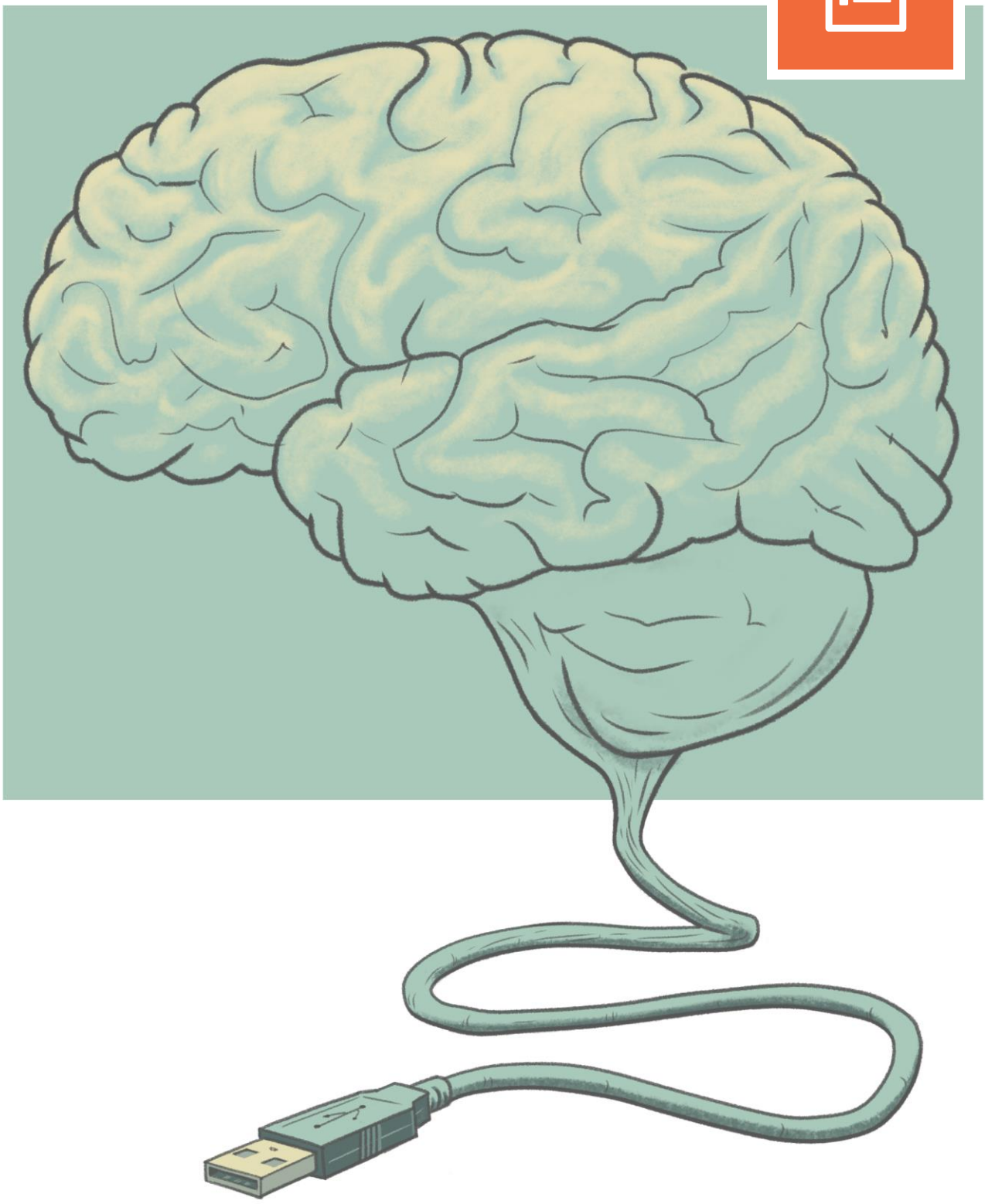


# Artificial intelligence and privacy

Report, January 2018

Datatilsynet
The Norwegian Data Protection Authority

# Contents

# About this report

Most applications of artificial intelligence (AI) require huge volumes of data in order to learn and make intelligent decisions. Artificial intelligence is high on the agenda in most sectors due to its potential for radically improved services, commercial breakthroughs and financial gains. In the future we will face a range of legal and ethical dilemmas in the search for a balance between considerable social advances in the name of AI and fundamental privacy rights. This report aims to describe and help us understand how our privacy is affected by the development and application of artificial intelligence.

The Norwegian Data Protection Authority (DPA) believes it to be imperative that we further our knowledge about the privacy implications of artificial intelligence and discuss them, not only in order to safeguard the right to privacy of the individual, but also to meet the requirements of society at large.

If people cannot trust that information about them is being handled properly, it may limit their willingness to share information – for example with their doctor, or on social media. If we find ourselves in a situation in which sections of the population refuse to share information because they feel that their personal integrity is being violated, we will be faced with major challenges to our freedom of speech and to people's trust in the authorities. A refusal to share personal information will also represent a considerable challenge with regard to the commercial use of such data in sectors such as the media, retail trade and finance services.

This report elaborates on the legal opinions and the technologies described in the 2014 report «Big Data – data protection principles under pressure»[1]. In this report we will provide greater technical detail in describing artificial intelligence (AI), while also taking a closer look at four relevant AI challenges associated with the data protection principles embodied in the GDPR:

- Fairness and discrimination
- Purpose limitation
- Data minimisation
- Transparency and the right to information

The above list is not exhaustive, but represents a selection of data protection concerns that in our opinion are most relevance for the use of AI today. In addition, the report considers the role of the DPA as the supervisory body for AI applications. Finally, we provide a number of examples of methods and tools and recommendations for safeguarding privacy in the development and use of AI.

The target group for this report consists of people who work with, or who for other reasons are interested in, artificial intelligence. We hope that engineers, social scientists, lawyers and other specialists will find this report useful.

Producing this report has been a learning process for the staff of the Norwegian DPA, and we learned a lot from the experiences and appraisals of artificial intelligence and data protection from the stakeholders we were in touch with during the process. We are most grateful to Inmeta, Privacy International, the Financial Supervisory Authority of Norway, Google, Sintef, the Norwegian University of Science and Technology (NTNU), Big Insight at the University of Oslo and the Norwegian Computing Center, Sparebank 1 Stavanger, the Information Commissioner's Office in the UK, the Office of the Privacy Commissioner in Canada, the Office of the Auditor General of Norway, and the Centre for Artificial Intelligence Research at the University of Agder.

## Legal sources and use of terminology

In this report we use artificial intelligence as a collective term describing its various aspects, including machine learning and deep learning.

The basis for this report is the **EU's General Data Protection Regulation (GDPR)**. This Regulation will be enshrined in Norwegian law in the form of a Personal Data Act which will come into force on May 25 2018.[2] We have also drawn upon the **Recitals of the Regulation** in interpreting the contents of the articles. The recitals are not legally binding, but explains the content of the articles.

Furthermore, we have also cited the **statements made by the Article 29 Working Party** and the guidelines it set for individually automated decisions and profiling.[3] The Article 29 Working Party is the European Commission's most senior advisory body on data protection and information security matters.

[1] https://www.datatilsynet.no/om-personvern/rapporter-og-utredninger/temarapporter/big-data/

[2] GDPR text: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:119:FULL

[3] http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083

# Artificial intelligence and data protection

Artificial intelligence (AI) is the concept used to describe computer systems that are able to learn from their own experiences and solve complex problems in different situations – abilities we previously thought were unique to mankind. And it is data, in many cases personal data, that fuels these systems, enabling them to learn and become intelligent.

The development of AI has made some major advances in recent years and its potential appears to be promising: a better and more efficient public sector, new methods of climate and environmental protection, a safer society, and perhaps even a cure for cancer.

We are in other words embarking on a venture that will without doubt have a considerable impact on society. Accordingly, it is important for us to engage in discussion now. What sort of regulatory framework do we need in order to grasp the opportunities offered by AI in an assured and just manner? For we cannot escape the fact that the use of AI raises a number of concerns with respect to ethics, security, legal responsibility, etc. This report is devoted to one such concern: the use of personal data in AI and the issue of privacy.

### From winter to spring – why now?

The concept of AI was known as far back as in the 1950s as a technology in which people had high hopes of success. The initial progress made was however followed by many decades that are often called the AI winter because the early expectations were not met. In recent years, though, we have witnessed the coming of spring.

Today we see that AI is used to solve specific tasks such as, for example, image and speech recognition. This is often called *specialised* AI. *General* AI refers to systems that are as versatile as humans when it comes to learning and problem solving. But it will probably be several decades before this is achieved.

The AI spring has dawned thanks to the availability of huge amounts of data, coupled with an increase in processing power and access to cheaper and greater storage capacity. Big Data often refers to vast volumes of data, extracted from multiple sources, often in real time.[4] These enormous data streams can be utilised for the benefit of society by means of analysis and finding patterns and connections.

This is where AI can make a difference. While traditional analytical methods need to be programmed to find connections and links, AI learns from all the data it sees. Computer systems can therefore respond continuously to new data and adjust their analyses without human intervention. Thus, AI helps to remove the technical barriers that traditional methods run into when analysing Big Data.

### Greater demand for data, more stringent regulations

The new data protection regulations that enters into force in May 2018 will strengthen our privacy rights, while intensifying the requirements made of those processing such data. Organisations will bear more responsibility for processing personal data in accordance with the regulation, and transparency requirements will be more stringent.

At the same time as the requirements are being intensified, demand for data is growing. AI-based systems can become intelligent only if they have enough relevant data to learn from.

