AI & Algorithmic Risks Report Netherlands

Edition 3, summer 2024

Autoriteit Persoonsgegevens | Department for the Coordination of Algorithmic oversight (DCA)

Periodic insight into risks and effects of the use of AI and algorithms in the Netherlands

AUTORITEIT
PERSOONSGEGEVENS

Table of contents

Key messages

1. The AI risk profile continues to call for vigilance from everyone – from Ministers to citizens and from CEOs to consumers – because (i) it is difficult to assess whether AI applications are sufficiently controlled and (ii) AI incidents can occur more and more frequently, especially as AI is increasingly becoming intertwined into society.

Consequently, these are still stormy times and this is understandable with the emergence of a new system technology. A year ago, the observation was that the Netherlands had to take steps to get a grip on algorithms. Meanwhile, the tumultuous growth of AI technology continues. In addition, the emergence of generative AI provides an incentive to experiment on a large scale with new AI applications. In the coming years, AI will become increasingly deeply intertwined with elements of society. This results, both in terms of size and nature, in more and newer risks that are still difficult to assess. The long-term effects of which are also not yet fully understood. Overall, the international policy response has been decisive so far. It focuses both on traditional supervision and on new forms of testing and controlling of for example, the safety of AI systems and in combating new cybersecurity risks. At the same time, setting up AI regulation is a long-term process. This means that organisations – and society as a whole – must continue to prepare for future AI incidents. And as long as organisations still have doubts about their risk management, this calls for restraint in the use of AI systems. This concerns both systems based on simple static algorithms and complex self-learning AI, both of which the AP classifies as an 'AI system'.

2. Many new AI systems and risks (or possible risks) stand out. From experimentation by big tech companies to the widespread use of AI in situations where people are vulnerable.

It is striking that big tech companies want to bring new applications based on generative AI to the market as quickly as possible. This has often been accompanied by mistakes and vulnerabilities in the past period that led to emergency repair or even the withdrawal of systems. Fundamental rights such as non-discrimination and data protection may be at stake and there are legality issues in relation to existing regulations such as the GDPR, copyright and consumer protection. Meanwhile, Dutch organisations are fully exploring the further possibilities of more classic AI systems, for example for behavioural monitoring via camera analysis, employee management, risk selection in the social domain and advice on the appropriate level of education for children. It is important that information about these systems becomes increasingly available through public algorithm registration. This is the basis for proactive transparency and oversight. See also Chapter 1.

3. Information provision is essential for the functioning of democracy, but is under pressure from the deployment of AI systems. This applies to both moderation and distribution of content and, more recently, to content creation with generative AI.

The use of AI systems affects the online provision of information on a large scale. Generative AI makes it possible for malicious parties to generate disinformation on a large scale. In addition, generative AI has inherent technological weaknesses, which also contribute to misinformation and discriminatory and stereotypical content. Furthermore, disinformation and misinformation have a major impact on public debate and the Dutch are very concerned about this. Verifying the origin and 'authenticity' of content is therefore a critical link in both being able to trust content and being able to deal with its effects. The information on offer is simply too overwhelming, which makes the use of filtering

what is on offer necessary. However, this moderation is largely in the hands of big tech platforms and this could jeopardise a diverse range of information. The European Digital Services Act demands very large online platforms, among other things, to provide openness about moderation and to tackle disinformation. It is striking that the number of Dutch-speaking moderators on these platforms is decreasing. Because the use of AI (and generative AI) in the online provision of information influences the public debate on a large scale, a common 'information base' is needed to counter polarisation. The extent to which the role of AI systems actually affects the functioning of democracy at present (or in the future) is difficult to measure, which makes it important to keep a finger on the pulse through active monitoring and analysis. See also Chapter 2.

4. Conditions for adequate democratic control of AI systems are currently insufficiently met.

The design of the process for democratic steering and supervision of AI systems determines the way in which representatives of the people – from the House of Representatives to the municipal council – can have control over AI systems used by the government. This guidance and control should be possible at every stage of the development, deployment and evaluation of an AI system. This report explores this topic on the basis of the situation in local government. Within the public sector, decentralised authorities use most AI systems. Democratic control of public AI systems is carried out by the representatives of the people, together with the court of auditors, the ombudsperson and the media. However, these authorities have limited capacity and expertise at their disposal. This complicates their supervisory role. Survey results show that municipal organisations have limited oversight of their AI systems, that council members have doubts about the adequacy of their AI knowledge, and that only a few local audit institutions conduct sporadic research into AI systems. Nationally, investments are desirable in a supporting infrastructure for national and local actors, for example via an AI coordination centre or via AI centres of expertise. Both to strengthen the responsible use of AI, as well as the democratic supervision. See also Chapter 3.

5. Random sampling is a valuable tool to reduce risks in profiling and selecting AI systems.

Many organisations use algorithms for risk profiling or similar processes that distinguish between people and this entails fundamental rights risks. This report explores this topic using examples in the field of fraud detection.

It is important to always see these algorithms as part of a broader process. Virtually everyone is subjected to these fraud detection algorithms in different places in society. In addition to legality issues – such an algorithm may be used in certain situations and certain indicators may be used – it is an essential point of attention that errors in these algorithms have a major impact. Discrimination and over-reliance on the fraud algorithm are two key risks, and to counteract them, embedding a random sample in the fraud detection process can help to monitor discrimination risks. The random sample also contributes to measuring efficiency and exploring new types of fraud. The design and operation will vary by context but in many cases it is a measure worth considering when using an AI system for profiling and selecting AI systems. See also Chapter 4.

6. The entry into force of the AI Act (early August 2024) is a milestone, with concerns about (i) the long transition period (up to 2030) for existing high-risk AI systems within the government and (ii) whether robust and workable product standards will be in place in a timely manner.

Some provisions under the AI Act enter into force as early as 2025, for example for prohibited AI applications and AI literacy within organisations. So there is work to be done here, noting that AI applications that will soon be banned under the AI Act may already be in conflict with other legislation, for example the GDPR. The AP emphasizes that the product standards must be completed under high time pressure. Timeliness is of the utmost importance but should not be at the expense of the content. The product standards are decisive for the actual effectiveness and practicability of the AI Act. In the meantime, supervisors in the Netherlands are working on preparing for new supervisory tasks under the AI Act. This has also led to initial recommendations to the government. See also Chapter 5.

7. With regard to the further elaboration of the coalition agreement, the AP advises to continue to give priority to algorithm registration by government organisations and to discuss registration by semi-public organisations.

The main principles of the coalition agreement contain important provisions on algorithms and AI that can strengthen current policies. For example, on the use of a scientific standard for the use of models and algorithms. It is important to see these requirements as part of the provisions of the AI Act and the further elaboration in product standards. For example, to prevent proliferation in standards (see next section). In the short term, it is important that algorithm registration remains a priority. The AP remains in favour of making it soon mandatory for government organisations to register algorithms. The AP also stresses that the scope of such a register must be sufficiently broad and that it must be seen in conjunction with the European registration obligation for high-risk AI systems under the AI Act. This trade-off is not always so simple, even if a particular AI system (or algorithm) is a high-risk system or not.
The focus is on the use of AI systems by organisations, in for example, healthcare, education, public housing and public transport. This is an essential service but insight into the use of AI systems in this sector is cloudy. See also Chapter 5.

8. The AP is committed to increasing the control of AI systems, in which (i) a proliferation of frameworks should be avoided and (ii) a recalibration of the national AI strategy can contribute to the further ecosystem for development and control of AI systems.

There are various policy developments at home and abroad that contribute to identifying and reducing AI-related risks, such as the establishment of AI safety institutes, cooperation between supervisors and policy makers and the setting up of AI Advisory Councils. This can contribute to timely and harmonised action when new risks arise. And it provides guidance for responsible development and deployment of AI systems. At the same time, there is a risk of an abundance of initiatives and frameworks. It creates confusion or can even be harmful – for example through both intentional and unintentional ethics washing – if frameworks are not concrete enough or can be interpreted in multiple ways. In addition to the legality issues, the AP makes an effort, through the coordinating task, to provide guidance in the control of AI systems.

This is being done precisely to support promising and responsible applications. Specific attention is paid to the upcoming AI Act. The government can contribute to strengthening the entire AI ecosystem by recalibrating the national 2019 AI strategy, through maintaining the positive aspects while at the same time paying attention to the new challenges posed by the more complex, modern AI systems. See also Chapter 5.

‘AI system’ broadly defined

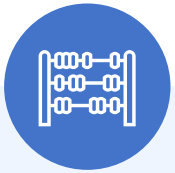
Recently, there has been a consensus on the meaning of the term ‘AI system’. The term AI system is included in the AI Act and is based on the OECD global recommendation. Briefly worded, an AI system can – for explicit or implicit purposes – infer from input received how to generate output. Like making predictions, generating content, making recommendations or making decisions. This output affects the physical or virtual environment. AI systems vary in their degree of autonomy and adaptability when deployed (after the development phase).

Based on current understanding, the term AI system refers to a wide range of applications, from simple (static) algorithms to complex AI or self-learning AI. In a recent explanation of the recalibration of the definition, the OECD explains, for example, that the model that forms the basis of an AI system (there could also be several models), can be either manually built by human programmers or automatically created, for example by unaccompanied, guided or self-reinforcing machine learning techniques. The OECD emphasizes in the explanatory memorandum that model adjustments are often part of the development phase and that the model is usually fixed during deployment. Different AI systems vary in their level of autonomy and adaptability in deployment. Traditional and simple software development, in which people have fully determined which rules are applied, falls outside the scope of the term.

The term AI system covers both applications that ‘only’ make recommendations to people and applications that can make drastic decisions on their own. It is often the context and the embedding in a broader task or objective that determines the form of the output of an AI system. An example is a support system for car drivers. Such a system can ‘predict’ that another car is nearby, then ‘recommend’ the driver and warn him to brake. But the system can also be used in such a way that it independently ‘decides’ to brake. In both cases, there is an AI system. The knowledge needed to deal responsibly with these systems, to be aware of their risks and effects and to make responsible decisions about their deployment is called ‘AI literacy’.

Guidelines from the European Commission will provide further clarification. Guidelines will be published on the practical application of the definition of an AI system. This will ensure further clarification, for instance through examples.

Overview Risk profile AI & Algorithms Summer 2024



Which AI systems stood out?

- **Tech innovations.** Impetuous and incident-rich launches of generative AI systems by big tech companies.
- **Behavioural monitoring.** Targeting customers and visitors via cameras in supermarkets, gyms and public transport.
- **Algorithmic management.** A management system for road authorities at Rijkswaterstaat.
- **Housing.** Use of AI and scraping for detection of housing fraud by housing corporations.
- **Testing in education.** The attainment test for primary school leavers, with an important role for adaptive testing.
- **Public services.** A system for filtering requests for social security through machine learning.

What are the noticeable risks?

- **'Rat race' in tech.** Big tech companies are striving for rapid market dominance in AI. Quality standards and risk management are under pressure.
- **Abuse of generative AI.** Technology is a panacea for malicious people. Risks such as cybersecurity are on the rise.
- **Provision of information threatened.** The use of AI systems in the production, moderation and consumption of online information has an impact on diversity and reliability – possibly with a major impact on public debate.
- **Democratic control over AI.** Governments are not sufficiently equipped to control the deployment of AI systems in the public sector. Incidents may therefore be noticed too late (or not at all).
- **Discrimination in AI systems for selection.** Profiling and selection systems, for example for fraud detection, are still under public scrutiny and detecting discrimination is often difficult.
- **Timeliness of detailed regulation.** It is questionable whether clear and solid product standards under the AI Act will arrive on time.
- **Long transition periods.** There is a transition period until 2030 for existing high-risk AI systems in the public sector.
- **You cannot see the forest for the trees.** There is a proliferation of frameworks and standards.

What needs to be done?

- **European approach to generative AI.** Going forward with standards for generative AI and striving for global convergence.
- **AI safety institute.** Exploring whether this can be set up, and how, in connection with existing supervisory tasks.
- **Think before you act.** To counteract, in a broad sense, an unsurgical urge to experiment. Responsible use of AI requires due care.
- **AI literacy.** Relevant to every citizen. The basis for understanding AI systems and being aware of their impacts and risks.
- **Traceability of information.** In a world with AI, the origin of the information needs to be known in order to be able to verify it.
- **Investing in democratic control.** Support knowledge, capacity and processes in local government, so that they use AI systems responsibly and these can also be controlled.
- **Random samples.** A useful validation tool for selecting systems that can be embedded in the work process.
- **Mandatory algorithm registration.** Maintain registration of algorithms as a priority for public organisations and make it mandatory, with possible extension to the non-commercial service sector.
- **AI strategy.** The tumultuous development of AI calls for a reassessment of the national AI strategy.

1. Overarching developments

QUICKLY TO THIS SUBJECT

1.1 Main points

Risk management of AI systems is not progressing at the same pace as the development and deployment of the technology. Policymakers, administrators and society must face this reality. This does not mean that we have to accept all future incidents in advance but that we have to prepare for them.

New technology is best adapted when it is still in its infancy. Once the technology has been further developed and is already fully deployed, it is much more complicated to achieve appropriate risk management. Just as requiring cars to have safety features in advance, such as an airbag, has more effect than doing so afterwards, if when everyone already has a car without an airbag. Then it takes a disproportionate amount of time and costs to adjust and the situation is already unsafe.

