. *Quip#: Quip with Lattice Codebooks*. Dec. 2023.

[33] Uriel Singer et al. *Make-A-Video: Text-to-Video Generation without Text-Video Data*. 2022. arXiv: 2209.14792 [cs.CV].

[34] Levon Khachatryan et al. *Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators*. 2023. arXiv: 2303.13439 [cs.CV].

[35] Andreas Blattmann et al. *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets*. 2023. arXiv: 2311.15127 [cs.CV].

[36] Joon Sung Park et al. *Generative Agents: Interactive Simulacra of Human Behavior*. 2023. arXiv: 2304.03442 [cs.HC].

[37] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL].

[38] Yuanzhi Li et al. "Textbooks Are All You Need II: phi-1.5 technical report". Sept. 2023. URL: https://www.microsoft.com/en-us/research/publication/textbooks-are-all-you-need-ii-phi-1-5-technical-report/.

[39] Google DeepMind Gemma Team. *Gemma: Open Models Based on Gemini Research and Technology*. 2024. URL: https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf.

[40] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL].

[41] Jinze Bai et al. *Qwen Technical Report*. 2023. arXiv: 2309.16609 [cs.CL].

[42] Ben Sorscher et al. *Beyond neural scaling laws: beating power law scaling via data pruning*. 2023. arXiv: 2206.14486 [cs.LG].

[43] Europol. *Facing reality? Law enforcement and the challenge of deepfakes. An Observatory Report from the Europol Innovation Lab*. 2022. DOI: 10.2813/158794|QL-02-24-129-EN-N. URL: https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf.

[44] Partha Pratim Ray. "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope". In: *Internet of Things and Cyber-Physical Systems* 3 (2023), pp. 121–154. ISSN: 2667-3452. DOI: https://doi.org/10.1016/j.iotcps.2023.04.003. URL: https://www.sciencedirect.com/science/article/pii/S266734522300024X.

[45] Independent High-Level Expert Group on AI set up by the European Commission in June 2018. *Ethics guidelines for trustworthy AI*. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[46] Forbes. *24 Top AI Statistics And Trends In 2023*. Apr. 2023. URL: https://www.forbes.com/advisor/business/ai-statistics/%5C#sources_section.

[47] Statistikaamet [Statistics Estonia]. *Tehisintellekti tehnoloogiate kasutamine ettevõtetes on tõusutrendis [Use of AI technologies by businesses is on the rise]*. Sept. 2023. URL: https://www.stat.ee/et/uudised/tehisintellekti-tehnoloogiate-kasutamine-ettevotetes-tousutrendis.

[48] Ada Lovelace Institute and The Alan Turing Institute. *How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain*. June 2023. URL: https://www.turing.ac.uk/sites/default/files/2023-06/how%5C_do%5C_people%5C_feel%5C_about%5C_ai%5C_-%5C_ada%5C_turing.pdf.

[49] Euractiv. *EU top court's ruling spells trouble for scoring algorithms*. Dec. 2023. URL: https://www.euractiv.com/section/data-privacy/news/eu-top-courts-ruling-spill-trouble-for-scoring-algorithms/.

[50] Politico. *Dutch scandal serves as a warning for Europe over risks of using algorithms*. Mar. 2022. URL: https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

[51] Maxi Scherer. "'Artificial Intelligence and Legal Decision-Making: The Wide Open?'" In: *Journal of International Arbitration* 36 (5 2019), pp. 539–573. URL: https://kluwerlawonline.com/journalarticle/Journal+of+International+Arbitration/36.5/JOIA2019028.

[52] Maja Brkan. "Opinions. Artificial Intelligence and Judicial Decision-Making". In: *European Data Protection Law Review* 9.3 (2023). DOI: 10.21552/edpl/2023/3/5. URL: https://doi.org/10.21552/edpl/2023/3/5.

[53] Matthew Dahl et al. *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*. 2024. eprint: 2401.01301.