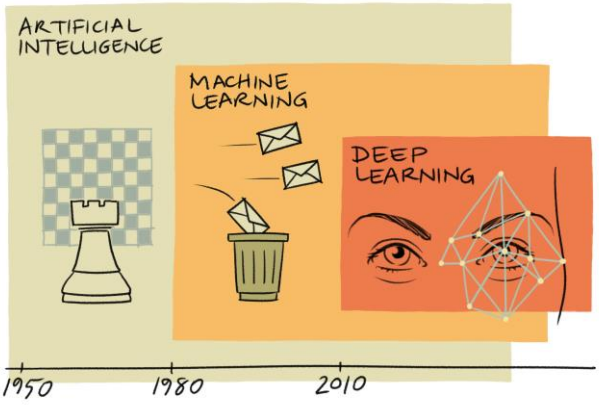
An intelligent chatbot (a computer program that people can interact with by means of ordinary speech, or through written input) analyses all the information it is fed – a combination of questions posed by customers and responses communicated by customer service. From its analysis the chatbot can "understand" what a customer is asking about and is therefore able to give a meaningful answer. The greater the volume of information the chatbot can base its analysis on, the better and more precise will be the reply it gives.

---

[4] https://ico.org.uk/for-organisations/guide-to-data-protection/big-data/

### Artificial intelligence, machine learning and deep learning

Artificial intelligence, machine learning and deep learning are terms that are often used as synonyms even though they are conceptually imprecise. The illustration depicts the relationship between the terms and their development over time.




Artificial intelligence is an umbrella term that embraces many different types of machine learning. Machine learning can be described as "a set of techniques and tools that allow computers to 'think' by creating mathematical algorithms based on accumulated data".[5] The system can reason independently of human input, and can itself build new algorithms.

Deep learning is a form of machine learning. Some types of deep learning build on the same principles as the brain's neural network. Systems of this type are often based on a known set of training data that helps the self-learning algorithms to carry out a task. This is conditional on the network itself being able to determine the correct response for solving the task.[6] This method was crucial in enabling the AlphaGo computer program to defeat one of the world's best players of the Chinese board game Go (see fact box). This was considered to be an important milestone in the continuing development of AI.

### Is it possible to combine artificial intelligence and appropriate data protection?

In compiling this report, we have spoken to a number of AI developers and users. The impression we are left with is that most sectors have adopted AI in a relatively restrictive manner, and that the techniques frequently used are limited. This corresponds fairly well with the limited case portfolio of the Data Protection Authority and the requests for guidance received with regard to AI and privacy.

We are still in the early phase of AI development, and this is the right time to ensure that AI technologies comply with the rules society lays down. The answer to the question as to whether it is possible to use AI, and protect people's data while doing so, is yes. It is both possible and necessary in order to safeguard fundamental personal data protection rights.

#### AlphaGo

AlphaGo is the computer program that defeated one of the world's best players of the Chinese board game Go.

Go is a game of so many possible combinations that it is currently impossible to calculate them all, and what was needed was therefore a more intelligent approach to the game than basic calculating capacity could offer. AlphaGo was developed by Deepmind, who are deep learning experts and could apply it as part of the program.

The program was developed by reviewing historical data drawn from many games played by humans. Then the program played against itself to learn more about the moves and strategies that produced the best results.

One of the most interesting results, apart from the fact that AlphaGo won, was that the program adopted new strategies that were previously unknown. These were published and are now used by Go players.

(Kilde: https://www.blog.google/topics/machine-learning/alphago-machine-learning-game-go/)

[5] https://iq.intel.com/artificial-intelligence-and-machine-learning/

[6] https://no.wikipedia.org/wiki/Nevralt_nettverk, https://en.wikipedia.org/wiki/Deep_learning

# How does artificial intelligence work?

There are two main aspects of artificial intelligence that are of particular relevance for privacy. The first is that the software itself can make decisions, and the second is that the system develops by learning from experience.

In order for a computer system to learn, it needs experience, and it obtains this experience from the information that we feed into it. This input may be in several different formats. If a system is sought that will only perform image recognition and analysis, the experiential data input will naturally enough consist of images. For other tasks the input data will consist of text, speech or numbers. Some systems utilise personal data, while other systems use data that cannot be linked to individuals.

## Machine learning

In order to understand why AI needs huge volumes of data, it is necessary to understand how the system learns.

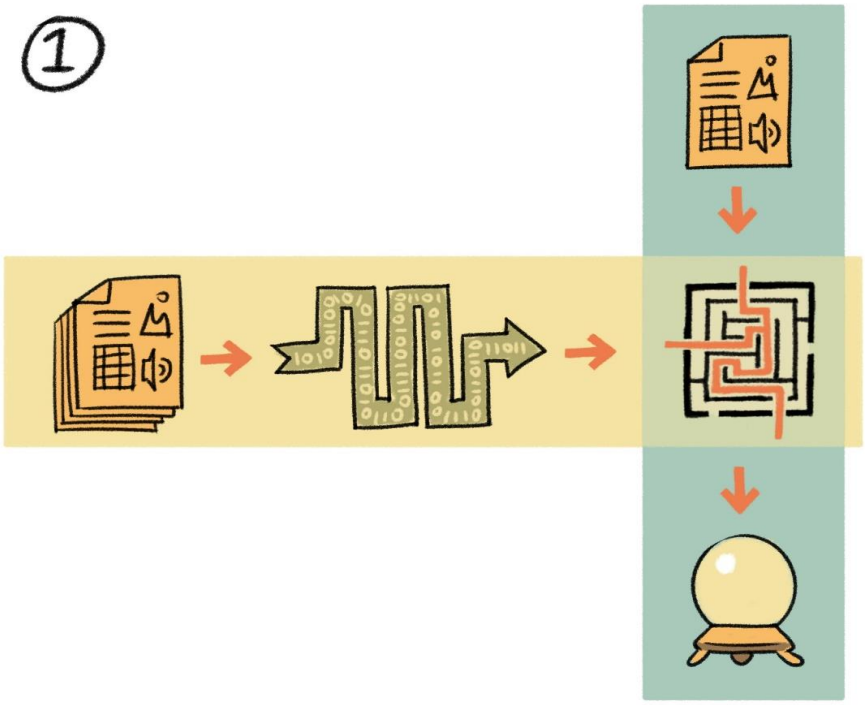
Developing AI requires the input of experiential data. Machine learning generally proceeds in this way: (Illustrated by Figure 1, *from left to right*):

1. Learning starts with selected information containing patterns or similarities.
2. By using machine learning, the patterns found in the information are identified.
3. A model is generated that can recognise the patterns that emerge when fresh data is processed by the model.

**Model** is an umbrella term for the final outcome of learning. There are many different types of models and it is these that are used in commercial applications — such as predicting the type of streamed TV series a consumer prefers. What these models have in common is that they contain essential training data. As the data that the model will process in the future will seldom be completely identical with the training data, a generalisation is required. Certain data that deviate from the main bulk of training data, will therefore usually be removed from the model.

This is how the model works: (Illustrated by Figure 1, *from top to bottom*)

1. The model receives data similar to that used for learning.
2. The model decides which pattern the new data most resembles.
3. The model produces an estimated result.



There are several forms of learning that can be used, depending on whether the information has been labelled or not. Labelled data is tagged data: if the data consists of images, the labels or tags may for example be gender, ethnicity, dog or cat.

Below we have listed the main forms of learning, and we describe how the data is used in these.

### Supervised learning

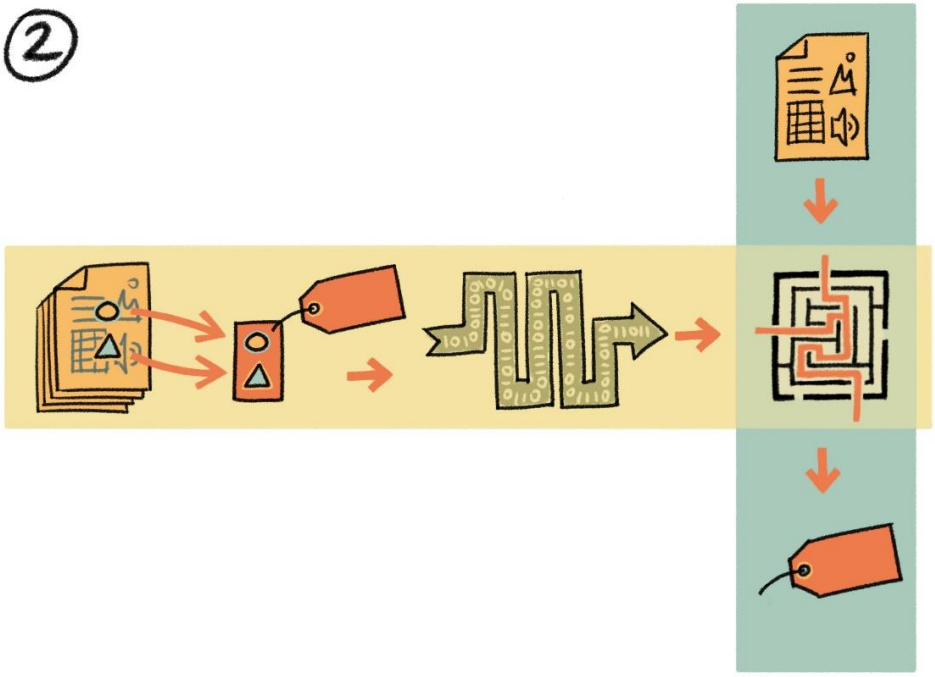
Supervised learning involves the use of labelled data, by means of which the supervision is performed. The dataset is split into two, usually an 80/20 split, with 80 per cent of the data used to train the model. The remaining 20 per cent is used to verify how precisely the model processes unknown data. It is no good if the model performs accurately using the training data and inaccurately using new and unknown data. If the model is too well adjusted to the training data, which we call overfitting, it will not produce satisfactory results using new data. Therefore, the model requires a certain degree of generalisation.

Training data may for example consist of images labelled with information about the contents of each image. Supervised learning may be compared to teaching a child. For example, we point to a number of objects to the child and give them names. If we show a number of cats to a child, the child will gradually learn to recognise other cats than those originally shown. In similar fashion, a machine learning model will develop the same ability to recognise objects based on labelled images.