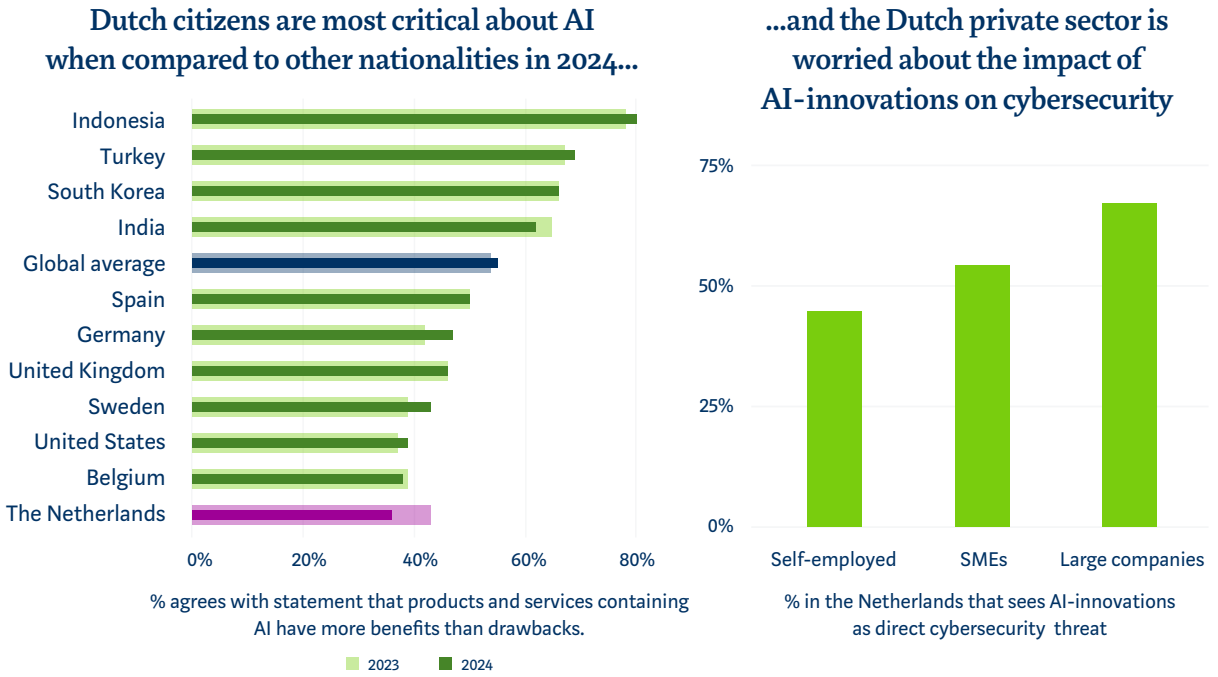
This is cross-border technology, so there must be consensus on risk management at least within Europe, but preferably worldwide. Without clear principles, regulations, standards and social norms, risk management will become increasingly difficult.

Trust in AI systems among Dutch citizens is comparatively low and concerns about some AI risks in business are growing. More than half of people worldwide respond positively to the statement that products and services with AI have more advantages than disadvantages. In the Netherlands, this number drops to 36%. A year ago, it was 43%. The Netherlands is at the bottom of a list of 32 countries. This is shown by an annual global AI monitor from Ipsos. A concern that is receiving more attention within the Dutch business community is the impact of AI innovations such as, generative AI technology, on cyber security. Research by ABN Amro, in collaboration with MWM2, shows that more than 50% of Dutch companies have these concerns. A year ago, it was less than a quarter. The biggest concerns are with large companies. See also Graph 1.1.

1.2 Turbulent growth of AI technology

Many investments in AI technology will ensure the further development and spread of AI in the coming years. Generative AI in particular has seen an explosive rise in venture capital investments over the past year and its centre of gravity is in the United States. New companies are emerging, but established organisations in particular are helping the general public to further embrace new AI systems. For example; Microsoft is working on the integration of language models into commonly used Office packages, Google

GRAPH 1.1: PERCEPTION OF AI-RISKS AMONGST DUTCH CITIZENS AND COMPANIES



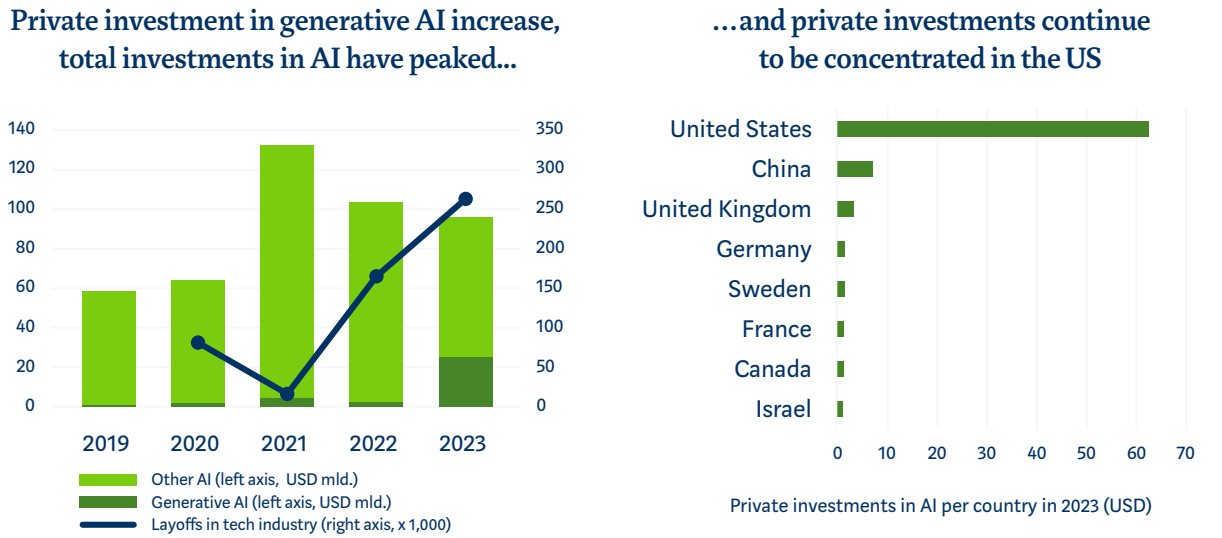
SOURCE (LEFT): IPSOS AI MONITOR (N = 23.685, 32 COUNTRIES)
SOURCE (RIGHT): ABN AMRO (2024) IN COOPERATION WITH RESEARCH AGENCY MWM2

recently overhauled the most widely used search engine to deploy more AI, and Apple is working with OpenAI on the integration of language models into Apple's operating systems. According to forecasts by US bank Morgan Stanley, AI PCs will take up the majority of the market as early as 2028.[2]

However, not all expectations of growth are being met. Research by consulting firm BCG shows that the use of generative AI can improve employee performance but also reduce it. This depends on the task.[3] Whereas a year ago there was sometimes a perception that generative AI can help everywhere, a more realistic picture of opportunities in specific application areas is now emerging. In addition, total investment in AI fell further last year and the peak was in 2021. At that time, the technology sector was boosted by various circumstances, including the accelerated digitalisation caused by the coronavirus pandemic. The end of the pandemic in 2023 coincided with the adjustment of too high expectations within the technology sector. This was reflected, among other things, in many redundancies in that sector.

In the competition that exists between large tech companies, new technology is often launched headlong. For Google, more than half of its revenue comes from search advertising. Imagine OpenAI developing a more user-friendly search engine, for example through collaboration with the Microsoft Bing search engine. In such a scenario, Google's business is at stake. That is why tech companies are in a hurry to launch new products quickly. They see exposure of the product to large parts of society as an experiment. For example, Google recently launched the new search engine AI-Overviews. When it received a lot of criticism, Google responded to this a few weeks later in a blog. Among other things, the search engine gave advice on the number of stones a person should eat per day.[4] Microsoft announced a large-scale introduction of Recall, a tool that can retrieve everything you have seen on the screen. After controversy over the security of this tool, Microsoft rolled back the launch.[5] OpenAI recently took Sky's voice offline. This voice was launched a week earlier for communication with ChatGPT. Sky's voice sounded too much like the voice of actress Scarlett Johansson, with whom OpenAI appeared to have been in hasty contact with, in the days before the new release.[6]

GRAPH 1.2: MARKET INDICATORS FOR INVESTMENT AND DEVELOPMENT OF THE AI INDUSTRY

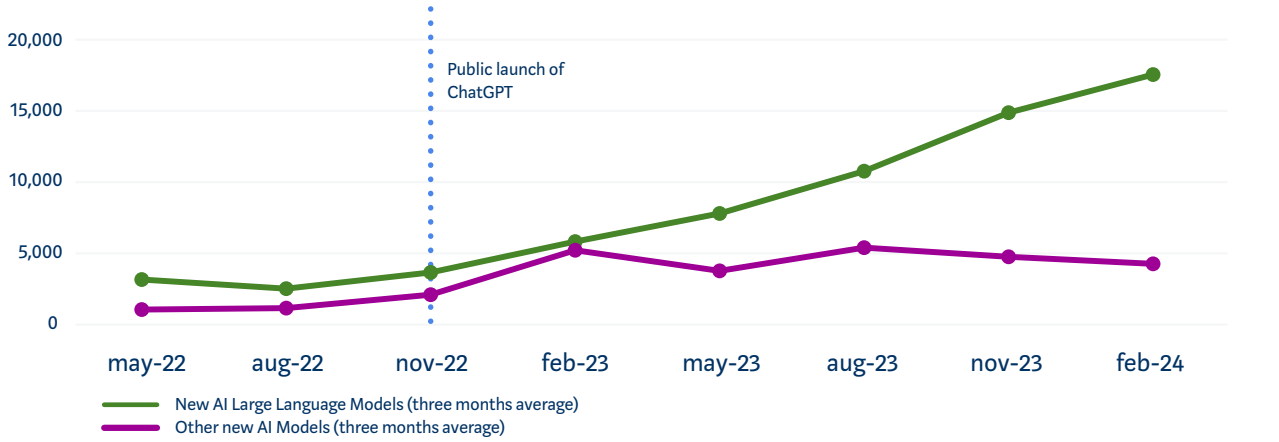


SOURCE: STANFORD UNIVERSITY, 2024 AI INDEX REPORT AND LAYOFFS.FYI

1.3 Developments in general-purpose AI

A lot of attention from developers focuses on covering up fundamental weaknesses of language models. The main basis of language models is the transformer architecture.[7] This model form can supplement text with following words based on patterns that the model has come to know from existing texts. State-of-the-art models generate text that sounds completely natural. However, the models do not reason with all the knowledge about the real world. This poses fundamental challenges. For example, language models can generate text with a completely illogical meaning, while the model is certain about the combination of words. Many efforts by developers are now aimed at allowing the language models in the response to make use of additional information, which is verifiably correct.

GRAPH 1.3: SINCE EARLY 2023, SUPPLY OF LARGE LANGUAGE MODELS HAS OUTPACED OTHER TYPES OF AI MODELS



SOURCE: OECD.AI (2024) ON THE BASIS OF HUGGING FACE DATA (22 OF APRIL 2024)

One method of using verifiable information is to fine-tune models based on human knowledge. This can be done, for example, by showing carefully prepared question-answer combinations to a model. Or by having a model learn from scores that people give to generated outcomes.[8] Another method to improve the results is to give more context to the question. For example, by including the content of an authoritative source in a question, a language model can be directed to base answers on that content. A next step in this is automating the search for and adding the right authoritative sources as context for a question. This is called a Retrieval Augmented Generation (RAG) system.

New AI models, such as for generative AI, are a so-called frontier technology and therefore difficult to predict. Frontier technologies are technologies that are at the intersection of scientific breakthroughs and implementation in society. AI is the umbrella term for one such technology that has emerged from science and, after a number of breakthroughs, is being applied in practice, such as recently generative AI. This rapid development and adoption offers us only a first glimpse of what things this technology will make possible to us and will mean in society. As with many developments in technology or society, the focus is primarily on applications that demonstrate the power of technology. For example, how generative AI can generate realistic material. But the focus is also on applications that bring convenience to people and society or that are particularly valuable, such as in the medical sector. There are also surprising applications, which could not be foreseen or predicted in advance.

AI systems are also used for malicious practices, such as social engineering scams. By focusing on certain elements of trust, scammers can operate more easily. Where messages or e-mail traffic can still be distrusted, many people do rely on their senses. People quickly trust an image of a person or a digital conversation, provided that the image is credible enough. This shows, for example, a recent incident in which a multinational company transferred more than $ 25 million to fraudsters after an employee was invited to a video call with a deepfake version of the CEO. Bona fide applications can be regulated on the basis of benevolent actors. Malicious applications, on the other hand, are much harder to get a grip on. Generative AI has already been used extensively for this purpose, for example by making deepfakes for scams, pornography or vengeful pornographic images. The opportunities of frontier technology can also be expressed in threats and CEO fraud is a good example of this.

1.4 Rise of AI safety institutes

In response to the awareness of the risks of generative AI, several AI safety institutes have been launched in the past six months. An AI Safety Institute has been established in the United Kingdom and the United States. In Europe, there is the AI Office, which stems from the European AI Act. An important objective of these AI safety institutes is to expose the dangers of AI language models and AI language models, by testing and evaluating them in different ways.[9] Sometimes these tests take place before AI language models become publicly available. The work of these institutes aims, among other things, to investigate the possibilities for malicious users to misuse advanced AI technologies. The policy approach and strategies for these types of institutions are still in full development. Due to the fact that advanced and multifunctional AI models have a direct global impact, it is imperative that AI safety institutes work together as much as possible. An important step in this is that a large group of countries have recently indicated that they want to work on

interoperability AI safety institutes work.[10] For example, by developing interchangeable test systems and comparable assessment frameworks. This enables knowledge sharing and a common foundation for testing.