[54] Matthew Dahl et al. *Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive*. Jan. 2024. URL: https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive.

[55] Shaun Lim. "Judicial decision-making and explainable artificial intelligence. A Reckoning from First Principles". In: *SAcLJ* 280 (2021). URL: https://law.nus.edu.sg/trail/wp-content/uploads/sites/9/2022/03/9777_09.-Shaun-Lim-Judicial-Decision-Making-and-Explainable-AI.pdf.

[56] European Commission. *Ethics guidelines for trustworthy AI*. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[57] "Proposal for a regulation of the European Oarliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts". In: (). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[58] European Parliament. *EU AI Act: first regulation on artificial intelligence*. June 2023. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[59] European Union. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)". In: *Official Journal L119* 59 (May 4, 2016), pp. 1–88.

[60] The White House. *Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*. Oct. 2023. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[61] UK Parliament. *Parliamentary Bills*. 2023. URL: https://bills.parliament.uk/bills/3464.

[62] UK Parliament. *Artificial intelligence and employment law*. URL: https://commonslibrary.parliament.uk/research-briefings/cbp-9817/.

[63] Official Website of the International Trade Administration. *UK AI regulations 2023*. 2023. URL: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach.

[64] Australian Government. *Supporting responsible AI: discussion paper*. June 2023. URL: https://consult.industry.gov.au/supporting-responsible-ai.

[65] Australian Government. *Positioning Australia as a leader in digital economy regulation. Automated Decision Making and AI Regulation. Issues Paper*. Mar. 2023. URL: https://storage.googleapis.com/converlens-au-industry/industry/p/prj211c4e81fb27d147epublic%5C_assets/automated-decision-making-ai-regulation-issues-paper.pdf.

[66] Dentons. *Australian Government requests public feedback on regulating 'Safe and Responsible AI'*. June 2023. URL: https://www.dentons.com/en/insights/alerts/2023/june/6/australian-government-requests-public-feedback-on-regulating-safe.

[67] eSafety Commissioner. *Tech Trends Position Statement. Generative AI*. URL: https://www.esafety.gov.au/sites/default/files/2023-08/Generative%5C%20AI%5C%20-%5C%20Position%5C%20Statement%5C%20-%5C%20August%5C%202023%5C%20.pdf.

[68] *Australia to require AI-made child abuse material be removed from search results*. Sept. 2023. URL: https://www.reuters.com/technology/australia-require-ai-made-child-abuse-material-be-removed-search-results-2023-09-08/.

[69] The Guardian. *Search engines required to stamp out AI-generated images of child abuse under Australia's new code*. June 2023. URL: https://www.theguardian.com/technology/2023/sep/08/search-engines-required-to-stamp-out-ai-generated-images-of-child-abuse-under-australias-new-code.

[70] Government of Canada. *Artificial Intelligence and Data Act*. Sept. 2023. URL: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act.

[71] *Government of Canada launches consultation on the implications of generative artificial intelligence for copyright*. Oct. 2023. URL: https://www.canada.ca/en/innovation-science-economic-development/news/2023/10/government-of-canada-launches-consultation-on-the-implications-of-generative-artificial-intelligence-for-copyright.html.

[72] Runway Strategies. *Global AI Regulation Tracker*. Dec. 2023. URL: https://www.theguardian.com/technology/2023/sep/08/search-engines-required-to-stamp-out-ai-generated-images-of-child-abuse-under-australias-new-code.

[73] ISO. *ISO/IEC 22989:2022*. 2022. URL: https://www.iso.org/standard/74296.html.

[74] International Electrotechnical Commission. *Two new foundational standards for artificial intelligence*. July 2022. URL: https://www.iec.ch/blog/two-new-foundational-standards-artificial-intelligence.

[75] ISO. *ISO/IEC 23053:2022*. July 2022. URL: https://www.iso.org/standard/74438.html.