If one is working with a dataset and wishes to separate men and women, one can use different features that are of relevance. The features used will depend on the basic data available. For example, women live longer than men on average, so life duration is of relevance when differentiating between genders. This feature will, however, prove to be somewhat narrow in most cases, and is mentioned here only as an example. If one's data basis consists of images, then hair length, or the use of make-up or jewellery, may be relevant features. The example below illustrates how two different features are used in learning.

Learning takes place as follows (Illustrated by figure 2, *from left to right*):

1. A set of labelled data is used.
2. Depending on data type, and what is considered relevant, the features (circles and triangles) to be used for learning are selected. The data is labelled to denote the right answer.
3. A model is built that, based on the same features, will produce a label.



We will often also know which features of labelled data are most decisive for correct categorisation or for producing the right result. It is important to have persons with sound knowledge of the field in question in order to identify the most relevant features. The correct selection of relevant features may be of much more importance than the amount of data, an issue we will be addressing later. One advantage of labelled data is that it enables an easy check of the model's precision.

When we use the model, the following takes place (Fig. 2, *top to bottom*):

1. New data of the same type as the training data is fed into the system.
2. The relevant features are fed into the model and processed.
3. The model produces a result that corresponds with the labels used in training.

### Unsupervised learning

In unsupervised learning, data is used that has not been pre-labelled, as the aim is for the system to group data that is similar. If, for the sake of simplicity, we again consider data consisting of cat and dog images, the goal would be for this data, to the greatest extent possible, to be sorted into two groups – one consisting of images of dogs, and the other of cat images.

Learning proceeds as follows (Fig.3, *left to right*):

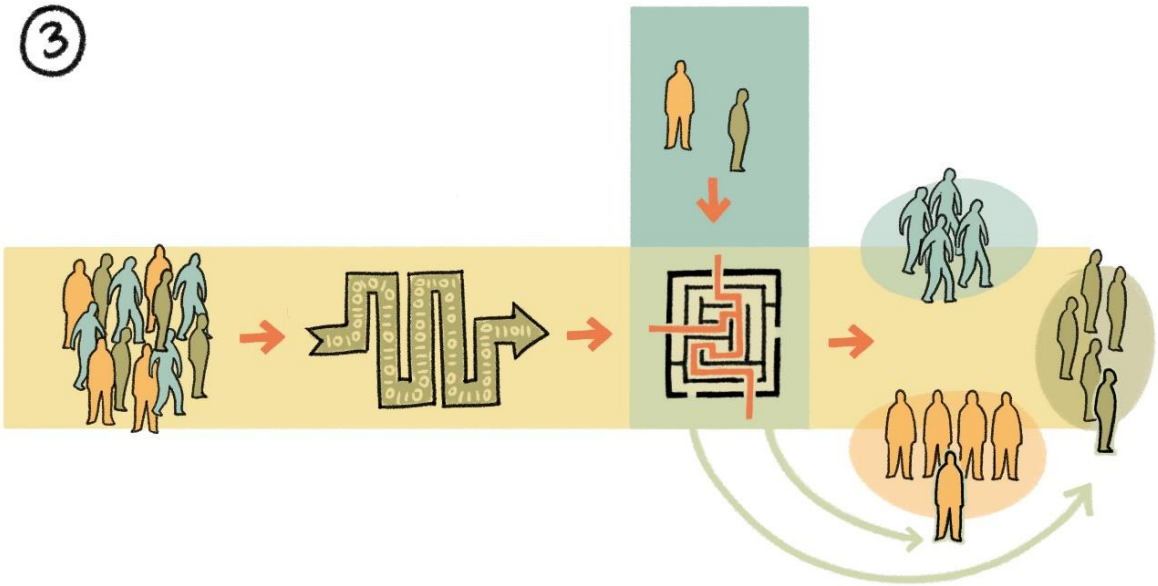
1. A dataset is used in which there must be a certain number of similarities, or patterns, if it is to be meaningful.
2. The patterns are revealed.
3. A model is built that can recognise and differentiate patterns.

This is what takes place when using the model (Fig. 3, *top to bottom*):

1. New unlabelled data of the same type as the training data is fed into the system.
2. The model identifies the data patterns.
3. The model tells which group the new data belongs to.

A disadvantage of this method is that the model cannot place data in other groups than those discovered during the learning process. It is therefore very important that the training basis is representative.

### Reinforcement learning



This form of learning is based on trial and error, as well as on optimisation, as the model learns which actions are targeted towards the goal. This means that less data, or no data at all, is needed for the system to learn.

### AlphaGo Zero

Earlier we mentioned AlphaGo as an example of machine learning. AlphaGo was first trained using a data set consisting of 30 000 games of Go. In order to further improve AlphaGo's ability to play Go, it was programmed to play against itself. Its experiential basis could be enlarged considerably through trial and error, without it needing to obtain data from more games. It also gave AlphaGo the opportunity of discovering moves and strategies not in the original training set.

The latest version – AlphaGo Zero – was devised in order to start playing without using training data. It was programmed only with the rules for Go, and was not fed any information about previously played games It then learned to play against itself. After 40 days it was able to beat the previous AlphaGo version 100-0. It is also interesting to note that the Zero version of AlphaGo requires much less computing power to achieve these results.

(Source: https://deepmind.com/blog/alphago-zero-learning-scratch/)

## Results of learning

Regardless of the algorithms or methods used for machine learning, the result will be a "model", which is in fact an umbrella term for all machine learning. The model can then be fed with new data to produce the desired type of result. This may be, for example, a labelling, or a degree of probability, or similar.

It is worth noting that the model does not normally hold the source data directly. It holds an aggregate representation of all the data used to train the system.

Decision trees represent one exception to this, as they contain a varying degree of the model's data basis. The limits here depend on whether the tree is "pruned" after learning, or a level limitation is set for learning. One or the other will normally be chosen, as the model should generalise and not overfit. In a deep-learning model, the basic data will be represented as numerical values in the neural network. It should, therefore, not be possible to retrieve any personal data used to train the model. We shall take a closer look at these models a little later, in the section entitled the Black Box.

### Model use – static and dynamic (offline/online)

A model may be used in two ways. The first way is to use a **static, or offline model**, that will not change through use. A static model will always, as the name suggests, operate in the same way and produce the same results throughout its entire lifecycle. All new model training will take place in a test environment, and all changes require that the model is replaced by a new version. This means that full control is maintained of the model in use.

The other possibility is provided by a **dynamic, or online model**. The model is used in a similar fashion to the static model. However, the difference is that the dynamic model is able to avail itself of input data in order to improve and adjust to changes. This may, for example, be necessary in connection with the monitoring of credit card transactions in order to reveal fraud. The transactions may change according to the user's life situation, or in relation to his job, by for example taking place in completely new locations. These new usage patterns could well be labelled suspicious by a static model and potentially result in a blocked credit card. A model can therefore become less accurate over time if it is not continuously updated.

A spam filter provides a good example of a typical area of application for a dynamic model which can be improved by the user indicating emails that have been wrongly labelled. The disadvantage of dynamic models is that there is less control over the model's development and the changes have immediate effect. A good example of this is the Microsoft chatbot Tay which learned from conversations with Internet users. After a brief period on Twitter the chatbot was described as a "Hitler-loving sex robot" by the media. Microsoft decided to remove Tay only 24 hours after it had been launched.[7]

[7] http://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/

## The more training data the better?

The more training data we can feed into the model, the better the result: this is a typical mantra frequently heard in connection with machine learning. In most instances the computer will require a lot more data than humans do in order to learn the same thing. This currently sets a limit for machine learning, and is compensated for by utilising considerable amounts of data – often greater than a human being would be able to manage.

It is worth noting that the quality of the training data, as well as the features used, can in many instances be substantially more important than the quantity. When training a model, it is important that the selection of training data is representative of the task to be solved later. Huge volumes of data are of little help if they only cover a fraction of what the model will subsequently be working on.

Correct labelling is extremely important when conducting supervised learning. If data has been incorrectly labelled, there will obviously be a negative impact on the training outcome. As the saying goes: garbage in, garbage out.

### Breadth and depth of data

The efficiency of machine learning can be heavily influenced by how the basic data is presented to the algorithms that develop models, and also by which features one chooses to use.

Like a spreadsheet, a dataset for machine learning may consist of rows and columns. If one has person-related data, the columns may well denote a person's age, gender, address, marital status, height, weight, nationality, etc. The rows will represent individual persons. Consideration must be given to the quantity of personal information needed in order to train the desired models, as well as its relevance to the chosen purpose.

In selecting relevant features, there will often be a need for persons who are expert in the relevant fields. It's not always the case that the basic data tells the whole story.

Good selection is important, otherwise one risks ending up with too many features, or what specialists call "The Curse of Dimensionality». Put simply, this means that an excessive number of features will result in matches being lost amongst all the non-matching data. This will mean that enormous volumes of data will be needed by way of compensation.

### Example

A US hospital undertook a trial to categorise the risk of complications for patients suffering from pneumonia. The result was that patients suffering from *both* asthma and pneumonia were categorised as low-risk patients – to the doctors' great surprise.

Though these patients ran a higher risk, their survival rate was better. What the model was not able to detect, was that the apparently low risk was a result of these patients getting better care and more intensive treatment.

This illustrates the risks inherent in using data without domain knowledge, and that the basic dataset does not always tell the whole story.

(Kilde: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf)

One disadvantage of reducing the scope of feature selection is that one may lose possible matches, or patterns, that were not previously known or which had not been thought of. This is partly why it is necessary to include persons with domain knowledge in this project phase. One should also consider what constitutes a good enough result.

It is worth mentioning here that deep learning is somewhat of an exception in this respect. The selection and adjusting of features are not as important as they are in other learning methods. For example, feature selection is conducted via value weights in a neural network. The disadvantage of not making selections means that one needs a vastly greater volume of training data.