Large language models are still vulnerable to the most basic circumvention techniques. This is shown in an initial study by an AI Safety Institute into the protective measures in large language models. The purpose of these protection measures is to prevent malicious users from abusing the system to obtain or generate sensitive and harmful information circumvention. Such as confidential information and/or personal data. But also sensitive information about, for example, cyber security or terrorist issues. A first test by the British AI Safety Institute of four prominent language models shows that these are all very vulnerable to simple circumvention techniques. And that with more advanced techniques, it is almost always possible to bypass the model limitations at least once every five attempts.[11] The model then provides answers to questions for which it has received instructions not to answer them.

AI safety institutes should also look at more risks than just misuse. For example, by paying attention to the socio-technical impact. A finding of the U.S. National AI Advisory Committee is that assessing and testing AI safety should go beyond just assessing (technical) model vulnerabilities, because AI technology is deeply embedded in society. AI systems are part of broader processes and are used by people. The safety of AI technology must therefore also be viewed from a broad sociotechnical perspective.[12] Think of how doctors, civil servants, teachers or judges assess cases in practice and build on suggestions made or partly made by generative AI. However, evaluations of this kind are still limited. So far, most assessments of advanced AI language models have had a technical impact. The suggestion is that sociotechnical evaluations offer scope to also look at how people deal with this type of AI system and, for example, how bias and discrimination are overcome. This also requires pilots, phased implementation and impact studies. The American AI Safety Institute gives substance to this broad interpretation in the recently adopted strategy of this institute.[13]

A policy approach in which AI safety is actively tested as part of the public task can also be considered in the Netherlands – this is closely linked to supervision based the AI Act. Supervision on the basis of the AI Act ensures compliance of individual systems with product standards. From an AI safety task perspective, broader and comparative research can be done, for example by testing. This identifies risks and guides the further development of standards and the rules in the AI Act. Setting up an AI safety task at national level also contributes to the cooperation that will be needed with the European AI Office and AI safety institutes in other countries. If the Netherlands has this knowledge and skill, it will also contribute to an ecosystem with international appeal for AI developers.[14]

Box 1.1

AI offers opportunities for people with disabilities

The rise of AI brings more and more initiatives which can enable people with disabilities to participate independently in society. An example is AI glasses that can describe the environment to a visually impaired user. For example, an obstacle on the street or the text on a product in the supermarket. A recent study by the Organisation for Economic Cooperation and Development (OECD) provides an overview of 142 of these types of AI applications for people with disabilities.[15]

AI applications offer opportunities for the inclusion of people with disabilities. For example, the UWV (the Dutch Employee Insurance Agency), in collaboration with various employers, has tested an AI app for people with voice problems and a speech recognition system (with machine learning) for the deaf and hard of hearing. Based on the pilots, the UWV states that the technology helps to increase the employment rate of people with disabilities, as well as their job satisfaction and autonomy.[16]

AI also offers opportunities to help people with disabilities participate in the democratic process. To this end, the EU has, for example, launched the research project iDEM, with the aim of developing AI language models that make information on public affairs more understandable. It also examines whether language models can support people with intellectual disabilities to express their opinions.[17]

At the same time, people with disabilities are often at risk of being discriminated against by AI systems. The Special Rapporteur on Disabled Persons of the United Nations has warned society about this. Examples include face detection software that does not recognise people with facial abnormalities or banking AI systems that see incorrect capitalisation in the written loan application as an indicator of poor payment behaviour – the latter will mostly affect people with dyslexia. The message is therefore to be alert to the risks of AI systems for people with disabilities. A recommendation is to explicitly take limitations into account in the development of AI, for example by actively involving people with disabilities in the development process.

The AI Act contains accessibility requirements. Providers of high-risk AI systems should ensure that those systems comply with accessibility requirements. Those requirements concern the way information is provided, but also the user interface and functionality.

1.5 Domestic lessons and developments (Netherlands)

Generative AI is definitely making its way into various public organisations, for example in healthcare. For example, according to the government-wide vision on generative AI, the government wants to 'facilitate knowledge sharing about the possibilities for the safe use of generative AI by sharing knowledge and practical experience.'[18] For the time being, the public sector seems to use generative AI mainly in experiments and pilots. This at least applies to the healthcare sector.[19] Examples from that sector also show how diverse generative AI is being applied. Healthcare institutions are using generative AI for administrative work, communicating with patients and summarizing patient records.

Every field of application requires context-specific regulation in addition to generic regulation for responsible AI use. After all, each field of application has its own existing standards. In the financial sector, for example, the use of AI should not jeopardise the financial soundness and integrity of financial institutions.[20] Furthermore, the protection of public values and fundamental rights requires different practices in different contexts. Sector-specific standards can complement general laws that protect fundamental rights, such as the AI Act and the GDPR. According to DNB and the AFM, there are still very few sector-specific standards for AI use in the financial sector. However, there is a growing need for sector-specific regulation of AI use.[21]

There is a lot of experimentation with camera systems to recognize undesirable behaviour in, for example, shops, gyms or public transport. This is relevant because behavioural monitoring can affect emotion recognition. The use of such applications in the workplace and in education will become a prohibited application under the AI Act from February 2025. In other contexts, emotion recognition is a high-risk application within the AI Act. However, many current experiments use AI-enabled cameras mainly to monitor visitors or users. For example, a supermarket chain announced the use of a camera system that should be able to recognize possible theft. A gym chain plans to use a system that could detect aggressive behaviour, emergencies, overcrowding, and visitors without a membership.[22] Also in public transport, a pilot is currently running with a camera system that should recognize aggressive behaviour at Amsterdam Central Station.[23] Such systems may qualify as high-risk emotion recognition systems under the AI Act, but are in many cases also unlawful under the GDPR.

The Netherlands has a leading role in the development and deployment of responsible AI systems and associated supervision. Due to major incidents in recent years, risk awareness among Dutch policymakers is high when developing responsible use of AI. According to the Global Index on Responsible AI, the Netherlands[24] is at the top of the list of responsible users of AI. At the same time, incidents, risks and negative effects in data chains still regularly occur and the ability to learn from previous problems does not yet seem to be in place everywhere. This not only shows that the right road has been taken but also that continuous attention is needed to manage risks from AI systems and seize opportunities. For example, assisting people. New products for people with disabilities use AI to improve quality of life (see Box 1.1).

The effects and obligations of deploying an AI system are rarely one-dimensional. This is visible, for example, in the system used by Rijkswaterstaat to optimise the use of road authorities in the event of incidents. This system advises the reporting centre on available road authorities and the shortest arrival times for emergency services in the event of an incident. The system also advises where road authorities can best be positioned for optimal coverage and minimum arrival times. However, this system has caused a stir among employees, because it would allow for monitoring of behaviour and would impair the autonomy and knowledge of road authorities.

According to the Netherlands Court of Audit, Rijkswaterstaat's lacks control over its AI system. Specifically, on the model quality, measures and privacy safeguards, the Court of Audit found that the system does not meet set requirements or criteria. The opinion is that Rijkswaterstaat does not have sufficient insight into the accuracy of the model. As a result, the organisation does not know how much the deployment of the AI system contributes to the faster handling of an incident.[25] It is remarkable that – in addition to insufficient privacy safeguards, which are mandatory under the GDPR – the points that score insufficiently largely correspond to the points from additional regulations for high-risk systems that the AI Act will soon make mandatory. That is why it is all the more important to get the control of this AI system in order.

Many organisations are deploying new forms of fraud detection with algorithms. The risks of this have been highlighted in particular by the childcare benefits scandal.[26] In the social security domain, for example, housing corporations use algorithms to detect housing fraud.[27] Due to the importance of housing, this is a high-risk application for the fundamental rights of the individual. One point of attention is that such a fraud algorithm is a high-risk system under the AI Act. This entails specific requirements for control and transparency. Chapter 4 elaborates on the risks of fraud algorithms and a measure to control them.

The new attainment test for primary school leavers has caused a stir this year. It is remarkable that many of these tests are AI systems.[28] Ten years ago, digital, adaptive tests were introduced in the Netherlands. Schools were given the ability to choose between different test providers.[29] With an adaptive test, each student gets a test with individualized questions. The questions gradually become more difficult or easier, until the AI system has sufficient certainty to come to an opinion about the level of the student. These attainment tests have been applied in a new way this year. A new standard for the various tests should make them comparable, by means of mandatory standard questions. At the same time, from this year onwards, schools are obliged to comply with the test results when the advice based on the transfer test (e.g. pre-university education) is higher than the preliminary school advice (e.g. senior general secondary education). Schools are now only allowed to deviate from the advice on a comply or explain basis. In response to the first transfer test, the umbrella organisation in primary education (PO raad) indicated that schools could not make much sense of the test results and that the results did not appear to correspond to the results in the student monitoring system.[30] The Ministry of Education, Culture and Science has since indicated that this year differences between the tests have surfaced, that were less clear previously. That is why the Ministry, together with the Board of Testing and Examination (Commissie Toetsen en Examens), is investigating the cause of these differences. Attention

is also being paid to the adaptive nature of the attainment tests.[31]

A topical issue in this type of adaptive testing is the extent to which current or future requirements for AI systems are met. AI systems to determine the appropriate level of education, including which system children should have access to, are high-risk applications under the AI Act. This entails obligations, for example on how transparent the test results should be. It is striking that some adaptive attainment tests provide only limited information to children. For example, in some tests, students are not shown which questions they have answered correctly or incorrectly. The ability to keep control is also limited in some adaptive tests. For example, students cannot go back to previous questions if they want to improve them. Other relevant requirements are, for example, that the system predicts with sufficient precision and that it does not discriminate. The AI system must also be used in such a way that a human controller, such as a teacher, can decide not to use the outcome of the AI system. The relevant question is how this relates to the mandatory character that the test has been given in the event of deviations upwards.

1.6 Steps in quality and risk management

AI systems are becoming increasingly intertwined in even more processes and systems in our society. There is a growing awareness that the deployment of these systems often has an impact beyond their mere application. For example, when replacing a human assessment in a process. Or when further automating checks with an AI system. This often has effects that go far beyond just the assessment or only the control act. It affects other elements of a process or chain. These are changes or effects that are often not mapped out in advance and are rarely noticed.

It is understandable that organisations are looking for guidance to be able to control these changing processes or chains. Innovation also has to go hand in hand with responsibility. Legislation and regulations offer more and more tools to set up control. However, as the effects are diverse, so is the accompanying legal framework and rarely does an organisation that deploys AI systems have a single legal framework to adhere to. It will almost always be an interplay of general and specific or sectoral legislative frameworks but the roles are also more diverse than is often thought at first glance. Developers are not always one and the same party. For example, if a developer bases applications on a foundation model of a third party. At the same time, an organisation that deploys an AI system can also be or become the developer itself and have multiple roles at the same time.

A conclusive model for adequate control of all AI systems is not possible or useful, given the speed of development and innovation. However, it is clear that steps are being taken by organisations that invest in innovation and control. Current applications will only grow in complexity, both technologically and in the multiplicity of actors and scale of application. Applications that are now visible are just the tip of the iceberg. The challenges of controlling current systems need to be tackled quickly, so that the backlog in control does not make new applications impossible. This requires a joint and holistic approach, from multiple legal frameworks, but also from supervision, business and government. And above all from the common desire to be able to take advantage of the opportunities offered by AI systems in a responsible manner.

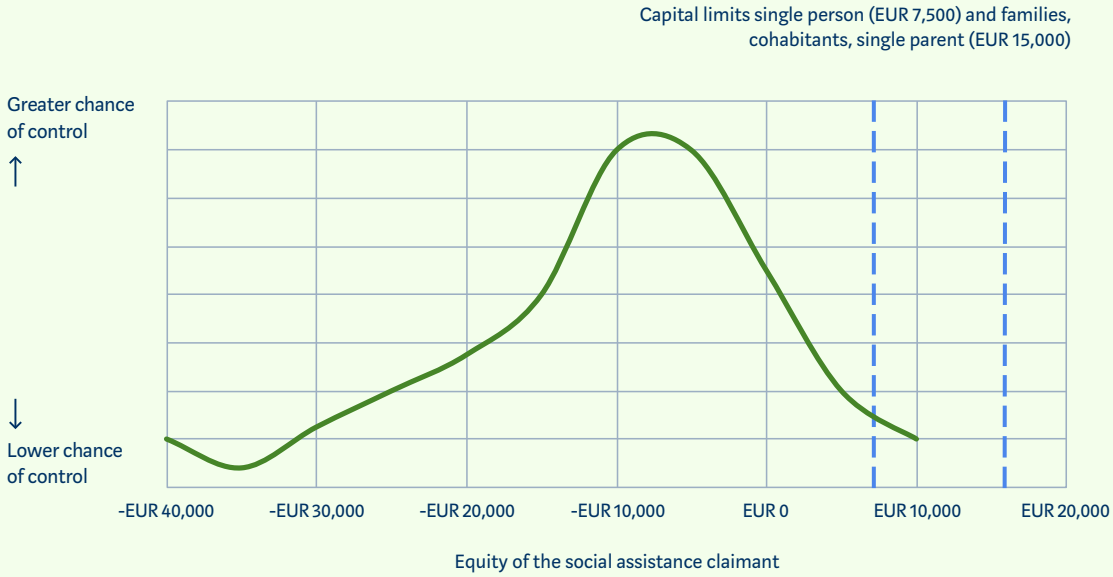
Box 1.2

Challenges to fairness in using machine learning, without arbitrariness and explainability

Practical initiatives illustrate very well how difficult it really is to use AI systems responsibly. An example is the 'Slimme Check' (smart check) of the municipality of Amsterdam. This is an algorithm that indicates whether applications for social benefits are worth assessing for illegality. The Amsterdam algorithm register shows how this algorithm would make the assessments more equal and effective. The technical documentation, process overviews, social considerations and bias analysis are publicly available. However, the uncertainty about the fairness of the algorithm turned out to be too great and that is why the city council decided not to use the algorithm as a precaution.[33]

The complexity and ethical dilemmas with such an algorithm become clear when you look at the importance of the 'total capacity' indicator in the research worthiness of an application. People with too much money are not entitled to welfare benefits. However, the data show that, according to the algorithm, someone with a debt of around EUR 8000 is more research-worthy than someone without debt. There may be a statistical explanation for this, but it is probably hard to accept for the people subjected to an assessment. Through progressive insight gained from these kinds of practical initiatives, more and more standards and quality frameworks are emerging. Chapter 5 discusses this further.