[76] ISO. *ISO/IEC FDIS 5259-1. Artificial intelligence. Data quality for analytics and machine learning (ML). Part 1: Overview, terminology, and examples*. URL: https://www.iso.org/standard/81088.html.

[77] ISO. *ISO/IEC DIS 5259-2. Artificial intelligence. Data quality for analytics and machine learning (ML). Part 2: Data quality measures*. URL: https://www.iso.org/standard/81860.html.

[78] ISO. *ISO/IEC TS 4213:2022. Information technology. Artificial intelligence. Assessment of machine learning classification performance*. URL: https://www.iso.org/standard/79799.html.

[79] ISO. *BS 30440:2023. Validation framework for the use of artificial intelligence (AI) within healthcare. Specification. Current. Published: 31 Jul 2023*. July 2023. URL: https://knowledge.bsigroup.com/products/validation-framework-for-the-use-of-artificial-intelligence-ai-within-healthcare-specification.

[80] IEEE. *IEEE Standards Association. Ethically aligned design, Version 1, Translations and reports*. URL: https://standards.ieee.org/industry-connections/ec/ead-v1/.

[81] Google. *Google AI. Responsibility: Our principles*. URL: https://ai.google/responsibility/principles/.

[82] Google. *Google AI. Responsibility: Responsible AI practicess*. URL: https://ai.google/responsibility/responsible-ai-practices/.

[83] Microsoft. *Microsoft Responsible AI Standard, v2. General Requirements. For external release. June 2022*. July 2022. URL: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf.

[84] OECD. *Multilayer Framework for Good Cybersecurity Practices for AI*. June 2023. URL: https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai.

[85] European Commission. *Ethics guidelines for trustworthy AI*. Apr. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[86] Independent High-Level Expert Group on AI set up by the European Commission in June 2018. *Ethics guidelines for trustworthy AI*. Apr. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[87] A. Vassilev et al. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Artificial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2023*. 2024. DOI: https://doi.org/10.6028/NIST.AI.100-2e2023.

[88] European Commission. *High-level expert group on artificial intelligence*. URL: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai.

[89] High-level expert group on artificial intelligence. *Policy and investment recommendations for trustworthy Artificial Intelligence*. June 2019. URL: https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

[90] High-level expert group on artificial intelligence. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. July 2020. URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[91] European AI Alliance. *Welcome to the ALTAI portal!* URL: https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal.

[92] High-level expert group on artificial intelligence. *AI HLEG - Sectoral Considerations on Policy and Investment Recommendations for Trustworthy AI*. July 2020. URL: https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai.

[93] European Commission. *Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee. Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics. Brussels, 19.2.2020, COM(2020) 64 final*. Feb. 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%5C%3A52020DC0064.

[94] European Commission. *White Paper On Artificial Intelligence - A European approach to excellence and trust. Brussels, 19.2.2020, COM(2020) 65 final*. Feb. 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%5C%3A52020DC0065&qid=1664351767552.

[95] European Parliament. "haping the digital transformation: EU strategy explained". In: (). URL: https://www.europarl.europa.eu/topics/en/article/20210414STO02010/shaping-the-digital-transformation-eu-strategy-explained.

[96] European Commission. "Commission welcomes political agreement on Artificial Intelligence Act". In: (Dec. 2023). URL: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence.

[97] European Parliament. "Legislative Train Schedule. Artificial intelligence act". In: (). URL: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence.

[98] Luca Bertuzzi. *EU countries give crucial nod to first-of-a-kind Artificial Intelligence law*. Feb. 2024. URL: https://www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/.

[99] *Analysis of the final compromise text with a view to agreement. Interinstitutional File: 2021/0106(COD). No. Cion doc.: 8115/21*. Jan. 2024. URL: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

[100] European Commission. *European AI Office*. 2024. URL: https://digital-strategy.ec.europa.eu/en/policies/ai-office.

[101] European Commission. *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). Brussels, 28.9.2022. COM(2022) 496 final. 2022/0303(COD)*. Sept. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0496&from=EN.