### Feature engineering

An important factor in achieving good results is how the dataset is presented. Relevant correlations may be concealed if the data is not used properly. In many instances there is a great deal more to be gained by smart data use than by increasing the amount of data.

Dates are one example. Let us consider the date 1.10.2017, which tells us it is the first day in the month and the tenth month of the year. It might well be that the information would be more useful if we could

## Norwegian Tax Administration

The Norwegian Tax Administration (NTA) has developed a predictive tool to help select which tax returns to check for errors or tax evasion. They tested roughly 500 different variables that revealed information regarding a tax payer's demography, life history, and other details in his/her tax returns. Only 30 variables were built into the final model. They include details regarding deductions made in the current and previous year, age, financial details such as income and assets, as well as details regarding individual tax return items.

This provides a good example of how it is not always necessary to use all the available data in order to achieve the desired purpose. Without knowing how the NTA decided on feature selection for its project, we can see that they set limits and they confirm out that this was sufficient for them to achieve their goal.

(Source: Skatteetatens Analysenytt 1-2016, http://www.skatteetaten.no/globalassets/pdfer/skatteetatens_analysenytt/analysenytt-1_2016_web_hele.pdf)

convert it to show what day of the week it is: a Sunday in this case.

In Norway, where there are four quite distinct seasons, we might consider grouping the months in order to represent the data in a better way. Month 10 could then be represented as autumn. Autumn itself could be represented as numerical value 3, while spring would be 1, summer 2, and winter 4. In this way we could derive more features from a data item, or reduce the number of different values. If data is extracted from multiple sources, steps should be taken to ensure that they are in the same format. In US data for example, the month will be denoted by the 1 and the day by the 10 in the data formula 1.10.2017.

The normalisation of features may also be necessary in order to ensure that certain features do not create an imbalance in the training data, or that a few extreme values do not adversely affect the rest. Put simply, we can say that it is important to ensure that everything is similarly scaled. If there are features where a change of 0.1 signifies as much as a change of 1000 for another feature, it is essential that they be re-aligned to the same scale.

### Enough is enough?

It can be difficult at the outset to estimate the amount of learning data that will be needed. It will depend on the type of machine learning employed, the number and characteristics of the features selected, and the quality of the basic data. Also of relevance here is the degree of accuracy a model needs for the objective to be achieved. If a person doing the job is 75 per cent accurate, will that be good enough for the model? If the goal is 100 per cent accuracy, a substantial amount of data will be needed.

The area of application will define what is reasonable when using personal information as training data. One would pursue the objective of diagnosing fatal illnesses differently than one would go about profiling someone in order to target advertisements for them as accurately as possible.

If we stick to the data minimisation principle, it would be natural to commence with a restricted amount of training data, and then monitor the model's accuracy as it is fed with new data. The learning curve is one tool used for assessing this.[8] These enable one to see, having started with a limited set of data, when a curve flattens and new data ceases to add training value.

## The Black Box

One concern in relation to machine learning is that one does not always know how the result is produced. Which features, or which combinations of features, are the most important? A model will often produce a result without any explanation. The question then arises as to whether it is possible to study the model, and thus find out how it arrived at that specific result.

As mentioned above, specialists at the Norwegian Tax Administration have built a predictive model that helps them select the tax returns to be scrutinised more closely. They state the following: "When we build a model in this way, we don't necessarily know what it is that gives a tax payer a high ranking for error risk. The ranking is the result of complex data aggregation in the model."

[8] https://www.coursera.org/learn/machine-learning/lecture/Kont7/learning-curves http://www.ritchieng.com/machinelearning-learning-curve/

This statement by the NTA underscores the relevancy of the black box issue. In this case only 30 different features are used, but it is possible for a system to use a lot more than that. It would then be even more difficult to identify what was relevant for the outcome.

### How to understand and explain what's behind it

When machine learning is employed, the end product is a model. When it comes to machine learning models, the ease with which their results can be checked varies greatly, even though the same training data is used.

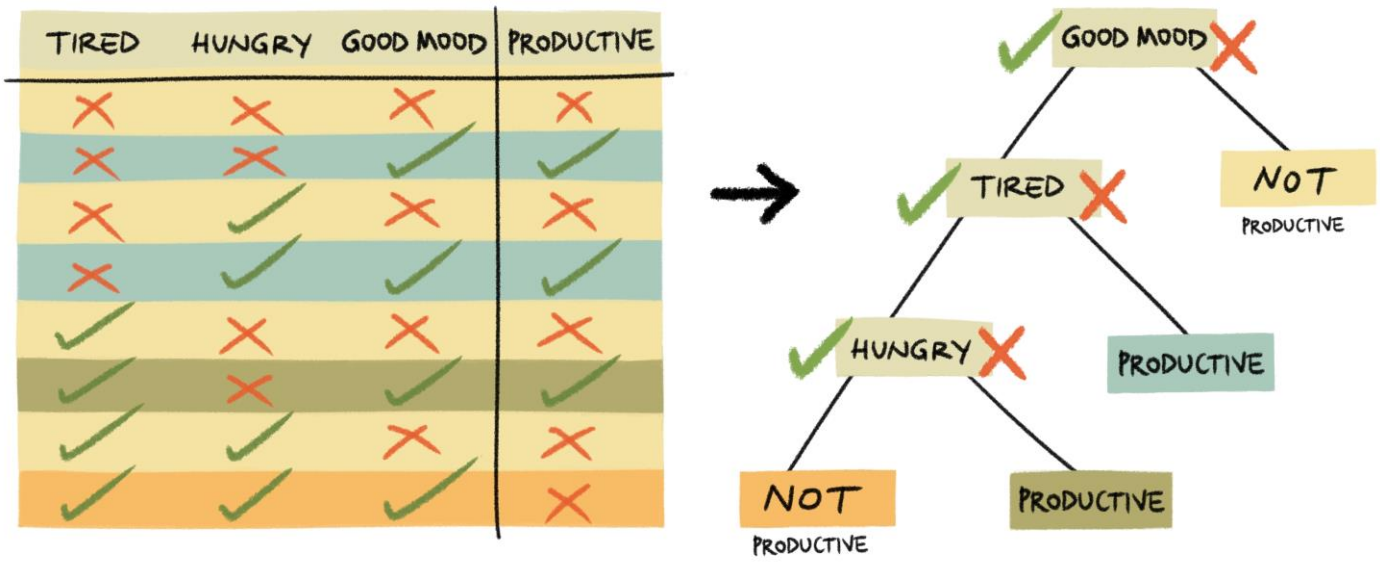
Deep learning and neural networks are often the first elements to be mentioned when black box issues are discussed, without their defining the issue fully.

We will now consider two examples that represent extremes of ease and difficulty in understanding and checking these models, namely so-called decision trees and deep neural networks.

**Decision trees**

A decision tree is one of the simplest models. In its most basic form all the data is broken down in such a way that it can be placed in the tree. One starts at the top and at each level selects a branch based on a particular feature's value. One continues right to the base of the tree, where the final outcome – the decision – is found (see figure below).

This type of model affords a high degree of transparency, at least when the tree is based on a manageable amount of data. It is possible to move up through the tree to see the criteria on which the result is based. With increasing amounts of data, however, a point will be reached where it will be difficult for a person to obtain an overview and understanding.

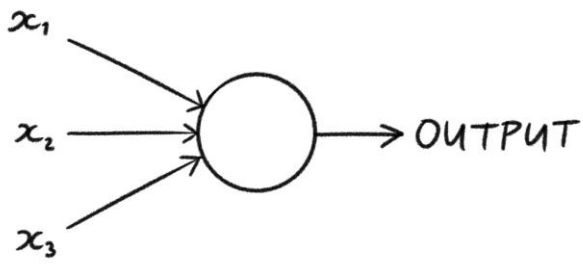


**Neural networks**

Neural networks are used in a methodology that is largely inspired by our understanding of the way the human brain functions. These networks are built by what is basically a very simple component (a perceptron), but very many of these components can be used to create large and complex networks.

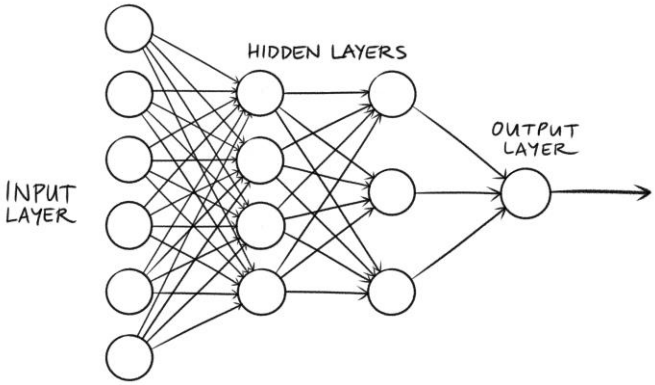
A perceptron, illustrated here below, has a variable number of inputs and one output:

Each «leg» of the perceptron has a weight value. This value determines how great will be the influence of the input feature on the final result. These values are adjusted when the network is trained to give the desired results. This is often carried out by working backwards in the network to adjust the values of the relevant perceptrons so that the final result is right




(backpropagation). This is an automated process that is a part of the learning process.

A neural network consists of three parts; an input layer, one or more hidden layers, and an output layer:




If there is more than one hidden layer, then this is considered to be deep learning. In the above figure we have a single neural network in which all the input data move from left to right, and emerge as a result. There are several variants of these neural networks. Some form loops and also send the data from right to left within the network before the final result is produced.

One of the challenges here is that the input data is viewed in isolation. In many situations we work with information that has a context. For example, some words carry different meanings depending on their context. This context does not need to be formed by the same sentence. This is part of the reason why some neural networks have a form of short-term memory. This allows them to produce different outputs based on the data that was processed previously, which of course makes it more difficult to determine how a result was derived. This also means that it can be very difficult to merely examine the algorithms to find out how they work and what decisions they reach.