GRAPH 1.4: THE USE OF MACHINE LEARNING IN PRATICE: BASED ON THE LAW, THERE IS A MAXIMUM EQUITY POSSIBLE IN ORDER TO RECEIVE ASSISTANCE, BUT MACHINE LEARNING IS ENGINEERED TOWARDS RESEARCHING NEGATIVE EQUITY.



Explanation: *This graph is a simplified and stylised representation of the documentation on the 'Researchworthiness Algorithm Smart Check', an AI system set up as a pilot by the municipality of Amsterdam. Through machine learning based on historical data, if all other circumstances remain the same, the risk model assigns a greater chance of research worthiness to assistance applicants with a negative equity (i.e. a debt) of around 10,000 euros than to assistance applicants with a positive equity or a negative equity of, for example, 20,000 or 30,000 euros.*

SOURCE: CITY OF AMSTERDAM ALGORITHM REGISTER

In addition, the use of machine learning creates additional challenges. In the above example, after documented consideration, the assets of the applicant for benefits has been selected as an appropriate risk indicator. The reason for this is that this is a 'core fact' in the application for benefits and that '[an applicant] is not entitled to assistance if the assets are too large'.[34]

However, after calibrating the risk model through machine learning, the algorithm gives a greater research worthiness to negative abilities than to positive abilities. See Graph 1.4 for a simplified representation of the marginal contribution of this indicator to risk profiling. The effect is counterintuitive, because it is too much wealth that prevents social assistance benefits. The developers also acknowledge this in the explanation of these indicators and put forward various arguments.

For example, debts are a complicating factor in determining the legality of welfare benefits and debts should only be deducted from assets if debts have to be repaid. At the same time, someone with total assets that are positive can also have debts.

With such an indicator, (i) explainability and (ii) the prevention of arbitrariness can be problematic. Precisely because the model has been trained with historical data (on whether or not someone qualified for an assessment), the possibility that this indicator – in how it has been calibrated – is¬ a proxy for something else must be taken into account. Or that it perceives a statistically significant connection to something that is not relevant. If there is no clear logic with which the actual contribution to the risk assessment of an indicator can be explained, problems arise with explainability and the prevention of arbitrariness. The AI Act will set high standards for explainability: everyone who is subjected to a decision taken with AI has the right to a substantive explanation of the role of the AI system and the main elements of the decision taken. We can assume that individual risk indicators are part of these key elements. For a further discussion of the challenges in managing AI systems, see the Annex.

1.7 National AI Agenda

AI systems offer opportunities for society, which need to be actively explored and exploited. Not only by the government or the business community but jointly within the frameworks and values of society. Legislation and regulations provide relevant frameworks but a national strategy is needed to define and steer opportunities. In the ARR Edition 2, Autumn 2023, the AP[35] called for a Delta Plan for risk management. However, a strategic delta plan is also essential for exploiting opportunities. The current Strategic Action Plan on Artificial Intelligence was published in 2019.[36] Relevant developments in technology, and knowledge about applications and effects, are missing from this action plan due to the rapid pace of development. TNO Vector published four papers at the annual symposium 'Strategic autonomy in an open economy' that provide insights into the current role of technologies in national security, energy security, knowledge and innovation, but also into the costs and benefits of digital autonomy.[37] It indicates that targeted policy measures are needed, but that they are lacking or that there is a lack of understanding of their effects. This supports the need to have a clear strategy in addition to policy and regulation, which can provide guidance in turbulent times to take advantage of opportunities but also to protect citizens and society. This requires national choices about investments, focus areas and core values. But with fast-developing technologies, continuous adjustments and updates to the strategy also need to be made.

An overarching strategy will strengthen responsible developments in the field of AI in the Netherlands. Several Dutch initiatives in recent years have shown that the Netherlands is part of the movement towards a more responsible use of AI. Fundamental Rights and Algorithms Impact Assessment (FRAIA) shows that there are relatively many students graduating in AI here. Researchers from the Netherlands are also very well represented at global conferences on responsible AI.[38] Even more development, such as the aforementioned Smart Check pilot at the city of Amsterdam, is being carried out transparently and relevant supervisors are thinking together about effective next steps.[39] By further formulating the Dutch ambitions for the coming years, policymakers can strengthen these positive developments. The second edition of the ARR mentions, under 'Deltaplan Algorithms & AI: Ambition 2030', a number of possible themes to be reflected in the strategy.[40] These range from 'human control' to the 'national ecosystem and infrastructure'.

2. Information provision in democracy threatened by AI systems

QUICKLY TO THIS SUBJECT

2.1 Creation / production →

2.2 Moderation and distribution of information →

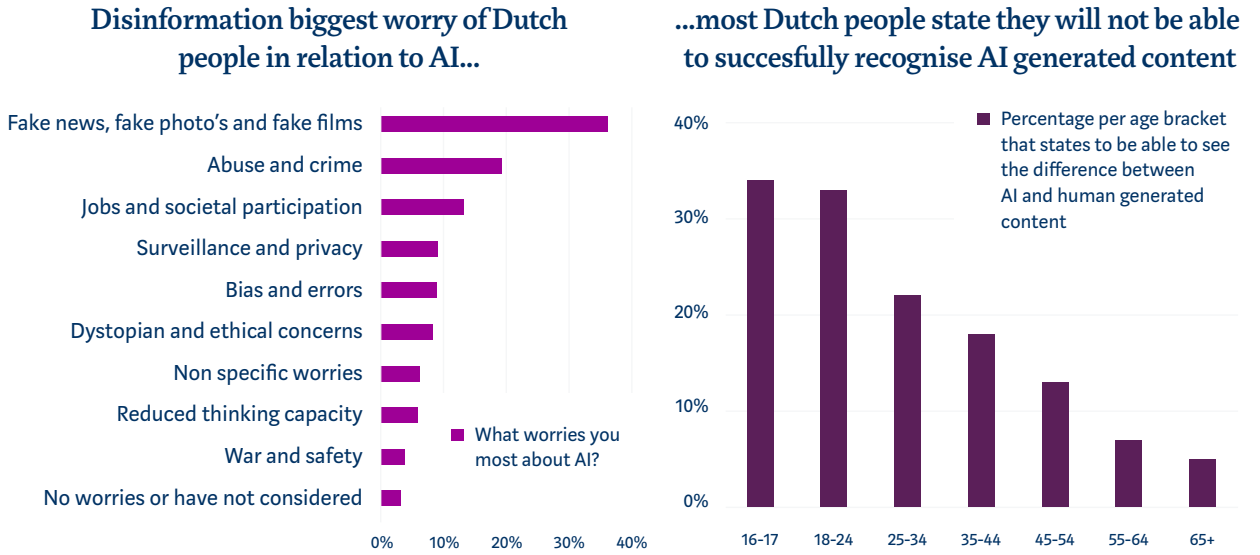
2.3 Consumption of information →

News collection and consumption are increasingly happening online. Young people in particular use social media as the main source of news.[41] AI systems have made a big mark on online information provision. In addition, platforms have a lot of power in this by deploying addictive AI recommendation systems. AI can also generate information on a large scale. A major risk is the generation and spread of disinformation and misinformation. The power of platforms and the misuse of generative AI endanger the diversity of the information landscape, putting democratic processes such as elections under pressure. The extent to which the functioning of the democratic system is actually affected is currently difficult to measure. Partly motivated by the further rise of AI in the coming years, it is therefore important to actively monitor this.

This chapter is partly based on public input.
In the spring of 2024, the AP issued a public call for input asking for visions, insights and concerns regarding AI systems in the provision of information. A dozen responses were received, each bringing their own insights, expertise and specific concerns to the fore. This knowledge and these insights have been used in the development of this chapter, in the preparation for discussions held on behalf of this chapter, and as inspiration for the internal thought process for this chapter.

Recent events show that AI is increasingly influencing public debate. In January 2024, residents of New Hampshire in the United States received a deepfake robocall from President Biden. An AI-generated robotic voice on the phone imitated Biden's voice and called on Democratic Party supporters not to go to the polls during the primary elections.[42] In India, deepfakes flooded the internet during the spring 2024 elections. There are many different kinds of misleading information: from humour and satire to illegal, offensive and harmful content.[43] Another example is an AI-generated image of a refugee camp in Rafah, which was shared tens of millions of times on social media in May

GRAPH 2.1: DUTCH ARE MOST CONCERNED ABOUT DISINFORMATION AND THEIR ABILITY TO RECOGNISE AI



SOURCE: WAAG FUTURELAB, APRIL 2024 (LEFT) EN ALGOSOC AI OPINION MONITOR 2024 (RIGHT)

2024 with the text All Eyes on Rafah. The role of AI in this is prominent: a virtual, 'clean' image that refers to human suffering without shocking encourages users to speak out against the war in this way. In addition, it slips through the automated moderation systems, which suppress bloody photos and critical expressions. Reactions to the post going viral are mixed. On the one hand, the AI-generated image may not accurately reflect the gravity of the situation in Rafah. On the other hand, this ensures a greater reach for the message that the post conveys.[44]

The Dutch are concerned about the risks of (generative) AI on the provision of information in our society. When Dutch people are asked about their biggest concerns about AI, most people think about the possible influence, spread and abuse of fake news, photos and videos. There are also concerns that generated information is indistinguishable from real (see Graph 2.1).[45] [46]This has a major impact on trust in the provision of information. The AP analyses the role of AI in the provision of information chapter using the information provision cycle: creation, moderation/distribution and consumption. To consider the different stages of the cycle and the influence through different channels. And to see what measures can be taken to reduce the risks as much as possible.

2.1 Creation / production

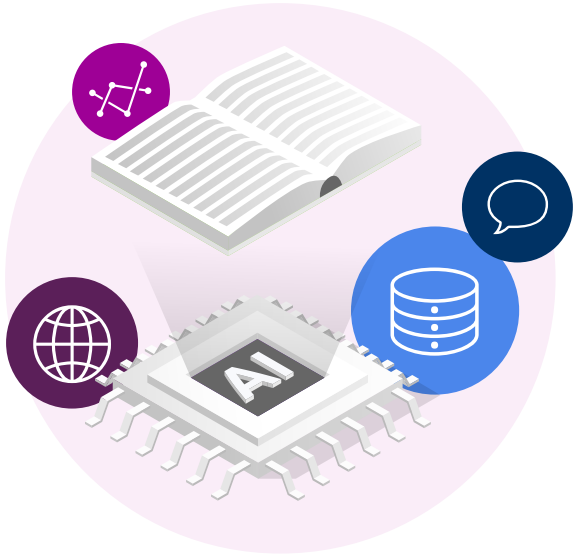
Generating and producing content is easier than ever with generative AI. This makes it possible to quickly create textual and audiovisual output that closely matches a specific request.[47] The problem is that generated content is increasingly difficult to recognize.

Generative AI makes it possible to generate disinformation on a large scale. Disinformation is untrue, inaccurate or misleading information that is intentionally spread to confuse or manipulate people. For example, to gain political or ideological support, to put other ideologies in a bad light or to create distrust and polarization.[48] This can affect democratic processes, the economy and national security.[49]

Generative AI systems also unintentionally produce misinformation. Misinformation is spread unintentionally, for example because people do not realize that it is false or misleading information.[50] However, the effects of misinformation can still be harmful. Because generative AI 'invents' answers, and cannot distinguish between what is true and what is not, errors quickly arise. A language model such as ChatGPT regularly provides incorrect information, while presenting that information as factual. This spring there was a stir because AI systems recommended the spread of disinformation and fear-mongering as a campaign strategy for the European elections.[51] In this way, AI-generated misinformation can be unknowingly brought into the world. Especially if developers, government organisations and users consider such systems to be all knowing or actually intelligent.

AI-generated content is increasingly taking the place of source content. An example is Google Overview, which makes the 'old' Google Search disappear by offering an AI summary instead of website links as the first result.[52] The quality of the AI search service still leaves something to be desired, but in the coming years these kind of AI systems will play an increasingly important role on different platforms.[53] The result is that the classic search engine, which organizes the results in a followable way, disappears as a result. Sometimes sources are missing, which makes it more difficult to fact check where the generated information comes from. The reliability of the information is therefore unknown and this increases the risk of misinformation.