[102] European Union. "Regulation (EU) 2023/988 on general product safety, amending Regulation (EU) No 1025/2012 and Directive (EU) 2020/1828, and repealing Directive 2001/95/EC and Directive 87/357/EEC". In: *OJEU L 135* 66 (May 23, 2023), pp. 1–51.

[103] European Commission. *Proposal for a Directive of the European Parliament and of the Council on liability for defective products. Brussels, 28.9.2022. COM(2022) 495 final. 2022/0302(COD)*. Sept. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0495.

[104] Jude Ume. "What AI Means for Intellectual Property". In: *ITNOW* 65.2 (May 2023), pp. 44–45. ISSN: 1746-5702. DOI: 10.1093/combul/bwad059. eprint: https://academic.oup.com/itnow/article-pdf/65/2/44/50367250/bwad059.pdf. URL: https://doi.org/10.1093/combul/bwad059.

[105] Peter Georg Picht and Florent Thouvenin. "AI and IP: Theory to Policy and Back Again – Policy and Research Recommendations at the Intersection of Artificial Intelligence and Intellectual Property". In: *IIC - International Review of Intellectual Property and Competition Law* 54.6 (July 1, 2023), pp. 916–940. DOI: 10.1007/s40319-023-01344-5. URL: https://doi.org/10.1007/s40319-023-01344-5.

[106] Winston Cho. *AI Companies Take Hit as Judge Says Artists Have "Public Interest" In Pursuing Lawsuits A federal judge rejected arguments from Stability AI, Midjourney and DeviantArt that the suit is intended to suppress its free speech*. Feb. 2024. URL: https://www.hollywoodreporter.com/business/business-news/artist-lawsuit-ai-midjourney-art-1235821096/.

[107] Bobby Allyn. *'New York Times' considers legal action against OpenAI as copyright tensions swirl*. Aug. 2023. URL: https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl.

[108] Will Bedingfield. *The Inventor Behind a Rush of AI Copyright Suits Is Trying to Show His Bot Is Sentient. Stephen Thaler's series of high-profile copyright cases has made headlines worldwide. He's done it to demonstrate his AI is capable of independent thought*. Sept. 2023. URL: https://www.wired.com/story/the-inventor-behind-a-rush-of-ai-copyright-suits-is-trying-to-show-his-bot-is-sentient/.

[109] European Parliament. *World Intellectual Property Organization. WIPO CONVERSATION ON INTELLECTUAL PROPERTY (IP) AND ARTIFICIAL INTELLIGENCE (AI). WIPO/IP/AI/3/GE/20/INF/5*. Jan. 2021. URL: https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_3_ge_20/wipo_ip_ai_3_ge_20_inf_5.pdf.

[110] S. Suganya and E. Prema. "Human centric intellectual property rights and legal personality of artificial intelligence". In: *International Journal of Intellectual Property Management* 13.3-4 (2023), pp. 252–267. DOI: 10.1504/IJIPM.2023.134051. eprint: https://www.inderscienceonline.com/doi/pdf/10.1504/IJIPM.2023.134051. URL: https://www.inderscienceonline.com/doi/abs/10.1504/IJIPM.2023.134051.

[111] Atif Aziz. "Artificial Intelligence Produced Original Work: A New Approach to Copyright Protection and Ownership". In: *European Journal of Artificial Intelligence and Machine Learning* 2.2 (Mar. 2023), pp. 9–16. DOI: 10.24018/ejai.2023.2.2.15. URL: https://www.ej-ai.org/index.php/ejai/article/view/15.

[112] Mauritz Kop. "TAI & Intellectual Property: Towards an Articulated Public Domain". In: *University of Texas School of Law, Texas Intellectual Property Law Journal (TIPLJ)* 28.1 (June 2019), pp. 44–45. ISSN: 1746-5702. DOI: http://dx.doi.org/10.2139/ssrn.3409715. eprint: https://academic.oup.com/itnow/article-pdf/65/2/44/50367250/bwad059.pdf.