The number of layers in a neural network may vary. An example of this is that in 2016 Microsoft won an image recognition competition using a network consisting of 152 layers.[9] The size of the network, and the number of connections, will depend on the number of input values and how the layers are interconnected. Clearly, the size of the neural network mentioned is way beyond what can be comprehended or examined without the help of suitable tools. We shall be looking at such tools in the final chapter.

[9] https://blogs.microsoft.com/ai/2015/12/10/microsoft-researchers-win-imagenet-computer-vision-challenge/

# Artificial intelligence meets the GDPR

The provisions of the GDPR govern the data controller's duties and the rights of the data subject when personal information is processed. The GDPR therefore applies when artificial intelligence is *under development* with the help of personal data, and also when it is used to *analyse* or *reach decisions* about individuals.

In this chapter we will review the principles of data protection and the articles of the GDPR that are especially relevant to the development and use of artificial intelligence.

## Fundamental principles of data protection

The rules governing the processing of personal data have their basis in some fundamental principles. Article 5 of the GDPR lists the principles that apply to all personal data processing. The essence of these principles is that personal information shall be utilised in a way that protects the privacy of the data subject in the best possible way, and that each individual has the right to decide how his or her personal data is used. The use of personal data in the development of artificial intelligence challenges several of these principles.

In summary, these principles require that personal data is:

- processed in a lawful, fair and transparent manner (principle of legality, fairness and transparency)
- collected for specific, expressly stated and justified purposes and not treated in a new way that is incompatible with these purposes (principle of purpose limitation)
- adequate, relevant and limited to what is necessary for fulfilling the purposes for which it is being processed (principle of data minimisation)
- correct and, if necessary, updated (accuracy principle)
- not stored in identifiable form for longer periods than is necessary for the purposes (principle relating to data retention periods)
- processed in a way that ensures adequate personal data protection (principle of integrity and confidentiality)

### § Personal data

Personal data means any information relating to an identified or identifiable natural person. (GDPR Article 4 (1))

The data may be *directly* linked to a person, such as a name, identification number or location data.

The data may also be *indirectly* linked to a person. This means that the person can be identified on the basis of a combination of one or more elements that are specific to a person's physical, physiological, genetic, mental, economic, cultural or social identity.

### § Processing

Processing means any operation or set of operations which is performed on personal data, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

(GDPR Article 4 (2))

### § Data controller

Data controller means 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))

In addition, the data controller is responsible for, and shall be able to prove, compliance with the principles (accountability principle).

In the following we will review the most important data protection challenges associated with developing and using artificial intelligence. We look at these challenges in the light of the data protection principles that are most relevant to artificial intelligence – namely the principles of fairness, purpose limitation, data minimisation and transparency.

## Algorithmic bias meets the fairness principle

It is easy to think that artificial intelligence will be able to perform more objective analyses and therefore reach better decisions than human beings. After all, artificial intelligence will not be affected by low blood sugar, by having a bad day, or by the desire to help a friend.

And yet algorithms and models are no more objective than the people who devise and build them, and the personal data that is used for training. The model's result may be incorrect or discriminatory if the training data renders a biased picture reality, or if it has no relevance to the area in question. Such use of personal data would be in contravention of the fairness principle.

This principle requires all processing of personal information to be conducted with respect for the data subject's interests, and that the data be used in accordance with what he or she might reasonably expect. The principle also requires the data controller to implement measures to prevent the arbitrary discriminatory treatment of individual persons. The Regulation's preface describes the use of suitable mathematical or statistical procedures as possible measures here.

This would not, however, be sufficient of itself to ensure compliance with the principle. The model must also be trained using relevant and correct data and it must learn which data to emphasise. The model must not emphasise information relating to racial or ethnic origin, political opinion, religion or belief, trade union membership, genetic status, health status or sexual orientation if this would lead to arbitrary discriminatory treatment.

If it is suspected, or claimed, that use of a model will entail unfair or discriminatory results, the Data Protection Authority can investigate whether the principle of fairness has been safeguarded in the processing of personal data. These investigations may include a review of the documentation underpinning the selection of data, an examination of how the algorithm was developed, and whether it was properly tested before it came into use.

### Example

A claim of AI-based discrimination was levelled against a US system for setting bail conditions and sentencing. The system is used to predict the risk of a convicted person committing a new crime.

The journal ProPublica studied the decisions reached by the system and concluded that it discriminated against black defendants. The number of blacks erroneously flagged as being high re-offending risks, was twice as high as the number of whites so classified.

The company that developed the software disagreed with ProPublica's conclusion, but it was unwilling to allow the criteria and calculations used in developing the algorithm to be examined. It is therefore impossible for the convicted person, or the general public, to obtain clear information as to why and how such decisions are reached.

(Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)

## Artificial intelligence meets the principle of purpose limitation

Many of the models developed using artificial intelligence will be used in connection with good causes, such as cancer diagnosis. Are we permitted to use personal data unrestrictedly as long as it is for a good cause?

The purpose limitation principle means that the reason for processing personal data must be clearly established and indicated when the data is collected. This is essential if the data subject is to exercise control over the use of his/her information. The purpose of the

processing also needs to be fully explained to the data subject if he or she is to be able to make an informed choice about whether or not to consent to it.

Yet the development and application of artificial intelligence often requires many different types of personal data – information that in some cases has been collected for other purposes. For example, it is possible that a person's Facebook activities are built into an algorithm that determines whether she will obtain a mortgage from the bank. Such recycling of information may be useful and provide more accurate analyses than those which were technically feasible previously, but it can also be in contravention of the purpose limitation principle.

In cases where previously-retrieved personal data is to be re-used, the controller must consider whether the new purpose is compatible with the original one. If this is not the case, new consent is required or the basis for processing must be changed. In the Facebook example discussed above, the data subject must consent to Facebook information being used by the bank in connection with mortgage applications in order to ensure that processing is conducted in compliance with the purpose limitation principle.

### New technology – new science?

The purpose limitation principle is highly important in ensuring that the data subject exercises control over his or her own personal information. There are, however, exceptions to the principle. Further processing of data is, for example, considered to be compatible with the original purpose if it takes place in connection with scientific or historical research, or for statistical and archival purposes in the public interest. This begs the question as to what constitutes scientific research, and to what extent the development and application of artificial intelligence is scientific research.

More and more university and hospital research environments are working on developing tools that use artificial intelligence. Examples include models that identify the risk of tax or social benefit fraud, or image recognition software that diagnoses cancer in tumours. But how do we really define scientific research?

The General Data Protection Regulation does not define what constitutes scientific research. A general understanding of the concept, however, is that it must relate to efforts aimed at discovering new knowledge or know-how.[10] The GDPR's preface (Recital 159) states that scientific research should be interpreted broadly and include technological development and demonstration, basic research, as well as applied and privately financed research. These elements would indicate that – in some cases – the *development* of artificial intelligence may be considered to constitute scientific research.

Applying artificial intelligence to assess a person's creditworthiness cannot, however, be said to be aimed at gaining new knowledge. In this case, the use of artificial intelligence cannot be defined as scientific research. But is it always possible to differentiate between the development and the application of AI?

### § Matters for consideration

The Regulation's preface (Recital 50) states that the following factors should be included when ascertaining whether the further processing of personal data is compatible with the original purpose:

- any connection between the original purpose and the purposes of the intended further processing
- the context in which the data was collected
- the data subject's relation to the controller and how this may affect the subject's reasonable expectations with regard to further processing
- the nature of the personal data
- the consequences for the data subject of the intended further processing
- whether the original processing operations and the new ones are subject to the appropriate safeguards

This list is not exhaustive and all issues that are relevant in the individual case must be included in the appraisal.

[10] Store Norske Leksikon

When the completed model is static (offline), development and use can be clearly differentiated. A model developed using training data is tested on similar data before it is used. Once the model is put into use, the training data is removed from the algorithm and the model will only process the personal data to which it is applied, such as information about loan applicants. Because the algorithm is static, it will not learn anything further from the personal data it is currently processing. Consequently, nor will it develop intelligence once it has been put into use.

Other models develop and improve continuously as they are fed more personal data. These include models that provide decision support for doctors. The model learns something new about every patient it receives data about, or every scientific article it reads. This new knowledge can then be used on the next patient.

When a model develops on a continuous basis, it is difficult to differentiate between development and use, and hence where research stops and usage begins. Accordingly, it is therefore difficult to reach a conclusion regarding the extent to which the development and use of these models constitute scientific research or not. The limits on what constitutes scientific research will need to be reviewed once the new data protection regulations come into force.

We emphasise that the use of personal data for scientific research is governed by specific rules in the GDPR (Article 89). Use in such instances must be subject to the appropriate safeguards to secure the data subject's rights and freedoms. The safeguards must ensure that technical and organisational measures are in place to protect the data minimisation principle in particular.

## Artificial intelligence meets data minimisation

It often takes huge amounts of personal data to develop artificial intelligence.

On the other hand, the principle of data minimisation requires that the data used shall be adequate, relevant and limited to what is necessary for achieving the purpose for which the data is processed. This means that a controller cannot use more personal data than is necessary, and that the information selected must be relevant to the purpose.

A challenge when developing AI is that it may be difficult to define the purpose of processing because it is not possible to predict what the algorithm will learn. The purpose may also be changed as the machine learns and develops. This challenges the data minimisation principle as it is difficult to define which data is necessary.

However, data minimisation is more than a principle limiting the amount of detail included in training or in the use of a model. The principle also stipulates proportionality, which restricts the extent of the intervention in a data subject's privacy that the use of personal data can involve. This may be achieved by making it difficult to identify the individuals contained in the basic data. The degree of identification is restricted by both the *amount* and the *nature* of the information used, as some details reveal more about a person than others. The use of pseudonymisation or encryption techniques protect the data subject's identity and help limit the extent of intervention.