At the same time, it is being investigated whether AI systems are also suitable for detecting whether AI has been used. Special AI detection tools may be able to recognize AI texts. For example, certain words and sentence constructs are statistically used much more often by AI than by human writers.[54]



Labelling or watermarking helps to distinguish real material from generated material. By 2025, the AI Act will make it mandatory to label generated images. This means that everyone who makes or distributes a deepfake must give openness about the origin and the technique that has been used. The aim of this is to stimulate transparency about the use of AI and thus prevent manipulative content. Developers and platforms are increasingly starting to label content.[55]

More transparency about the origin of information helps users to assess whether information is reliable or not. A joint initiative of several companies is the Content Authenticity Initiative (CP2A).[56] This is a coalition of media companies, technology companies and NGOs, which has developed technical standards that allow, among other things, cryptography to verify the origin of images. Companies such as the BBC, Microsoft and Google are affiliated with C2PA. In March 2024, the BBC used C2PA for the first time by providing, under a video, more information about its origin and explaining how the video was verified for authenticity.[57] Under the AI Act, providers of AI tools to generate material will also be obliged to apply these types of techniques to make generated content recognisable as such.

2.2 Moderation and distribution of information

The amount of information available is too large for an individual to decide what information is relevant, useful and reliable: this is called 'information overload'.[58] Before the digital age, newspaper and television newsrooms selected what would be offered to the public. Due to the increased amount of online information, the gatekeeper function of the traditional media has partly disappeared. In the digital age, citizens are not only consumers of information, they also produce and disseminate information themselves. For example, with accessible generative AI tools, which became widely available in 2024. In order to make these online information flows manageable, AI systems filter, as a form of editorial, for information that would be of personal interest to the consumer. Where newsrooms make substantive choices, AI systems are unable to do so. For example, AI cannot distinguish between truths and untruths, but it does record what kind of content appeals to the user.

The online information provision is in the hands of a few big tech platforms. More than 90 percent of the world's population uses Google's search engine.[59] Many young people are increasingly using the video platform TikTok as a search engine.[60] Platforms have an online gatekeeper function because of this power.

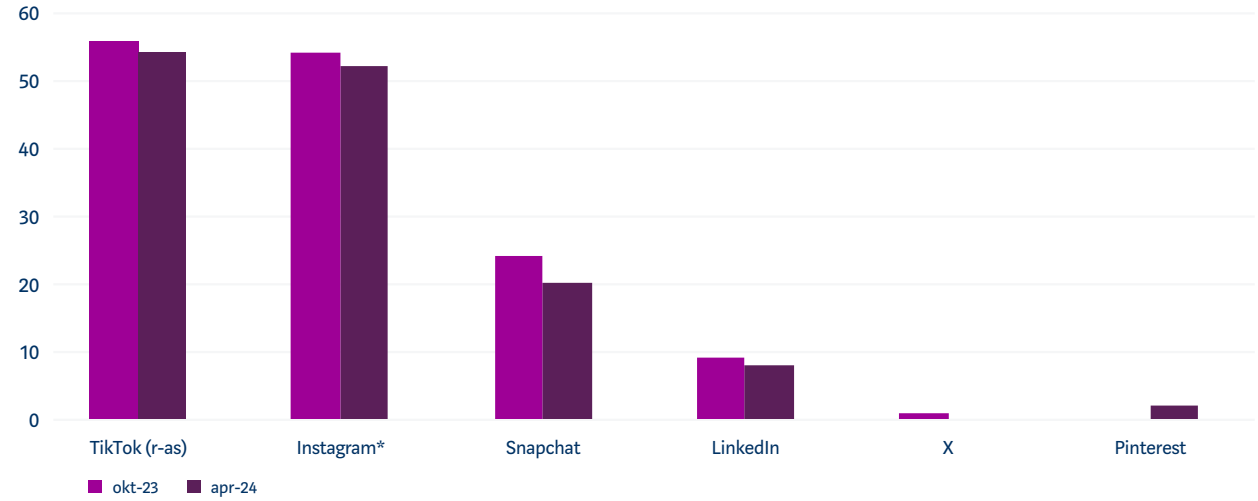
Big tech platforms largely shape our 'information diet': what we get to see, but also what we do not get to see. A personalized AI system should ensure that users stay on their platform for as long as possible, so that they see as many ads as possible because bigtech platforms rely heavily on advertising for their revenue. Consequently, the personalised AI systems are deliberately made addictive. This ensures, for example, that many messages are shown that evoke emotions.[61] Because these platforms mainly focus on revenue and not on the most diverse range of information possible, this can jeopardise the pluralism, reliability and independence of information provision.

Some governments use moderation to censor information on online platforms. This jeopardises freedom of expression and information.[62] According to a Freedom House report (2023), at least 22 countries surveyed are using AI to remove unwanted political and religious expressions from social media platforms.[63] The Information Technology Act in India has the provision that central government and other authorities may issue emergency orders to block and remove social media accounts, videos, messages or photos when the content is deemed harmful to the public order, peace, sovereignty and security of the country. This is widely interpreted and frequently used to block and remove critical content.[64] Social media platform X has had to make several critical messages from the opposition invisible to the Indian users of X.[65] The possible danger is that with the use of this provision, regulation censorship can be made easier, go unnoticed and widely applied.[66]

To limit the influence of large platforms on information provision, online platforms and search engines required to comply with the Digital Services Act (DSA) since August 2023. This law forces very large online platforms, designated by the European Commission, to tackle disinformation (see Box 2.2). Platforms are also forced to be more transparent about content moderation. Moderators check whether content is in line with the terms of the platform and they may have to remove content if it is illegal or harmful. Despite the

GRAPH 2.2: DECREASE IN DUTCH CONTENT MODERATORS AT LARGE SOCIAL MEDIA PALTFORMS

Number of Dutch content moderators per platform



SOURCE: DSA VLOP TRANSPARANCY REPORTS

fact that platforms indicated their intention to counter disinformation as much as possible, we see a decrease in Dutch-speaking moderators for almost all platforms (see Graph 2.2).

Platforms need to be transparent about the manner in which they moderate information. There should also be the possibility to contest the choices of the moderators if moderation seems to violate freedom of expression. The DSA stipulates that independent organisations can apply for to view the status of a trusted flagger. Trustworthy flaggers are organisations that detect illegal content and report it to the platforms. These notifications should be treated as a matter of priority by platforms. The flaggers are supervised by the Digital Services Coordinators, who play a key role in regulating the DSA. Trusted flaggers are required to publish annual reports.[67]

Users should have more influence on the information they consume. Positive first steps have been taken in the DSA. For very large online platforms, there is an additional obligation that allows users to disable personal recommender systems. Several very large online platforms currently offer this option.

The distribution of platforms should focus on a diverse range of information. The use of addictive recommendation and filtering systems may reduce the diversity of media offerings. This is detrimental to the public debate. Recommendation systems can also be used to show a diverse range of information. In order to protect the core values of media policy, the Dutch Media Authority (Commissariaat voor de Media), which supervises the media sector, recently published an exploration that maps out the effects of AI on this. The Box entitled 'How AI challenges the core values of media policy' provides insight into these effects.

2.3 Consumption of information

Disinformation and misinformation can lead to growing mistrust in the media. Even if the incorrect information has already been debunked. Not everyone who sees disinformation will know afterwards that it was incorrect information. The increasing amount of AI-generated content on online platforms increases mistrust of information. Many Dutch people also think they cannot recognize that content is generated by AI. This can cause them to question legitimate information. Journalistic evidence can then, for example, be put away as deepfake, even though it does contain real images or audio recordings.

Through the use of AI, large-scale automated influence is applied on the public debate. In its 2022 annual report, the Dutch General Intelligence and Security Service states that various countries are deliberately circulating disinformation in order to give the Dutch population a more positive, but incorrect, picture of actions by their country.[69] Disinformation is often spread by anonymous accounts. Increasingly, these are bots: accounts that are fully automated and controlled by an AI system. In 2023, bots accounted for 49.6 percent of all internet traffic.[70] Bots often operate on a large scale, spreading disinformation and thus disrupting the social debate. Bots can, among other things, create information, like and share messages and interact with users. In this way, they present themselves as real people and voters. By frequently using hashtags and liking certain messages,

Box 2.1

How AI challenges the core values of media policy

By: The Dutch Media Authority

For a functioning democracy, it is essential that everyone can form their own opinion. The Dutch Media Authority (hereinafter: the Authority) contributes to this by monitoring and stimulating an independent, pluralistic (various), accessible and safe media offer. To this end, the Authority supervises the rules set out in the Media Act 2008. However, the Authority also puts developments on the agenda that have an impact on the aforementioned values. AI is one of these developments. There are many opportunities in the use of AI, but the use of AI also poses risks. Below, we discuss the most important opportunities and risks per value.

Reliability of information is a guiding principle underlying media policy and closely linked to the core value of independence. On the one hand, the Authority sees that the use of generative AI can negatively affect the reliability of information. Think of the malicious use of AI to generate misleading information. Generative AI applications are also often not transparent about how they work. This is called a black box. This makes it difficult to understand how and by whom an image, text or video was created. On the other hand, the Authority sees that the use of AI can also support journalists and the media. In any case, AI must then be used responsibly, for example by means of guidelines. Media institutions can draw up these guidelines themselves. In doing so, they must indicate which use of AI is and is not permitted. A guideline can also indicate how and to what extent people are informed about the role of AI in the creation of the media offer they consume. However, the Authority notes that more transparency can also lead to more mistrust.

AI also poses some risks to media **plurality**. For example, the Authority sees that people may see fewer different types of media offerings online through recommendation and filtering systems. In addition, companies that develop AI applications have a lot of power in the market. For example, the infrastructure of AI is largely in the hands of a small group of companies. The 'opinion power' that these companies have is also increasing. This is because big tech affects the aforementioned recommendation and filtering systems, and that can determine what kind of media offerings people get to see. There are rules, for example in the Digital Services Act and the Digital Markets Act, that must ensure that the gatekeepers get less power.

The Authority also sees that AI offers many opportunities for the **accessibility** of the media offer. People with hearing or visual impairments in particular can gain easier access to the digital world through AI. For example, through the use of AI-based automatic subtitling, translation systems and audio description.

AI can also increase people's online **safety**. For example, by helping to automate age verification and moderate harmful media offerings. But the use of AI also carries risks with this value. Deepfakes can mislead people with realistic AI-generated edits of images, sound clips or videos. This can have harmful consequences for the democratic process. For example, if deepfakes are used to influence elections. To illustrate, in 2023, two days before the Slovak elections, a deepfake audio fragment of a politician was circulated. Finally, AI applications can have negative effects on mental health, especially among young people. For instance, look at the addictive algorithms that social media companies use.[68]

This Box was written by the Dutch Media Authority, supervisor of the Media Act. For more information on this topic, see the publication 'Between Bits and Principles: How AI challenges the core values of media policy' of June 2024.

they send the recommendation system, in order to influence what kind of messages are and are not displayed.[71] [72]

Disinformation often targets marginalised groups who are also disadvantaged or discriminated against in the offline world.[73] An example is increased anti-Semitism, which is increasingly visible online. Much anti-Semitic content consists of disinformation, ranging from conservative-nationalist reporting to conspiracy theories. AI systems play an important role in exposing users to this kind of disinformation, misinformation and hatred. This can also be seen in the large amount of misogynistic content, which is especially aimed at boys. Already after 5 days on the TikTok platform, the recommender system can recommend a quadrupling of that kind of content on someone's personal page, based on innocent interests such as mental health or fitness. In this way, young people can end up in online 'echo rooms' in which misogynistic rhetoric is normalised.[74] In addition, women are increasingly becoming victims of sexual deepfakes (manipulated images).[75] Even though artificial pornographic material has been a problem for some time, AI tools make it easier to create these videos and photos. In the Netherlands, too, this is a growing problem among female public figures and politicians.[76]

A lack of a common 'information base' can contribute to polarisation. In recent years, interest in following news has declined, particularly among 18-34-year-olds.[77] The news that they do come into contact with, they mainly get through social media. In addition, more and more people experience the amount of news as tiring. This is especially true for people who often avoid the news. This is a group that covers 8% of the Dutch population (a doubling compared to 2017). Social media is also the main source of news for them.[78] On social media, it is more likely that the coverage that does reach them contains incorrect information or is manipulative and one-sided. Although the extent of the effect of these 'filter bubbles' is uncertain, it makes it easier for people to come into contact with extreme views. Extreme content interacts with the social media, risks losing sight of the facts and belief in a shared reality. If this shared reality disappears as a common basis, it can contribute to a polarized political and social landscape, in which especially one's own community is trusted.[79] Social media and AI systems cannot be clearly identified as the cause of this development but they do make it more visible.

A combination of measures is needed to reduce the risks of AI for the online provision of information. Transparency about the origin of digital content is needed to determine the reliability of information. Citation, labelling and watermarking can help with this. AI can also be used to recognize generated content. Many AI detection tools are still relatively new and will be further developed in the future. In addition, media literacy and algorithmic literacy are necessary to be able to handle online information in the right way.

AI-literacy and media literacy are necessary to participate in the digital information society. Media literacy is the set of knowledge, skills and mentality that enables citizens to move consciously, critically and actively in a digital media society.[80] There are various initiatives from the government to make citizens resilient and media wise.[81] Digital literacy, for example, will become a mandatory part of the educational curriculum.[82] AI literacy should therefore also be an important part of digital literacy. Knowledge of AI systems and their risks and impacts is needed to be able to participate safely in the digital information society. The National Centre of Expertise for the Curriculum (SLO) has incorporated the goal of AI literacy into the concept core goals of digital literacy.[83] It is important that not only children and young people, but also adults become more aware of the influence of AI systems on the provision of information. The risks of AI for the provision of information can have an affect on everyone.