[113] Winston Cho. *AI-Created Art Isn't Copyrightable, Judge Says in Ruling That Could Give Hollywood Studios Pause. A federal judge on Friday upheld a finding from the U.S. Copyright Office that a piece of art created by AI is not open to protection*. Aug. 2023. URL: https://www.hollywoodreporter.com/business/business-news/ai-works-not-copyrightable-studios-1235570316/.

[114] European Parliament. *European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI))*. Oct. 2020. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_EN.html.

[115] European Union. "Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act)". In: *OJEU L 151* 62 (June 7, 2019), pp. 15–69.

[116] Centre for Cyber security Belgium. *The NIS 2 Directive: What does it mean for my organization?* 2024. URL: https://ccb.belgium.be/en/nis-2-directive-what-does-it-mean-my-organization.

[117] European Commission. *Proposal for a Regulation of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020. Brussels, 15.9.2022. COM(2022) 454 final. 2022/0272(COD)*. Sept. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0454.

[118] EU Presidency. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Analysis of the final compromise text with a view to agreement. Brussels, 26 January 2024. Interinstitutional File: 2021/0106(COD). No. Cion doc.: 8115/21*. Jan. 2024. URL: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

[119] European Commission. *Proposal for a Regulation of the European Parliament and of the Council laying down additional procedural rules relating to the enforcement of Regulation (EU) 2016/679. Brussels, 4.7.2023. COM(2023) 348 final. 2023/0202(COD)*. July 2023. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52023PC0348.

[120] European Union. "Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA". In: *OJEU L 119* 59 (May 4, 2016), pp. 89–131.

[121] European Union. "Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC". In: *OJEU L 295* 61 (Nov. 21, 2018), pp. 39–98.

[122] N. AllahRakha. *AI and the Law: Unraveling the Complexities of Regulatory Frameworks in Europe*. Aug. 2023. URL: https://irshadjournals.com/index.php/ibys/article/view/115/102.

[123] *AI: ensuring GDPR compliance*. Sept. 2022. URL: https://www.cnil.fr/en/ai-ensuring-gdpr-compliance.

[124] Information Commissioner's Office. *Guidance on AI and data protection*. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/.

[125] Information Commissioner's Office. *Information Commissioner's Office launches consultation series on generative AI*. Jan. 2024. URL: https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/01/information-commissioner-s-office-launches-consultation-series-on-generative-ai/.

[126] Information Commissioner's Office. *ICO consultation series on generative AI and data protection*. Jan. 2024. URL: https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-series-on-generative-ai-and-data-protection/.

[127] CNIL. *AI: ensuring GDPR compliance*. Sept. 2021. URL: https://www.cnil.fr/en/ai-ensuring-gdpr-compliance.

[128] Federal Trade Commission. *FTC Takes Action Against Company Formerly Known as Weight Watchers for Illegally Collecting Kids' Sensitive Health Data*. Mar. 2022. URL: https://www.ftc.gov/news-events/news/press-releases/2022/03/ftc-takes-action-against-company-formerly-known-weight-watchers-illegally-collecting-kids-sensitive.

[129] CNIL. *Artificial intelligence: CNIL unveils its first answers for innovative and privacy-friendly AI*. Oct. 2023. URL: https://www.cnil.fr/en/artificial-intelligence-cnil-unveils-its-first-answers-innovative-and-privacy-friendly-ai.

[130] Dan Svantesson. *The European Union Artificial Intelligence Act: Potential implications for Australia*. 2022. DOI: https://doi.org/10.1177/1037969X211052339.

[131] Adèle Azzi. "The Challenges Faced by the Extraterritorial Scope of the General Data Protection Regulation". In: *JIPITEC* 9.2 (2018), pp. 126–137. ISSN: 2190-3387. URL: http://nbn-resolving.de/urn:nbn:de:0009-29-47231.

[132] Michal Czerniawski and Dan Svantesson. "Challenges to the extraterritorial enforcement of data privacy law - EU case study". In: Jan. 2024, pp. 127–153.