This principle also forces developers to thoroughly examine the intended area of application of the model to facilitate selection of relevant data necessary for the purpose. Furthermore, the developer must consider how to achieve the objective in a way that is least invasive for the data subjects. The assessments performed need to be documented, so that they can be presented to the Data Protection Authority in the event of an inspection, or in connection with a preliminary discussion.

### § Data protection impact assessment

Before personal information is processed the impacts on data protection must be assessed if it is likely that the process will represent a risk to the rights and freedoms of natural persons. This is particularly the case when using new technology, and consideration must be given to the nature of the processing, its scope and purpose, and the context in which it is performed.

If the risk is high, and the data controller cannot limit it, he or she is duty bound to initiate preliminary discussions with the Data Protection Authority.

(GDPR Articles 35 and 36)

Although it is difficult to establish in advance the exact information that will be necessary and relevant to the development of an algorithm – and this may change during the project – it is essential for the data minimisation principle to be adhered to by means of continuous assessment of the actual requirements. This not only protects the rights of the data subjects, but also minimises the risk of irrelevant information leading the algorithm to find correlations that rather than being significant are coincidental and to which no weight should be attached.

The pressure to use personal data is intensifying as AI-based analyses are employed to promote increased efficiency and better services. The Data Protection Authority believes that the principle of data minimisation should play a major role in the development of artificial intelligence so that the rights of data subjects are protected and general confidence in the models retained.

## The black box meets the principle of transparent processing

Data protection is largely about safeguarding the rights of individuals to decide how information about themselves is used. This requires that controllers are open about the use of personal data, that such use is transparent.

Transparency is achieved by providing data subjects with process details. Data subjects must be informed about how the information will be used, whether this information is collected by the data subjects themselves or by others (GDPR Articles 13 and 14). Besides, the information must be easily available, on a home page for example, and be written in a clear and comprehensible language (GDPR Articles 12). This information shall enable the data subjects to exercise their rights pursuant to the GDPR.

It can be challenging to satisfy the transparency principle in the development and use of artificial intelligence. Firstly, this is because the advanced technology employed is difficult to understand and explain, and secondly because the black box makes it practically impossible to explain how information is correlated and weighted in a specific process.

It is also challenging that information about the model may reveal commercial secrets and intellectual property rights, which according to the GDPR's preface (Recital 63) the right of access must avoid. Consideration of others' rights, such as the commercial secrets of an organisation, may nevertheless not be used to deny a data subject access to all data relating to her. The answer is to find a pragmatic solution. In most cases, furnishing the data subject with the information she needs to protect her interests, without at the same time disclosing trade secrets, will not be problematical.

Although AI is complex, and difficult to understand and explain, the principle of transparent processing of personal data applies with full force in the development and use of artificial intelligence.

Below we will discuss the duty to inform and the rights of data subjects.

### General information

When personal data is collected, the data controller must always provide some *general information* such as

- the identity of the data controller
- how the data controller can be contacted
- the purpose of processing
- the legal basis for processing
- the categories of personal data that are processed
- and the data subjects' right to inspect the data

Information must also be provided regarding risks, rules, safeguards, and the rights of the data subjects in connection with processing, as well as how these rights can be exercised.

In addition, an *extended duty to inform* will apply when personal data is collected for automated decision-making. The use of artificial intelligence is a form of automated processing, and, moreover, in some cases the decision is taken by the model. It is important that we clarify what is required for a decision to be described as automated, before we take a closer look at the extended duty to inform.

### Individual automated decisions

Individual automated decisions are decisions relating to individuals that are based on machine processing. An example of this is the imposition of a fine on the basis of an image recorded by an automatic speed camera. Automated decisions are defined and regulated in Article 22 of the GDPR.

Essentially, automated individual decisions are not permitted. Exceptions apply, however, if the automated decision is a necessary condition for entering into a

contract, is permitted by law, or is based on the explicit consent of the data subject. The regulation does not define what constitutes explicit consent as opposed to ordinary consent, but the phrase indicates that an express gesture by the data subject is required.

In order to meet the requirements of the Regulation, the decision must be **based solely on automated processing,** and it must produce **legal effect,** or similarly **significantly affect a person**.

That an automated decision must be based solely on automated processing, means that there cannot be any form of human intervention in the decision-making process. "Human intervention" means that a natural person must have undertaken an independent assessment of the underlying personal data, and be authorised to re-examine the recommendations the model has produced. The rules governing automated decision-making cannot be circumvented by fabricating human intervention.

What is meant by legal effect is not defined in the preface. It would be natural to understand this phrase as meaning that the automated decision must impact on the data subject's rights or duties, such as legal rights, or the rights set out in a contract. See the examples listed in the fact box.

Nor is the alternative that the automated decision similarly significantly affects a person defined more closely. We assume that the decision must have the potential to affect the circumstances, behaviour or choices of the person who is subject to the automated decision. Yet it is difficult to state precisely where the line should be drawn, as there are considerable subjective elements in such an appraisal.

When automated decisions are applied, measures must be implemented to protect the data subject's rights, freedoms and rightful interests. The data subject must be able to demand that a human being takes the final decision, and she must have the right of appeal.

Automated decisions that involve **special categories of personal data (sensitive personal data)** are permitted only if the data subject has consented, or if they are legally warranted.

It is important to be aware that the alignment of different types of personal data can reveal sensitive information about individuals. Operating with this data will involve the processing of special categories of personal data.

For example, one study combined "likes" on Facebook with information from a simple survey and predicted male users' sexual orientation with an accuracy of 88 per cent. Moreover, they predicted ethnicity with 95 per cent

## GDPR Article 22

Our interpretation of Article 22 is based on the latest draft of the Article 29 Working Party's guidelines on automated decision-making.

This draft is based on submissions from 64 organisations, and is planned for publication at the beginning of February 2018.

The Article 29 Working Party consists of representatives of the EU states' data protection authorities. As an EEA country, Norway has observer status. The working party's statements normally carry considerable weight.

(Article 29 Data Protection Working Party: xx/2017 on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679)

## Examples

**Legal effect:**

- If you are banned from entering a country
- If you satisfy the requirements for receiving unemployment benefit or social security benefit
- If your electricity supply is cut off because you have not paid your bills

**Decisions that similarly significantly affect a person:**

- Automatic rejection of a credit application on the Internet
- Electronic recruitment without human intervention

accuracy, and whether a user was Christian or Muslim with 82 per cent accuracy.[11] A study of this nature is subject to the same legal obligations pursuant to the GDPR as if sensitive personal data had been processed from the outset.

## § Special categories of personal data

Special categories of personal data include information about racial or ethnic origin, political convictions, religious or philosophical beliefs or trade union membership, as well as the processing of genetic and biometric data with the aim of uniquely identifying a natural person, health details or information regarding a person's sexual relationships or sexual orientation.

(GDPR Article 4)

### Right to information in connection with individual automated decisions

In addition to receiving the above-mentioned general information, data subjects must be informed that their personal data is being collected for use in an automated decision-making process. Relevant information must also be given regarding the underlying logic of the model, as well as the significance and anticipated impacts of the automated process.

The information given regarding the model's logic will cover, for example, such aspects as whether decision trees are to be used, and how the data is to be weighted and correlated. Because the information must be readily understood by the data subject, it is not always necessary to provide a thorough explanation of the algorithm, or even to include the algorithm.

Data subjects must also be informed about how automated decisions may affect them. An insurance company that employs automated decision-making to set its motor insurance premiums on the basis of policy holders' driving patterns, should inform its customers of the possible impacts of this, and that careless driving can lead to higher premiums.

The data subject must receive the information described here before automated processing commences. It will enable the data subject to lodge a complaint against processing, or to consent to it.

### The right to an explanation of an automated decision?

Can the data subject request an explanation of the content of the decision once it has been reached, in other words an explanation of how the model arrived at its result?

The preface states that the necessary guarantees given in cases of automated processing shall include "specific information … and the right to … obtain an explanation of the decision reached after such an [automated] assessment" (Recital 71). The preface states that the data subject is entitled to an explanation of how the model arrived at the result, in other words how the data was weighted and considered in the specific instance.

However, the right to an explanation does not appear in the GDPR itself. The implications of the linguistic differences between the preface and the wording of the articles are unclear,[12] but the preface itself is not legally binding and cannot of itself grant the right to an explanation.

Regardless of what the differences in language mean, the data controller must provide as much information as necessary in order for the data subject to exercise his or her rights. This means that the decision must be explained in such a way that the data subject is able to understand the result.

The right to an explanation does not necessarily mean that the black box must be opened, but the explanation has to enable the data subject to understand why a particular decision was reached, or what needs to

[11] Michael Kosinski, David Stilwell and Thore Graepel. «Private traits and attributes are predictable from digital records of human behaviour. Proceedings of the National Academy of Sciences of the United States of America»: http://www.pnas.org/content/110/15/5802.full.pdf

[12] See for example Andre Burt, «Is there a right to explanation for machine learning in the GDPR?»: https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/ cf. Sandra Wachter, Brent Mittelstadt, Luciano Floridi, International Data Privacy law, forthcoming, «Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation», available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2903469

change in order for a different decision to be reached.[13] The data subject must be informed as to how he can oppose the decision, either by appealing or by requesting human intervention.

### Does one have a right to an explanation when a human being makes a decision based on the model's recommendtion?

Sometimes an automated *process* occurs that does not lead to an automated *decision*. Instead, a human utilises the information produced by the automated process to reach a decision, for example by employing a decision support tool. The preconditions for an automated decision to have been taken will therefore not have been satisfied. The question will therefore be whether the data subject is entitled to the same explanation as in the case of an automated decision.

There are no articles in the GDPR, or statements in the preface, regarding the right to an explanation of a specific decision when the preconditions for automated decisions are not satisfied.