Box 2.2

The Digital Services Act

The Digital Services Act (DSA) requires very large online platforms and very large online search engines to take measures against systemic risks, such as threats to democracy. Dissemination of disinformation may classify as a systemic risk under the DSA. Very large online platforms need to take measures to counter this. The DSA is initiated by the European Commission (EC) and is applicable in all Member States of the European Union.

The EC publishes guidelines for very large online platforms and search engines to avoid negative effects on elections.[84] The guidelines further illustrate the obligations of platforms and search engines under the DSA with concrete examples and best practices. However, following the guidelines is not mandatory. Platforms and search engines may also comply with the DSA by mitigating systemic risks in ways other than those described in these guidelines. The success of the measures therefore depends heavily on the willingness and interpretation of the platforms and search engines.

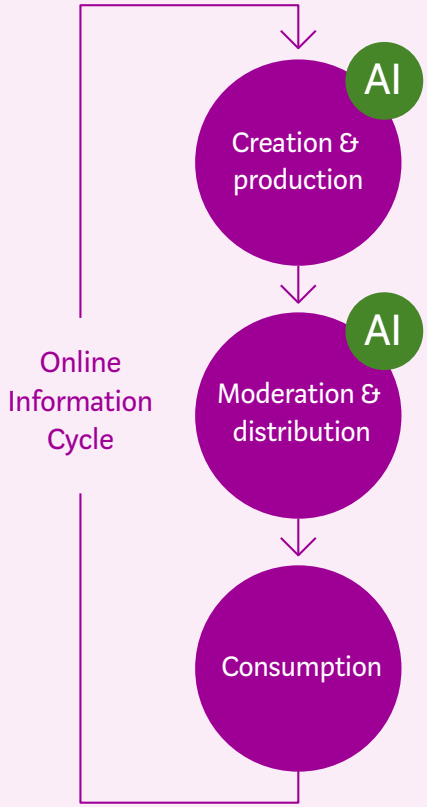
Some of the concrete proposals in the guidelines are specific adaptations of recommender systems to protect electoral processes. For example, very large online platforms can take measures to ensure that their recommender systems do not show demonstrable disinformation in elections. Very large online platforms must be transparent and clear about such measures. The guidelines also propose to increase the recognisability of official accounts. This can help to counter the spread of disinformation and misinformation on very large online platforms. Another example is designing recommender systems in such a way that users gain meaningful control over the information they consume.

The guidelines also include specific proposals for generative AI. The EC recommends very large online platforms to watermark AI-generated content. In this way, AI-generated content is easier to recognize. Information generated by AI systems should also be based as much as possible on reliable sources in order to avoid incorrect outputs. In addition, the guidelines repeatedly mention the importance of increasing media literacy. Improving algorithmic literacy can help citizens to be alert to the possible spread of misinformation or disinformation in elections.

The EC, national regulators, civil society organisations and the very large online platforms have carried out a stress test to see if large platforms are ready for election times. The EC ran through different scenarios in the stress test and looked at whether the internal procedures and practices of large online platforms and search engines effectively counter systemic risks in elections. The scenarios include, for example, the spread of political deepfakes to mislead voters.[85] The EC does not disclose the results of the stress test.

Based on the DSA, the EC actively monitors the management of systemic risks in elections. The EC has recently launched an investigation procedure against Meta. The EC suspects that Meta is not complying with its obligations to manage systemic risk.[86] For example, there are concerns about the spread of disinformation and the visibility of political messages on Meta platform users feeds.

AI influences the online information cycle at all stages and through different channels



Channels of AI-related risks

- **Abuse**: More and more complex disinformation by using AI
- **Hyperrealism**: AI-content can no longer be distinguished from real content
- **Unconscious errors**: Misinformation in AI-generated content

- **Concentration risks**: Market power makes information provision dependent on algorithms of tech companies
- **Algorithm design**: Incentive to prioritize engagement over quality
- **Moderation risks**: Balance needed between safeguarding freedom of information and protecting against mis- and disinformation and illegal content

May lead to

- **Loss of insight into reality** due to mis- and disinformation and filter bubbles
- **Social polarisation** due to lack of a common information base
- **Mistrust in media** due to fundamental uncertainty whether information can be trusted

Trust in information provision decreases and this puts democratic processes under pressure

Mitigation options

- **Watermarking** of both real content and AI-generated content
- **Quality standards for testing** generative AI models
- Make content and source material **verifiable**

- **Pluralism** of media needs to be ensured
- **Customizable algorithms** on platforms to give users more control over content
- **Transparency and contestability** of moderation together with independent oversight and trusted flaggers

- **Increasing media literacy** for young and old
- **Visible labelling** of both real content and AI-generated content

3. Challenges in democratic control of AI systems

QUICKLY TO THIS SUBJECT

The shape of the process for democratic supervision and setting frames of reference for the use of AI systems determines the way in which representatives of the people can exert democratic control on AI systems that are used by the government. This can range from the House of Representatives to the city council. Democratic supervision must be possible during every phase of development, deployment and evaluation of an AI system. This chapter explores this topic through the situation in local government. The insights and recommendations are relevant to all levels of government.

Governments use a diverse range of AI systems to automate all kinds of processes. This is evident, among other things, from an exploratory study by TNO.[87] In recent years, automation with AI systems has also led to incidents and fundamental rights violations in governments. The local government carries out many tasks, which can have a major impact on citizens. Dutch municipalities use a large part of the AI systems in the public sector.[88]

Survey results show that municipal organisations have limited oversight of their AI systems, that council members have doubts about the adequacy of their AI knowledge, and that only a few local audit institutions conduct sporadic research into AI systems. Municipalities need clear frameworks and rules to shape the democratic control cycle for AI systems. Specifically, it concerns clarification of questions such as: (i) how does the executive account for the use of AI within their organisation to councillors (ii) how can an external auditor, such as a court of auditors, effectively audit AI systems and (iii) what questions could representatives best ask about AI systems and at what point in time? This could include an overarching AI coordination centre and/or centres of expertise to support the democratic control cycle of AI systems in public administrations.

3.1 Case: Democratic control of AI systems in local democracy

Municipalities, like other public authorities, use AI systems for various purposes – this has led to fundamental rights violations in the past. Municipalities use AI systems to perform their tasks more efficiently. For example, preventing fraud when registering in the Personal Records Database,[89] communicating with people who do not speak Dutch[90] or proactively helping with debts.[91] And in many municipalities, scan cars are now driving around that use AI systems to check for parking violations.[92] At present, almost 75% of the algorithms registered in the national Algorithm Register come from municipal organisations (see Graph 5.4 in Chapter 5). Recently, the use of AI by municipalities has led to risks to or breaches of fundamental values and fundamental rights on a number of occasions. The AP has previously pointed this out, for example in the 2023 annual report.[93]

This chapter is partly based on a survey among municipal organisations, councillors, local courts of auditors and local ombudspersons. Together, these organisations ideally ensure a responsible and democratically anchored deployment of AI systems, while keeping an eye on risks and incidents. So that municipalities use AI technology in a way that contributes to fundamental values and the protection of fundamental rights. In total, the survey was answered by 85 municipal organisations, 35 courts of auditors and 27 councillors. The number of local ombudspersons that were able to respond to the survey was minimal. The survey was set up by the AP and was carried out in March and April 2024. Any quoted responses have been translated by the AP.

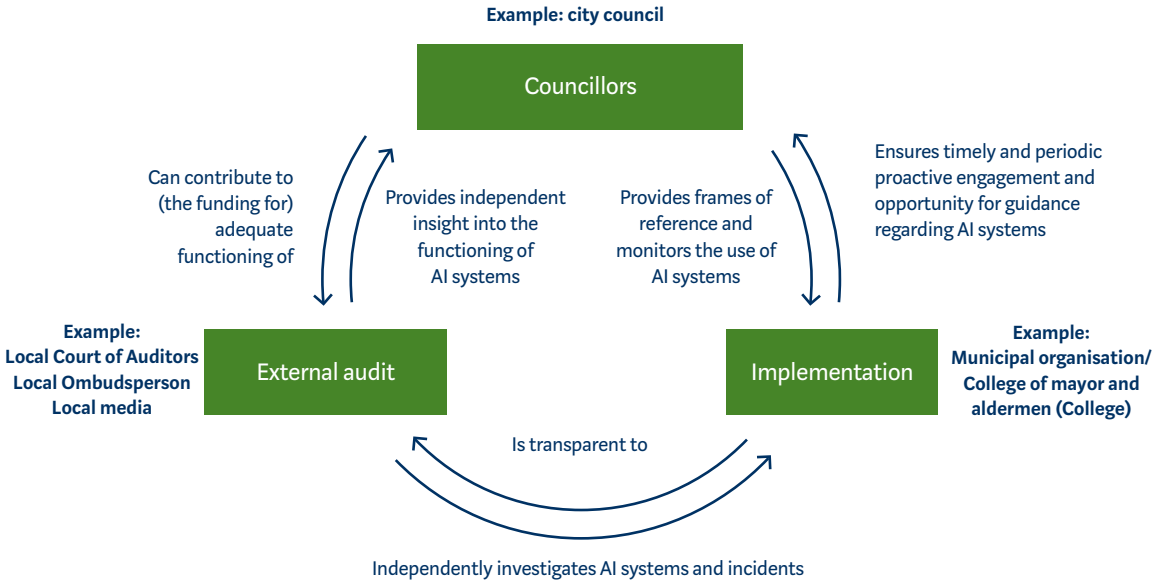
3.2 The democratic cycle for AI systems

The use of AI systems in the public sector must be part of a democratic cycle of direction and accountability. This applies at national, regional and local level. Representatives should supervise and set frames of reference for the performance of government tasks. This includes the use of AI systems in that implementation. In doing so, they can build on the work of other institutions, such as the media or regulators. Representatives of the people give frameworks to an executive authority. This can also indicate whether and, if so, how the government can use AI systems. Within these frames of reference, the government develops and implements an AI policy. The representatives of the people then assess the government's use of AI and adjust the policy frames of reference based on their judgement. This ideally creates a democratic cycle of control and accountability of the use of AI systems by governments (see Graph 3.1).

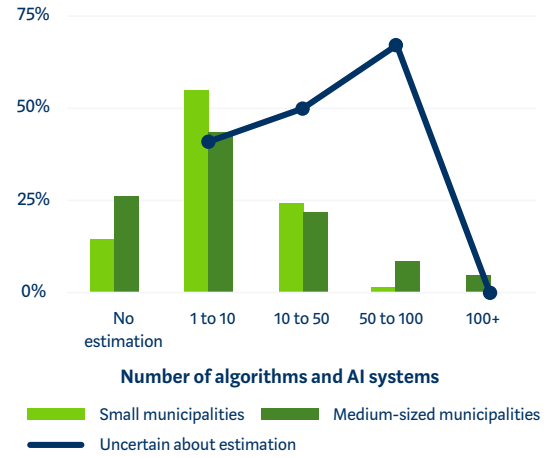
Municipal organisations receive frames of reference for their policy from the municipal council and are also accountable to the council through the college of mayor and aldermen. The city council has a directing (through the provision of policy-frames of reference) and a supervisory task, in addition to the task of representing the inhabitants.[94] If the college is accountable for how it uses AI systems, the city council can pass judgment on this through the supervisory task of the council.

Local courts of auditors, local ombudspersons and local media also monitor the college and can feed the frames of reference that are set by the council. For example, auditors independently examine the effectiveness, efficiency and legality of policies pursued by a municipality.[95] If they report on the legality of municipal use of AI systems, the municipal council can use this to assess the college. Local ombudspersons help citizens with complaints against the local administration. The aim is to protect citizens from actions by the municipality that harm them. For this purpose, the Ombudsperson has extensive investigative powers.[96] The results of investigations by the ombudspersons can also be used to assess and adjust the college. Local media can openly question the college critically and can put societal issues on the agenda.[97] The city council can also make use of this in assessing and redirecting the college.

GRAPH 3.1: CONCEPTUAL DEMOCRATIC CYCLE FOR THE USE OF AI-SYSTEMS



GRAPH 3.2: MANY MUNICIPALITIES PRODICT THAT THEY ONLY USE A LIMITED NUMBER OF ALGORITHMS AND AI-SYSTEMS, BUT ARE INSECURE ABOUT THESE NUMBERS



SOURCE: RESEARCH BASED ON AP-SURVEY AMONG MUNICIPAL ORGANISATIONS (N=85)

3.3 Challenges in the use of AI systems by municipalities

Many municipalities aim to use only a limited number of algorithms and AI systems, but are uncertain about their estimate. The survey of municipal organisations shows that more than half of the smaller municipalities expect to use no more than 10 algorithms and AI systems (see Graph 3.2). However, the estimates are diverse: Several small municipalities expect to use more than 100 algorithms and AI systems. Doubt about this is great, however, especially with smaller and medium-sized municipalities that use 50 to 100 algorithms. Almost three-quarters indicate that the estimate is uncertain or very uncertain. This uncertainty indicates that municipalities still have to take steps to get an overview. This is a prerequisite for risk management.