[133] Federico Fabbrini and Edoardo Celeste. "The Right to Be Forgotten in the Digital Age: The Challenges of Data Protection Beyond Borders". In: *German Law Journal* 21.S1 (2020), pp. 55–65. DOI: 10.1017/glj.2020.14.

[134] Chris Burt. "Clearview denies jurisdiction of French regulator in response to €20M fine". In: *Biometric Update* (Oct. 2022). URL: https://www.biometricupdate.com/202210/clearview-denies-jurisdiction-of-french-regulator-in-response-to-e20m-fine.

[135] IAPP. *Toward a risk-based approach? Challenging the 'zero risk' paradigm of EU DPAs in international data transfers and foreign governments' data access schedule*. Feb. 2024. URL: https://iapp.org/news/a/towards-a-risk-based-approach-challenging-the-zero-risk-paradigm-of-eu-dpas-in-international-data-transfers-and-foreign-governments-data-access/.

[136] European Commission. *Adequacy decisions. How the EU determines if a non-EU country has an adequate level of data protection*. URL: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en.

[137] European Commission. *Commission finds that EU personal data flows can continue with 11 third countries and territories*. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_161.

[138] European Commission. *Adequacy decision for the EU-US Data Privacy Framework*. URL: https://commission.europa.eu/document/fa09cbad-dd7d-4684-ae60-be03fcb0fddf_en.

[139] EDPB. *Recommendations 1/2022 on the Application for Approval and on the elements and principles to be found in Controller Binding Corporate Rules (Art. 47 GDPR). Adopted on 20 June 2023*. 2023. URL: https://edpb.europa.eu/system/files/2023-06/edpb_recommendations_20221_bcr-c_v2_en.pdf.

[140] Andmekaitse Inspektsioon [Data Protection Inspectorate]. *Isikuandmete edastamine välisriiki [International transfer of personal data]*. URL: https://www.aki.ee/isikuandmed/andmetootlejale/isikuandmete-edastamine-valisriikikuidas-aru-saada-mi#edastamineameerika.

[141] European Data Protection Board. *Guidelines 05/2021 on the Interplay between the application of Article 3 and the provisions on international transfers as per Chapter V of the GDPR. Version 2.0. Adopted 14 February 2023*. Feb. 2023. URL: https://www.edpb.europa.eu/system/files/2023-02/edpb_guidelines_05-2021_interplay_between_the_application_of_art3-chapter_v_of_the_gdpr_v2_en_0.pdf.

[142] *Risk management — Guidelines*. en. Standard ISO 31000:2018. International Organization for Standardization, 2018. URL: https://www.iso.org/standard/65694.html.

[143] *Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy*. en. Standard NIST SP 800-37 Rev. 2. US National Institute of Standards and Technology, 2018. URL: https://csrc.nist.gov/pubs/sp/800/37/r2/final.

[144] *Information technology — Information security, cybersecurity and privacy protection — Guidance on managing information security risks*. en. Standard ISO/IEC 27005:2022. International Organization for Standardization, 2022. URL: https://www.iso.org/standard/80585.html.

[145] *NIST Cybersecurity Framework 1.1*. en. Standard NIST CSF v. 1.1. US National Institute of Standards and Technology, 2018. URL: https://www.nist.gov/cyberframework/framework.

[146] *Information technology — Artificial intelligence — Guidance on risk management*. en. Standard ISO/IEC 23984:2023. International Organization for Standardization, 2023. URL: https://www.iso.org/standard/77304.html.

[147] Riigi Infosüsteemi Amet [Information System Authority]. *Eesti infoturbestandard (E-ITS) [Estonian Information Security Standard]*. 2023. URL: https://eits.ria.ee/.

[148] Pille Pullonen, Raimundas Matulevičius, and Dan Bogdanov. "PE-BPMN: Privacy-Enhanced Business Process Model and Notation". In: *Business Process Management*. Springer International Publishing, 2017, pp. 40–56. DOI: 10.1007/978-3-319-65000-5_3.