The data subject is nevertheless entitled to be given the information necessary for her to safeguard her rights. The transparency principle also sets information requirements.

The right to access information also gives the data subject the right to obtain information about the personal data used in reaching the decision. However, it does not grant the right to be given an explanation of the decision.

Even though there is no right to an explanation when a decision is not automated, the transparency principle requires that the data controller should give an explanation similar to those given for automated decisions.

### § Other relevant regulations

In addition to the GDPR, there are other regulations requiring that a decision is explained.

For example, the public sector is subject to the Public Administration Act that requires, *inter alia,* individual decisions to be substantiated. The person concerned has the right to be informed of the regulations and the actual circumstances underpinning a decision, as well as the main considerations that have been decisive. (Public Administration Act: Sections 24 and 25).

[13] See for example Sandra Wachter, Brent Mittelstadt and Chris Russel, «Counterfactual explanations without opening the black box: automated decisions and the GDPR».

# Controlling the Algorithms

In the future we will find that more and more decisions affecting us will be made by AI. They may be decisions regarding whether we can obtain a loan, what our motor insurance premium will be or which content our online newspaper shows us. At the same time, it is getting to be more and more difficult to comprehend and gain insight into the complex systems that make decisions on our behalf. So, we are dependent on service providers processing our data in the appropriate manner and in compliance with the data protection regulations.

The Data Protection Authority (DPA) is tasked with supervising organisations in both the private and the public sector and ensuring that they comply with the data protection regulations. But how can an algorithm hidden in a black box be supervised?

## The DPA's supervisory competence

The GDPR establishes the investigative authority vested in that the DPA in connection with its supervisory role. To control whether personal data is being processed in accordance with the regulations, the DPA may conduct an investigation. An inspection shall clarify whether the data controller has routines and guidelines in place designed to ensure compliance with the regulations, and whether the routines and guidelines are followed.

In connection with an investigation, representatives from the DPA may ask for all the information they require to perform their tasks. This might consist of documentation relating to organisational and technical measures, risk assessments, data protection impact assessments, employee training and the ways in which approaches made by data subjects are followed up.

The representatives may also require to be given access to premises, data processing equipment and means, as well as to the personal data that is being processed. Access to premises, data processing equipment and means shall be granted in accordance with the procedural rules that apply nationally. When consulted on the subject of the new Personal Data Act in Norway, the Norwegian DPA proposed that consideration be given to granting the Authority powers of securing evidence similar to those currently wielded by the Norwegian Competition Authority.

## Investigating the use of AI

An organisation developing or utilising AI is bound by the same legal constraints as any other organisation that is processing personal data. In the course of a normal inspection the DPA will check whether the organisation has a basis for processing, whether it has satisfactory internal controls and routines, that risk assessments have been carried out, and that technical and organisational measures are in place in order to protect the data.

There are some areas that may be particularly important to control at organisations utilising AI, such as compliance with the principles described earlier in this report; that data is not re-used for new purposes without an adequate processing basis; that organisations do not process more personal data than they need; that measures are in place to ensure fair treatment; and that the data subjects are informed as required by law.

If an organisation develops AI, it may be relevant to control the nature and quantity of the training data used, as well as how this data is applied during the training process. If an organisation uses an AI-based system, it may be relevant to check whether it tests the results and conducts audits to ensure that personal data is not being utilised in an unlawful or discriminatory manner. It will also be relevant to investigate whether the system has been developed on the basis of privacy by design.

## How deep can an investigation go?

In most investigations it will be sufficient for the DPA to obtain documentation to determine whether the organisation is in compliance with the regulations. An organisation must be able to explain and document, and in some cases, demonstrate, that they process personal data in accordance with the rules. This means that the organisation must know how a system processes personal data and be able to account for this. If an organisation cannot account for how it uses personal data, the DPA is authorised to impose a fine and temporary or definitive ban on processing activities.

If the DPA suspects that the account given by an organisation is wrong or contains erroneous information, it can ask the organisation to verify the details of its routines and assessments, for example by having the organisation demonstrate how their system processes personal data. This may be necessary when, for example, there is a suspicion that an algorithm is using data that the organisation has no basis for processing, or if there is a suspicion that the algorithm is correlating data that will lead to a discriminatory result.

The DPA currently carries out few IT system controls when out on inspection. In some cases where there is a need, the DPA checks what is taking place inside a system, for example by investigating how long a camera recording is stored for. We expect that the need to control IT systems will increase in the coming years in line with the greater use of automated analyses and decision-making in all sectors. Moreover, the Personal Data Act places greater emphasis on the data controller's duty to carry out responsible operations and internal controls, and less emphasis on preliminary controls conducted by the DPA.[14]

## How to inspect a "black box"?

"Ordinary" algorithms are relatively straightforward to deal with. They are programmed to carry out specific actions. If, for example, your income is x and your debts y, you can obtain a loan of z. This is a much-simplified example, but it shows how it is possible to see what the inputs are and how the data is processed in order to obtain a given result.

However, models based on deep learning and neural networks are complex and have low transparency, making it challenging to control what is actually taking place inside the system. Considerable knowledge of AI-based systems is required in order to know what to look for, as well as which questions to ask. In an inspection situation, where we identify a need to delve more deeply into the system, advanced technological expertise will be required.

From a resource utilisation standpoint, the solution may be to hire external expertise in those cases where a "deep" control is required of an AI-based system. It is important that the DPA has both the knowledge and the resources required to discover breaches of the regulations, so as to avoid algorithms that reinforce social differences or lead to arbitrary discrimination, as well as the unlawful re-use of data.

[14] See the guidelines on the responsibilities of enterprises under the GDPR on the Norwegian DPA's web site (in Norwegian), https://www.datatilsynet.no/regelverk-og-skjema/veiledere/virksomhetens-ansvar-etter-nytt-regelverk

# Solutions and recommendations

A data protection principle which underpins all AI development and applications, is accountability. This principle is central to the GDPR and places greater responsibility on the data controller for ensuring that all processing is conducted in compliance with the rules. Data processors are also bound by the accountability principle.

In this chapter we shall present examples of tools and solutions that can help the data controller to comply with the rules. But first, we will discuss two of the requirements in the GDPR that are especially important in connection with the development and application of AI; data protection impact assessment (DPIA) and privacy by design. Following that, we look at tools and methods that can help protect privacy in systems that use AI. Finally, we shall propose some recommendations for developers, system suppliers, organisations buying and using AI, end users and the authorities.

## Assess the data protection impact – and build privacy into your system!

The new data protection regulations will enhance the rights of individuals. At the same time, the duties of organisations will be tightened up. Two new requirements that are especially relevant for organisations using AI, are the requirements privacy by design and DPIA.

### Privacy by design

The data controller shall build privacy protection into the systems and ensure that data protection is safeguarded in the system's standard settings. These requirements are described in Article 25 of the GDPR and apply when developing software, ordering new systems, solutions and services, as well as when developing these further.

The rules require that data protection is given due consideration in all stages of system development, in routines and in daily use. Standard settings shall be as protective of privacy as possible, and data protection features shall be embedded at the design stage.[15] The principle of data minimisation is expressly mentioned in the provision relating to privacy by design.

### Data Protection Impact Assessment

Anyone processing personal data has a duty to assess the risks involved. If an enterprise believes that a planned process is likely to pose a high risk to natural persons' rights and freedoms, it has a duty to conduct data protection impact assessment (DPIA). This is described in Article 35 of the GDPR.

When a risk is assessed, consideration shall be given to the nature, scope, context and purpose of the process. The use of new technology must also be taken into account. Moreover, there is a requirement to assess the impact on personal privacy by systematically and extensively considering all personal details in cases where this data is used in automated decision making, or when special categories of personal data (sensitive personal data) are used in on a large scale. The systematic and large-scale monitoring of public areas also requires documentation showing that a DPIA has been conducted.

The impact assessment should include the following as a minimum:

- a systematic description of the process, its purpose and which justified interest it protects
- an assessment of whether the process is necessary and proportional, given its purpose
- an assessment of the risk that processing involves for people's rights, including the right to privacy
- the measures selected for managing risk identified

The DPA shall be involved in preliminary discussions should an impact analysis reveal that the planned process may represent a high risk for data subjects, and

[15] Read the Norwegian DPA's guidelines on software development with embedded privacy: https://www.datatilsynet.no/en/regulations-and-tools/guidelines/data-protection-by-design-and-by-default/

that the risk cannot be reduced by the data controller (GDPR Article 36).

## Tools and methods for good data protection in AI

Artificial intelligence is a rapidly developing technology. The same applies to the tools and methods that can help meet the data protection challenges posed by the use of AI. We have collected a number of examples to illustrate some of the available options. These methods have not been evaluated in practice, but assessed according to their possible potential. This means that technically they are perhaps unsuitable today, but the concepts are exciting, and they have the potential for further research and future use.

We have placed the methods in three categories:

- Methods for reducing the need for training data.
- Methods that uphold data protection without reducing the basic dataset.
- Methods designed to avoid the black box issue.

### 1. Methods for reducing the need for training data

One of the challenges we have pointed out in this report, is that there is often a need for huge amounts of data during machine learning. However, by selecting the right features, and adjusting them appropriately, the data requirement can be reduced. Here is a selection of methods that can help achieve this:

**Generative Adversarial Networks**[16]
Generative Adversarial Networks (GAN) are used for generating synthetic data. As of today, GAN has mainly been used for the generation of images. But it also has the potential for becoming a method for generating huge volumes of high quality, synthetic training data in other areas. This will satisfy the need for both labelled data and large volumes of data, without the need to utilise great amounts of data containing real personal information.

**Federated learning**[17]
This is a form of distributed learning. Federated learning works by downloading the latest version of a centralized model to a client unit, for example a mobile phone. The model is then improved locally on the client, on the basis of local data. The changes to the model are then sent back to the server where they are consolidated with the change information from models on other clients. An average of the changed information is then used to improve the centralized model. The new, improved centralized model may now be downloaded by all the clients. This provides an opportunity to improve an existing model, on the basis of a large number of users, without having to share the users' data.