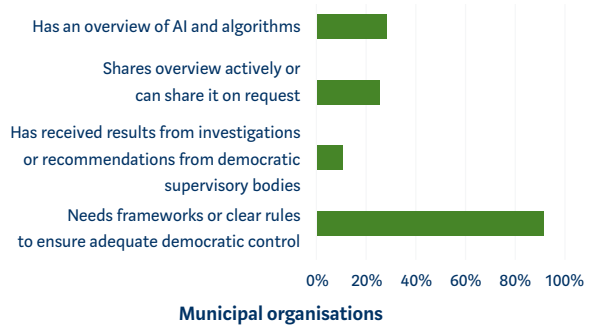
The current uncertainty is understandable, as municipal organisations face challenges in terms of knowledge, expertise and resources due to their size. The municipal challenges on the one hand make the choice to use AI systems attractive, but on the other hand make it difficult to use these AI systems responsibly. For example, civil servants need knowledge in order to assess the value of a purchased AI system and to be able to oversee its impact. One municipality indicated, in one of its answers to our survey that "it was not always known in the past - if we were using an application - wheter this [was or was not] the use of AI of algorithms." To take the next step forward, municipalities need to work on sufficient AI literacy (as required by the AI Act as of 1 February 2025, see Chapter 5). However, limited resources and a challenging labour market make this difficult. Purchasing or hiring expertise is expensive, as is the training of the officials who have to deploy and master an application on a daily basis.

Municipalities are not always aware of the impact and political nature of choices in the use of AI systems. This is a problem, especially since municipalities use AI systems in processes that affect people in vulnerable positions, for example in the social security domain.[98] Until recently, municipalities saw the application of AI systems primarily as a technical issue and an efficiency measure. Moreover, given their size, it is difficult for both small and large municipal organisations to get the right information about an AI system. As a result, the implications of deploying an AI system for fundamental rights and values remain unforseeable.[99] As stated by a respondent:

> "There is a strong need for independent certification for algorithms and AI. Most organisations rely on suppliers to gather information about an algorithm or AI. They are unable to technically and sufficiently test the underlying training dataset and the algorithm or AI for topics such as bias."

Municipalities that are aware of fundamental rights in their use of AI systems mainly focus on the right to data protection.[100]

GRAPH 3.3: MUNICIPAL ORGANISATIONS HAVE LIMITED INTERACTION WITH THE CITY COUNCIL, COURT OF AUDITORS AND OMBUDSPERSONS ABOUT AI AND ALGORITHMS



Explanation: *28% of the municipalities know for sure that they have an overview of algorithms and AI systems in the municipality. 26% share this overview at least once a year, they share it on request or are able to share the overview. 11% received recommendations or results from investigations into algorithms and AI from democratic supervisory bodies in the last three years. 92% of municipalities say they need frameworks or rules to ensure adequate democratic control regarding the use algorithms and AI. Source: AP-survey results (n=85)*

Municipal organisations are rarely sharing information about AI systems with the city council, courts of auditors and councillors. The survey shows that just over 20% of municipalities have an overview of AI and algorithms. Also just over 20% share this overview actively, or upon request, with the other parties in the democratic cycle. A smaller proportion (approximately 10%) subsequently also received results of research into or advice on AI systems from the city council, or auditors such as the Court of Auditors (see Graph 3.2). In the explanatory memorandum, for example, a municipal organisation does not take the involvement of municipal councils or citizens for granted: "Why are questions asked about the involvement of citizens and/or the Council? Will this soon be a legal obligation?" A possible best practice for municipal organisations is to consider embedding information about AI systems in the

regular provision of information. For example, a respondent indicates that the algorithms involved are appointed in decision-making documents to the city council:

"In the Council proposal, we include a paragraph on the algorithm used to arrive at this proposal (policy-making), or to implement this proposal (supervision and enforcement)."

This is in line with the observation that it is still a challenge for municipalities to organise their own use of AI systems. Responsible use is only possible if there is an overview of which AI systems their own organisation uses and who is responsible for it. However, research commissioned by the College of Human Rights revealed that the responsibilities for AI systems in some municipalities are so unclear that the researchers call it an 'administrative spaghetti':[101] in other words: it is hard to untangle. Some municipal organisations are also deterred by a broad interpretation of the concept of 'algorithms' and suggested legal obstacles. For example, a respondent replies that "each automated application contains a system of algorithms. It is not possible to map them, apart from the fact that the source code is copyrighted." It is clear that the mapping of AI systems must involve systems that have an impact through their output, for example by making predictions, generating content, making recommendations or making decisions. Hence the importance of the broad but specific definition of AI systems (see introduction of this ARR).

There is a great need among municipal organisations for frameworks or rules for shaping democratic control of AI systems. In the survey, more than 90% of municipal organisations agree with the need for such frameworks (see Graph 3.3). In explanatory questions, municipal organisations indicate that they mainly need concrete frameworks with little room for interpretation and, on the contrary, a lot of certainties. For example, one respondent indicates that the "disadvantage so far is that all frameworks in this area are extremely vague in terms of definitions or miss the mark in this area." Another respondent indicates that "the assessment framework as well as the estimates to be made to classify algorithms are rather complex and, above all, subjective." Furthermore, a need expressed by yet another respondent is to share best practices that municipalities can build upon.

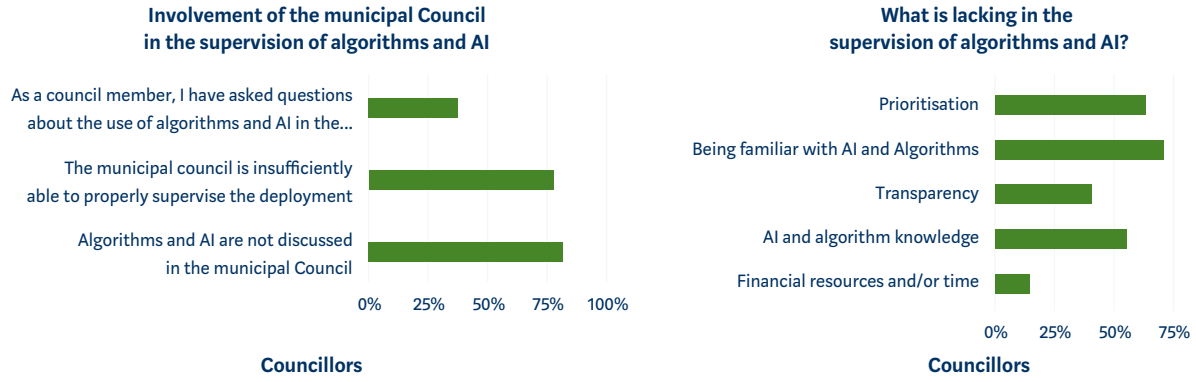
3.4 Similar challenges for directing and supervising parties

"Algorithms and AI are not actually discussed in municipal councils"

Three quarters of the councillors who answered the survey agree with this statement. Three quarters of the council members therefore believe that the council is insufficiently able to supervise the use of algorithms and AI (see Graph 3.5).

This is due to a lack of prioritisation, awareness of the functioning of AI systems and AI literacy. More than half of the council members who answered the survey see this as bottlenecks. According to the council members, there are various points to pay attention to here. For example, a council member indicates that "if it is not discussed, it cannot be checked." Another councillor indicated that "a lot is happening and no one seems to have the overview.

GRAPH 3.4: COUNCILLORS AGREE THAT ALGORITHMS AND AI ARE RARELY DISCUSSED – DUE TO LACK OF PRIORITISATION AND AWARENESS



SOURCE: RESEARCH BASED ON AP-SURVEY AMONG COUNCILLORS (N=27)

This means that we cannot carry out our monitoring task." The lack of in-depth attention is a concern, as it can also stand in the way of valuable applications. In the words of a councillor:

> "Councillors know nothing about AI. They find it scary and [that] creates a lot of resistance."

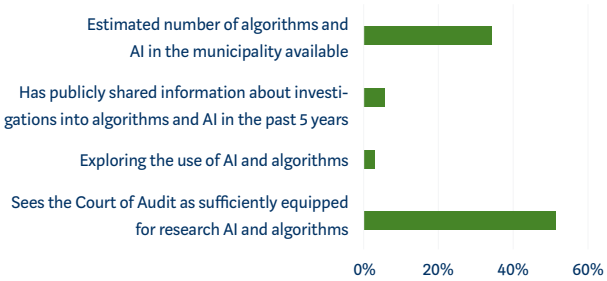
Municipal councils are also not always aware of the impact and political nature of choices when using AI systems. As a result, councils often do not take advantage of the opportunity to set frames of reference for and exercise democratic control over AI systems at the municipality. That is according to the Rathenau Institute based on its own research from 2020. The Rathenau Institute found that council members did not seem to be sufficiently aware of the impact of digital technology. According to the Rathenau Institute, councils rarely discuss digitalisation as a subject on which political and ethical choices can be made.[102] The Council for Public Administration (ROB) also concluded that the use of AI systems is not sufficiently approached as a moral and political issue. The ROB also noted that awareness about AI systems is increasing.[103] The survey conducted for this chapter still gives a similar picture on points. For example, a council member indicates that "implementation of the policy is taken away from the civil service". In that sense, it can only be called into question through the proper functioning of the civil service and not directly by the instruments used for that purpose.'

Knowledge and expertise in the field of AI technology is a challenge for representatives of the people. That makes it difficult to realize what questions they have to ask when using AI systems at the municipality to check the college. However, it is precisely for the task of setting the frames of reference that representatives of the people need knowledge. The Rathenau Institute found in 2020 that councillors often consider themselves insufficiently competent to judge on digitalisation issues.[104] And the survey confirms this picture. For example, a council member replies that "there is too little knowledge to ask good questions." However, practical solutions to this are conceivable. For example, another council member asks if "there is a sample list of questions available somewhere that we can use to ask targeted questions?".

Limited funding of local democratic institutions makes effective supervision and direction more difficult, for example for council members. Sufficient funding of the municipal council is important because of the part-time nature of council membership and the limited size of municipal councils, which means that specialization among council members is limited. That is why council members benefit from the help of, for example, group support. However, the average annual budget for this is very limited.[105]

Courts of auditors are competent but can only conduct limited research into algorithms and AI. Fewer than one in ten local courts of auditors have conducted research into the use of AI and algorithms in the past five years and shared public information about this research. More than half of the local audit institutions consider themselves competent on this subject (and therefore see the subject as part of the mandate) – see Graph 3.5. Prominently visible are the activities of the audit offices of the largest municipalities, such as Amsterdam, The Hague and Rotterdam. The Amsterdam Metropolitan Area Court of Auditors published an investigation into the application of algorithms in Amsterdam in October 2023.[106] The Rotterdam Court of Auditors published a follow-up study on the use of algorithms in March 2024.[107] In March 2024, the Court of Auditors announced that it would start an exploration into the use of algorithms in The Hague in order to arrive at a fully-fledged research design.[108] The working methods of these local audit offices in large municipalities can, through best practices, provide inspiration for local audit offices in smaller municipalities.

GRAPH 3.5: LIMITED ATTENTION FROM LOCAL COURT OF AUDITORS FOR ALGORITHMS AND AI



Explanation: *Where a local court of auditors is responsible for several municipalities, the answer to the first question concerns the largest municipality.*

SOURCE: RESEARCH BASED ON AP-SURVEY AMONG LOCAL COURT OF AUDITORS (N=35)

More than 90% of the local audit institutions do not yet know whether they will include algorithms and AI in the coming years. This is indicated by local courts of auditors in response to the survey. This is partly because many courts of auditors – given the new act on the strengthening of decentralised courts of auditors – have only been in existence for a short time in their current form. At the same time, most local courts of auditors indicate that they are open to the subject. In response to the survey, the Courts of Auditors indicate that it is an 'interesting topic', where they 'want to discuss whether this topic lends itself to an investigation' and also take into account 'that this investigation should be repeated on a regular basis'.

The lack of resources limits the possibilities for courts of auditors, councillors and local media to carry out their monitoring tasks. Many local audit offices have budgets for one or two audits per year. A respondent to the survey makes it concrete: "Local courts of auditors, especially of small municipalities [have] a very limited time and capacity, investigations like these [about algorithms and AI] are beyond our capabilities. The commitment of board members and total support includes a maximum of one or two days per week. The budget for all investigations is EUR 30,000 per year." The basic funding of local broadcasters is also insufficient, according to both the Fund for Journalism and[109] the Councils for Public Administration and Culture. Local private media also has little money and therefore do not fulfil their role as supervisor of the municipality.

3.5 Establishment of decentralised government and role of frameworks

A complex local organisation around AI systems complicates democratic control. For many implementation issues, colleges nowadays opt for collaborations with other municipalities.[110] Or they opt to outsource tasks to market parties. The use of AI systems is also often partially outsourced.[111] An opaque division of responsibilities not only makes it difficult for municipalities to control their own AI systems. It also makes it more difficult for supervising parties, such as the city council, to assess and adjust the value of municipal use of AI systems. This undermines the democratic legitimacy of municipal use of AI systems.

The establishment of the local administration also creates obstacles in supervising and directing the municipal use of AI systems. The financing of the council, court of auditors, local broadcaster and councillors presents the council with difficult dilemmas when determining the municipal budget.[112] In addition, many city councils have a tight budget.[113] Does the council opt for an extension of the budget of the council fractions, the local broadcaster, the councillors or the court of auditors? In the current situation, this is at the expense of the budgets for services that directly benefit the residents. National laws and regulations can however provide guidance on how the system of control and correction works at local level. An example is the act on the strengthening of decentralised audit institutions, which entered into force on 1 January 2023.