[149] HM Government. *Safety and Security Risks of Generative Artificial Intelligence to 2025*. URL: https://assets.publishing.service.gov.uk/media/653932db80884d0013f71b15/generative-ai-safety-security-risks-2025-annex-b.pdf.

[150] Richard Fang et al. *LLM Agents can Autonomously Hack Websites*. 2024. arXiv: 2402.06664 [cs.CR].

[151] Riigi Infosüsteemi Amet [Information System Authority]. *Eesti infoturbestandardi etalonturbe kataloog [E-ITS Baseline Security Catalogue]*. 2023. URL: https://eits.ria.ee/et/versioon/2023/eits-poohidokumendid/etalonturbe-kataloog.

[152] The New York Times. *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said*. Dec. 2023. URL: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

[153] TechTarget. *AI lawsuits explained: Who's getting sued? Authors, artists and others are filing lawsuits against generative AI companies for using their data in bulk to train AI systems without permission*. Jan. 2024. URL: https://www.techtarget.com/whatis/feature/AI-lawsuits-explained-Whos-getting-sued.

[154] The Fashion Law. *From ChatGPT to Getty v. Stability AI: A Running List of Key AI-Lawsuits*. URL: https://www.thefashionlaw.com/from-chatgpt-to-deepfake-creating-apps-a-running-list-of-key-ai-lawsuits/.

[155] Civil Resolution Tribunal of British Columbia. *Moffatt v. Air Canada, 2024 BCCRT 149 (CanLII)*. Feb. 2024. URL: https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html.

[156] Lei Huang et al. *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. 2023. arXiv: 2311.05232 [cs.CL].

[157] Rusheb Shah et al. *Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation*. 2023. arXiv: 2311.03348 [cs.CL].

[158] Xiangyu Qi et al. *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!* 2023. arXiv: 2310.03693 [cs.CL].

[159] José Luis Ricón. *Set Sail For Fail? On AI risk*. Nintil. Available at https://nintil.com/ai-safety/. Aug. 2022.

[160] Government Office of Science, UK. *Future Risks of Frontier AI*. Tech. rep. Technology & Science Insights and Foresight, Oct. 2023.

[161] Jonas B. Sandbrink. *Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools*. 2023. arXiv: 2306.13952 [cs.CY].

[162] Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. *Emergent autonomous scientific research capabilities of large language models*. 2023. arXiv: 2304.05332 [physics.chem-ph].

[163] OpenAI Research Team. *Building an Early Warning System for LLM-Aided Biological Threat Creation*. OpenAI. URL: https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation (visited on 02/08/2024).

[164] Joseph R. Biden Jr. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Executive Order. 14110. 2023.

[165] T. C. King, N. Aggarwal, M. Taddeo, et al. "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". In: *Sci Eng Ethics* 26 (2020), pp. 89–120. DOI: 10.1007/s11948-018-00081-0.

[166] Brigitta Dresp-Langley. "The weaponization of artificial intelligence: What the public needs to be aware of". In: *Frontiers in Artificial Intelligence* 6 (2023), p. 1154184. DOI: 10.3389/frai.2023.1154184.

[167] Mrinank Sharma et al. *Towards Understanding Sycophancy in Language Models*. 2023. arXiv: 2310.13548 [cs.CL].

[168] Emily Baker-White and Forbes Staff. "Who Is @BasedBeffJezos, The Leader Of The Tech Elite's 'E/Acc' Movement?" In: *Forbes* (Dec. 2023). Külastatud 04.12.2023. URL: https://www.forbes.com/sites/emilybaker-white/2023/12/01/who-is-basedbeffjezos-the-leader-of-effective-accelerationism-eacc/.

[169] Milad Nasr et al. *Scalable Extraction of Training Data from (Production) Language Models*. 2023. arXiv: 2311.17035 [cs.LG].