**Matrix capsules**[18]
Matrix capsules are a new variant of neural networks, and require less data for learning than what is currently the norm for deep learning. This is very advantageous because a lot less data is required for machine learning.

### 2. Methods that protect privacy without reducing the data basis

The optimal solution would be if one could use as much data as one wished for machine learning, without compromising privacy. The field of cryptology offers some promising possibilities in this area:

**Differential privacy**[19]
Let us, for example, start with a database that contains natural persons and features related to these persons. When information is retrieved from the database, the response will contain deliberately-generated "noise", enabling information to be retrieved about persons in the database, but not precise details about specific individuals. A database must not be able to give a markedly different result to a query if an individual person is removed from the database or not. The overriding trends or characteristics of the dataset will not change.

**Homomorphic encryption**
This is an encryption method that enables the processing of data whilst it is still encrypted. This means that confidentiality can be maintained without limiting the usage possibilities of the dataset. At present, homomorphic encryption has limitations, which mean that systems employing it will operate at a much lower rate of efficiency. The technology is promising, however.

[16] https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[17] https://research.googleblog.com/2017/04/federated-learning-collaborative.html

[18] https://openreview.net/pdf?id=HJWLfGWRb

[19] https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf, https://arxiv.org/abs/1412.7584

Microsoft for example has published a white paper on a system that uses homomorphic encryption in connection with image recognition.[20] Active efforts are also underway to standardise homomorphic encryption solutions.[21]

**Transfer learning**[22]
It is not the case that it is always necessary to develop models from scratch. Another possibility is to utilise existing models that solve similar tasks. By basing processing on these existing models, it will often be possible to achieve the same result with less data and in a shorter time. There are libraries containing pre-trained models that can be used.

**RAIRD**
Statistics Norway (SSB) and the Norwegian Centre for Research Data (NSD) have developed a system called RAIRD[23] that permits research to be carried out on data without having direct access to the complete dataset.

In short, this system works by means of an interface that allows the researchers to access only the metadata of the underlying dataset. The dataset may, for example, be a cancer diagnosis register containing fields for age, gender, date of and place of birth. The researcher can then submit queries based on the metadata and obtain a report containing aggregated data only.

This solution has been designed to prevent the retrieval of data relating to very small groups and individual persons. This type of system can therefore be used when data for machine learning is needed. Instead of receiving a report as an end result, one could obtain a model from the system.

### 3. Methods for avoidiing the black box issue

One of the issues mentioned is the lack of transparency in connection with machine learning and automated decision-making. This represents a challenge both for those using such a system and for the people whose data is processed by it. System developers who base their work on machine learning would derive great benefit from knowing what takes place under the bonnet, as it were, in order to quality assure and improve their products.

**Explainable AI (XAI)**[24]
XAI is the idea that all the automated decisions made should be explicable. With people involved in a process, it is often desirable that an explanation is given for the outcome. There are some interesting possibilities in two areas. There will also be a need to be able to control systems that do not have this embedded. It will probably also be attractive for developers employing transfer learning.

There is also a project underway in this field, being run by the Defense Advanced Research Projects Agency (DARPA), where the objective is to gain more knowledge about providing understandable explanations for automated decisions. They have sponsored Oregon State University, awarding an amount of USD 6.5 million over four years for research into this topic. The goal is to create AI that can explain its decisions in such a way that is understandable and promotes confidence in using the system. There are in any case good grounds for believing that this research will drive the field forward.

**LIME**[25]
LIME is an approach to XAI. It is a model-agnostic solution that produces explanations ordinary people can understand. In the case of image recognition, for example, it will be able to show which parts of the picture are relevant for what it thinks the image is. This makes it easy for anyone to comprehend the basis for a decision.

---

[20] https://www.microsoft.com/en-us/research/publication/cryptonets-applying-neural-networks-to-encrypted-data-with-high-throughput-and-accuracy/

[21] http://homomorphicencryption.org/

[22] http://www.cs.utexas.edu/~ml/publications/area/125/transfer_learning

[23] http://raird.no/

[24] https://www.darpa.mil/program/explainable-artificial-intelligence

[25] https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

# Recommendations for privacy friendly development and use of AI

In the following we propose a number of recommendations for protecting personal data when developing and using AI.

### Recommendations for developers of AI

These recommendations are meant for actors pursuing AI research and development. They will in the main be research milieus in universities and large commercial organisations. These constitute an important target group because they are developing the basic technology that is the basis for further application of AI.

- Conduct research into how intelligent systems can be made more privacy friendly, such as how AI systems can be designed in order to make it easy for users to comply with the regulations. Research can, for example, be carried out on solutions that use less training data, anonymisation techniques and on solutions that explain how systems process data and how they reach their conclusions. Other interesting research areas include how to conduct system audits to ensure the system isn't biased, especially audits by third parties.
- Adopt a multidisciplinary approach. AI is more than just technology. It is important to put together multi-disciplinary teams that can consider the consequences for society of the systems developed. Research can also throw light on the how the use of AI can be of considerable value to society as well as on the problematical areas.

### Recommendations for system suppliers

These recommendations are meant for organisations that use basic technologies developed by others – organisations that use AI in their own projects or in solutions supplied to others. These can be data controllers or merely a supplier of a service or product. Our recommendations are also relevant for research milieus utilising technologies developed by others.

- Get familiar with the GDPR – the duties you have, and the rights and duties of the users of the system.
- Select models that meet the privacy needs of the buyer. For example, not all types of model can explain how they reached a specific result.
- Limit the amount of personal data in the training data to what is relevant and necessary for the purpose.
- Ensure, and document, that the system you are developing meets the requirements for privacy by design.
- Document how the data protection requirements are met. Documentation is one of the requirements of the regulations, and will be requested by customers or users.
- Assist customers by showing how different systems protect personal data, by for example helping to fulfil the duty to provide information, and by showing the customer how to test or audit the system to ensure compliance with the regulations and internal requirements.

### Recommendations for organisations purchasing and using AI-based systems

These recommendations are aimed at organisations purchasing and using IT solutions based on AI technologies. This could be both commercial and public organisations.

- Carry out a risk assessment and, if required, carry out a DPIA before you purchase a system, before you start using it, as well as when in use.
- Demand that the system you order satisfies the requirements for privacy by design.
- Conduct regular tests of the system to ensure that it complies with the regulatory requirements, for example, for avoiding latent discriminatory treatment.
- Ensure that the system protects the rights of your users; for example, the right to demand limited processing.
- Ensure that you have good systems for protecting the rights of data subjects, such as the right to information, to access and deletion. If consent is the legal basis of processing, the system must also include functionality enabling consent to be given, and to be withdrawn.
- Consider establishing industry norms, ethical guidelines or a data protection panel consisting of external experts in the fields of technology, society and data protection. These can provide advice on the legal, ethical, social and technological challenges – and opportunities – linked to the use of AI.

### Recommendations for end users

These recommendations are aimed at end users. An end user is the data subject using a service or whose personal details are processed by using AI.

- **Right to information.** You are entitled to comprehensible and readily available information about the processing of your personal data. This right applies both when organisations retrieve information directly from you, and when it is retrieved from other sources. You shall know what the information is being used for (*the purpose*) and the legal *basis* on which that the organisation is processing the information.
- **Consent**. In many situations the controller must obtain your consent before processing can begin. The data controller is responsible for documenting that proper consent has been given, which means that you have given a voluntary, specific, informed and unambiguous declaration that you approve of your personal data being processed. You also have the right to withdraw any consent you have given previously.
- **Right of access to information.** You have the right to contact organisations and ask *whether* they are processing details about you, and, if so, *what* has been registered. As a rule, you are entitled to a copy of the details registered. There are, however, some exceptions to the right of access to information, for example within the judicial sector.
- **Right to rectify and delete information.** You are entitled to ask for incorrect or unnecessary details about you to be rectified or deleted.
- **Right to object** to the processing of details about you. You may have the right to protest against the processing of details concerning yourself. If you protest against direct marketing, it must be stopped without your needing to provide further grounds. In other situations, you may have to justify your right to object by explaining the circumstances affecting your situation. The organisation must then cease processing, unless they can prove they have compelling and justifiable grounds for processing the data, and that these grounds weigh more heavily than your interests, rights and freedoms.
- **Right to demand limited processing**. If you are of the opinion that some details are incorrect, or are being processed unlawfully, or you have exercised your right to protest against processing, the organisation may be compelled to stop the data being used, but continue to store them until the disagreement has been settled.
- **Data portability**. If, whether contractually or having given your consent, you have had personal data about yourself processed, you can ask for these details to be delivered to you by the organisation in a structured, generally applicable, and machine-readable format.

### Recommendations for authorities

These recommendations are for legislators and political decision makers as they set the terms and conditions for the development and use of AI.

- Ensure that the public sector sets a good example in using AI. This requires acute awareness of the ethical and privacy consequences of the systems they use, as well as expertise as buyers to make sure that the systems purchased have privacy by design and that they meet the legislative requirements.
- Allocate funds to research that ensure the technology processes the personal data in compliance with the regulations. Protecting personal data is not just a legal requirement, but can also be a competitive advantage for Norwegian industry.
- Ensure that the enforcement authorities possess the relevant expertise, and arrange for experience and knowledge sharing across sectoral boundaries.
- Ensure that the law keeps apace with technological developments. This applies to all legislation that has relevance for the use of personal data.


**The Norwegian Data Protection Authority**

**Visiting address**:
Tollbugata 3, 0152 Oslo, Norway

**Postal address:**
P.O. Box 8177 Dep.,
0034 Oslo, Norway

postkasse@datatilsynet.no
Telephone: +47 22 39 69 00

**datatilsynet.no**
personvernbloggen.no
twitter.com/datatilsynet