[170] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML].

[171] Jing Lin et al. *ML Attack Models: Adversarial Attacks and Data Poisoning Attacks*. 2021. arXiv: 2112.02797 [cs.LG].

[172] Jonas Geiping et al. *Coercing LLMs to do and reveal (almost) anything*. 2024. arXiv: 2402.14020 [cs.LG].

[173] Joseph Lucas. "Mitigating Stored Prompt Injection Attacks Against LLM Applications". In: *NVIDIA Technical Blog* (Aug. 2023). URL: https://developer.nvidia.com/blog/mitigating-stored-prompt-injection-attacks-against-llm-applications/.

[174] Kai Greshake et al. *Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. 2023. arXiv: 2302.12173 [cs.CR].

[175] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. "I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences". In: *ACM Computing Surveys* 55.14s (July 2023), pp. 1–41. ISSN: 1557-7341. DOI: 10.1145/3595292. URL: http://dx.doi.org/10.1145/3595292.

[176] Reza Shokri et al. *Membership Inference Attacks against Machine Learning Models*. 2017. arXiv: 1610.05820 [cs.CR].

[177] Boris van Breugel et al. *Membership Inference Attacks against Synthetic Data through Overfitting Detection*. 2023. arXiv: 2302.12580 [cs.LG].

[178] Ngoc-Bao Nguyen et al. *Re-thinking Model Inversion Attacks Against Deep Neural Networks*. 2023. arXiv: 2304.01669 [cs.LG].

[179] Kuan-Chieh Wang et al. *Variational Model Inversion Attacks*. 2022. arXiv: 2201.10787 [cs.LG].

[180] Liam Fowl et al. *Adversarial Examples Make Strong Poisons*. 2021. arXiv: 2106.10807 [cs.LG].

[181] Battista Biggio, Blaine Nelson, and Pavel Laskov. *Poisoning Attacks against Support Vector Machines*. 2013. arXiv: 1206.6389 [cs.LG].

[182] Shawn Shan et al. *Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models*. 2023. arXiv: 2310.13828 [cs.CR].

[183] Micah Goldblum et al. *Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses*. 2021. arXiv: 2012.10544 [cs.LG].

[184] Xinyun Chen et al. *Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning*. 2017. arXiv: 1712.05526 [cs.CR].

[185] Ruixiang Tang et al. *An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks*. 2020. arXiv: 2006.08131 [cs.CR].

[186] Megha Agarwal et al. *LLM Inference Performance Engineering: Best Practices*. https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices. Accessed on 11 Dec 2023. 2023.

[187] Jiaqi Ruan et al. *Applying Large Language Models to Power Systems: Potential Security Threats*. 2023. arXiv: 2311.13361 [cs.AI].

[188] Robin Staab et al. *Beyond Memorization: Violating Privacy Via Inference with Large Language Models*. 2023. arXiv: 2310.07298 [cs.AI].

[189] Cybernetica AS. *Privaatsuskaitse tehnoloogiate kontseptsioon [Privacy enhancing technology concept*. Tech. rep. Majandus- ja Kommunikatsiooniministeerium [Ministry of Economic Affairs and Communications], 2023. URL: https://www.kratid.ee/analuusid-ja-uuringud#pet.

[190] Rouzbeh Behnia et al. "EW-Tune: A Framework for Privately Fine-Tuning Large Language Models with Differential Privacy". In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, Nov. 2022. DOI: 10.1109/icdmw58026.2022.00078. URL: http://dx.doi.org/10.1109/ICDMW58026.2022.00078.

[191] OECD. *Accountability (Principle 1.5)*. URL: https://oecd.ai/en/dashboards/ai-principles/P9.

[192] OECD. *Inclusive growth, sustainable development and well-being (Principle 1.1)*. URL: https://oecd.ai/en/dashboards/ai-principles/P5.

[193] OWASP Foundation. *LLM AI Cybersecurity & Governance Checklist*. 2024. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.pdf.