Regulation provides guidance for the deployment of AI systems in the public domain. In addition to existing regulations, such as the GDPR, the upcoming AI Act will impose obligations on many AI systems used in the public sector. In terms of content, it is mainly about anchoring control measures that were already advisable or unavoidable. For example, it will be mandatory to have a risk management system, monitor risks and register AI systems. It is wise to make use of the guidance provided by the AI Act as soon as possible. See the annex to this ARR for an indication of how this control framework is being shaped.

3.6 Recommendations

National politicians and policy makers can help local government by supporting local authorities with sufficient resources, knowledge and flexible executive capacity for AI supervision and direction. By analogy, these recommendations apply to democratic control at national level, and also in relation to the deployment of AI systems in implementing organisations and other governance bodies. Automating operations through AI systems often seems an attractive option, given staff shortages, and the need for efficiency and cost-saving operations. However, the AI paradox is that this requires significant investment in an infrastructure for controlling AI, personnel that is sufficiently trained and sufficient personnel for human control of AI systems. Consequently, targeted requirements, funding and support from national level can contribute to this.

An overarching AI coordination centre and/or a centre of expertise, which can be used by local authorities to supervise and direct the use of AI systems, may have an important role to play in this regard. Advice has already been published on the possibilities for the distribution of knowledge and expertise. For example, in 2021, the Dutch Scientific Council for Government Policy (WRR) recommended setting up an AI policy infrastructure, starting with a national AI coordination centre. This would, among other things, increase the learning capacity of governments in AI systems.[114] Note that the British Alan Turing Institute could provide inspiration for this (see Box 3.1). In recent years, many new AI systems, for example for generative AI, have only become more complex. This increases the added value of national and regional structures to support the deployment of AI. Think of centres of expertise that can structurally help municipal organisations and local audit institutions to supervise and direct AI systems. Given the technical complexity of AI systems, organisations should not have to reinvent the wheel over and over again.

Municipal organisations can help the council and other supervisory institutions along the way by means of clear AI governance and by proactively involving the council. Basically, an overview of your own AI use is the basis for AI governance. This enables the council, the court of auditors, the councillors and local media to take a critical look at the use of AI within the municipality and, if necessary, to adjust it. Because such an overview makes it easier, for example, to inform the council and gives council members better structured information. It is the responsibility of the college to further promote the democratic cycle by involving the council timely and proactively in issues about AI within the municipality. In addition, the college could present choices about AI systems as choices in which democratic control is desirable, because these choices can have an impact on fundamental values and fundamental rights. See also the annex to this ARR, including the emphasis on explicit target definition and a balanced decision on exploring the deployment of an AI system.

The Ministry of the Interior and Kingdom Relations is developing an algorithm framework. This is currently the most important instrument that is being worked on to support governments. At the end of June 2024, the State Secretary for Digital Affairs and Kingdom Relations indicated that the algorithm framework should provide an overview of the main requirements for the use of algorithms and AI systems. In doing so, the framework is developed in an open source manner, in broad working groups in which inter-administrative parties participate. The intention is to apply the framework to the entire government. The State Secretary indicates that the supervisory bodies, such as audit services and courts of auditors, supervise the setting of standards.[115]

The AP notes that the framework must be as concrete as possible to help government organisations. In addition, it is important that the framework is not only about (i) the requirements for the (quality) control of individual AI systems, but also about (ii) requirements for the AI governance of government organisations and (iii) how the democratic control cycle is shaped. For example, by providing basic information that parliamentarians and external bodies can use in their supervising and directing tasks.

The AP points out that it is crucial to align the algorithm framework as closely as possible with the binding requirements of European regulations, such as the AI Act and the GDPR. In the second half of this year, the AP will provide a further analysis of the algorithm framework under development based on the coordinating algorithm task.

Strengthening control obligations specific to AI systems can increase learning capacity. This can be done, for example, by requiring public AI systems with impact – from simple algorithms to complex models – to be audited. The results can be input for the supervisory task for AI systems within municipalities. A professional party must then carry out these audits according to clear criteria. Pre-established frames of reference can be included in the auditing of public AI systems, in order to optimally provide the city council and citizens with relevant information. The introduction of a statutory audit obligation that strengthens the auditing parties and optimally provides them with knowledge and information can contribute to the learning capacity of municipalities.

Box 3.1

Supporting local and regional authorities through AI knowledge institutes – UK experience

The Alan Turing Institute (ATI) is the UK's national institute for data science and is the best place for public organisations in the UK to address data science questions. The institute focuses on gathering knowledge and using it in projects in the field of data science and AI. In this way, the ATI advises, among other things, the public sector in the field of data science. Through the ATI, public organisations have one institute to turn to for knowledge and assistance in complex AI issues. The ATI is financed to a large extent by public research and innovation funds.

The London district of Camden approached the ATI to develop a vision on data use together with residents in 2021. Camden wanted to develop a vision to handle personal data more ethically and thoughtfully and to use it responsibly in algorithms. During this process, Camden was in constantly engaging with residents. Their ideas and contributions, collected with a survey and input sessions, formed the basis for the final vision.

The external expertise of the ATI contributed to the level of public debate and the final policy. The ATI helped to inform residents or other lay people well and to initiate an entertaining debate. In this way, governments and residents were given the opportunity to learn from each other and they were able to work towards a form of agreement and a broadly supported approach. The multi-disciplinary approach of the ATI also contributed to asking as many relevant questions as possible.

For more information, see the Alan Turing Institute .
Including information on the Camden Council project.

4. Profiling and selecting AI systems: Risks and the random sample

QUICKLY TO THIS SUBJECT

4.1 Risk profiling and selection →

4.2 Discrimination and excessive trust →

4.3 Random sample →

4.4 Worth considering →

Many organisations use algorithms for profiling or similar processes that distinguish between people. This chapter explores this topic through examples in the field of fraud detection. It is important to always view these algorithms as an AI system and therefore part of a broader process. The risk of discrimination is a recurring theme in this context it can occur at various points in the process around a profiling AI system. For example, through non-representative data and overreliance on algorithmic outcomes. A random sample can be used as a control tool. The advantages of this technique are that algorithms can be monitored better and that this technique ensures that there is a human decision in the process. Complementing fraud detecting algorithm processes with a random sample is therefore recommended in many cases.

4.1 Risk profiling and selection

Many organisations use algorithms to profile and select. On the basis of cases from the past, organisations make an estimate. This allows them to take action, such as a targeted investigation into certain persons. The estimate serves in that case as a selection tool to select persons who will be examined by an inspector.

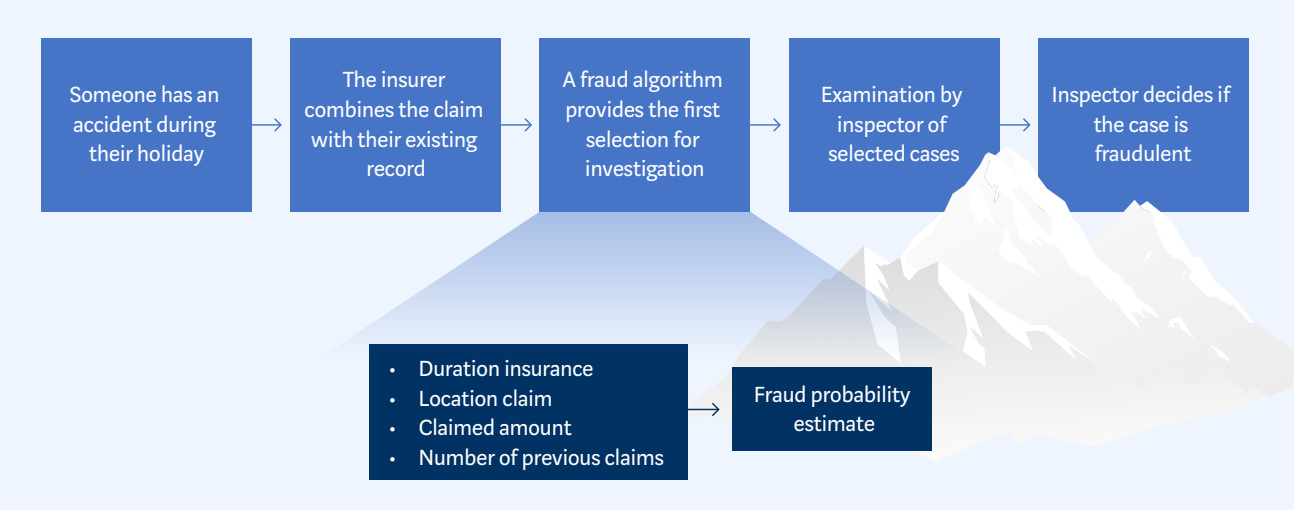
An example of profiling and selecting systems are fraud detecting algorithms, which almost everyone comes into contact with. Insurers use algorithms to search for insurance fraud,[116] banks use fraud models to check transactions[117], and online platforms search for fraud within new user accounts.[118] Often a citizen, customer or user does not know that a fraud check is taking place. As long as there is no suspicion of fraud, the fraud detecting algorithm remains an invisible step in the process.

"The lack of safeguards in the implementation practice of risk-based supervision and the violation of laws and regulations have the consequence that some people are more likely than others to be exposed to the public authorities and checked in the context of fraud prevention."

Parliamentary Committee of Inquiry: Blind voor mens en recht, 7.1 (translation: AP)

Errors in fraud detecting algorithms can have a major impact on individuals. In addition to legality issues – may such an algorithm be used in certain situations and may certain indicators be used– it is important to note that errors in these algorithms have a major impact. The evidence of this can be seen in the childcare benefits scandal. Research by the Dutch Data Protection Authority (AP) showed that fraud detection in the childcare allowance involved unlawful use of data on dual nationality and involved discriminatory processing of personal data.[119] The Parliamentary Committee of Inquiry concluded that this was a violation of the fundamental rights to privacy and to equal treatment.[120] A second example is the system called Systeem Risico Indicatie (SyRI). SyRI was used to detect social security fraud. The court ruled that the method violates the right to privacy. The court considered that the system was not sufficiently transparent and verifiable, while the use of SyRI could (unintentionally) entail discriminatory effects[121] A third and more recent example is in the controlling of fraud in the student housing grant process , where DUO (Dutch excecutive body for education services) used a selection algorithm.

GRAPH 4.1: A FRAUD ALGORITHM IS PART OF A BROADER PROCESS. IN THIS VISUALISATION THE FRAUD ALGORITHM IS ONE OF THE STEPS IN THE PROCESSING OF A CLAIM FOR A TRAVEL INSURER



An investigation report concluded that there was discrimination, for which the government and DUO subsequently apologized.[122] In this case, this is an addition concern to discrimination associated with the use of indicators such as a level of education as a distinguishing factor for fraud risk.[123]

4.2 Discrimination and excessive trust

Risks in the use of fraud algorithms often arise in the process steps surrounding the algorithm. The outcomes of an algorithm depend on the steps taken in the process beforehand and the final impact depends on how the outcomes are handled. To illustrate this, a fictitious example of a fraud checking process by a travel insurer is schematically presented.

The algorithm here depends on the submitted claim and existing data sources. Next, the algorithm serves here as a pre-selection in the fraud checking process: an inspector then investigates the cases with a highrisk indication (see Graph 4.1).

Discrimination is an important risk when using fraud algorithms. The logical purpose of an algorithm is to distinguish. The selection must include fraudsters and not people who do not commit fraud. The term 'discrimination' here refers to Article 21 of the Charter of Fundamental Rights of the European Union, which states: 'Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited."[124] A clear example of discrimination is disadvantaging a candidate in an application process on the basis of gender or nationality.[125] Discrimination can occur both directly and indirectly. In the case of indirect discrimination, no direct distinction is made on the basis of a prohibited ground. For example, a postcode as a selection criterion. There are postcode areas where many people with a migrant background live. If high-risk zip code areas in an algorithm coincide with this, there may be indirect discrimination based on migration background.[126] In addition to discrimination, there are other risks to fundamental rights associated with the deployment of AI systems. These are often linked to the field of application of the system. See also Box 4.1.

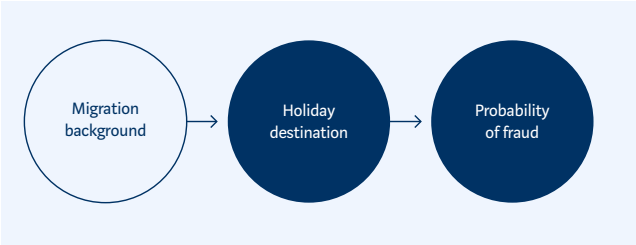
Unrepresentative data leads to unfair outcomes. A statistical model depends on the data with which it is trained. If a group of people is rare in the data, the algorithm will make more false predictions for that group. This leads to discriminatory outcomes when an application is unfavourable.[127]

Example: Suppose that the travel insurer from the aforementioned example does not have a good system for claims that come in another language. These claims first go through the fraud algorithm but the data ends up in a different form in the database without anyone noticing. The algorithm cannot properly learn to deal with claims in another language and will make more mistakes in this area. Consequently, this may have discriminatory effects.

If a protected group does appear in the data but not in a representative way, this can also lead to discrimination.

Example: Suppose the process at the insurer was unconsciously biased for some time. As a result, people in a protected group are more often referred to as fraudsters.

GRAPH 4.2: IF A PROHIBITED GROUND (UNOBSERVED) HAS A RELATION TO AN INDICATOR IN THE ALGORITHM, THERE IS A RAISED RISK OF DISCRIMINATION



The algorithm will adopt this pattern and assign this group too high a risk in the future.

A negative feedback loop exacerbates discrimination. The algorithm learns from previous fraud cases to find new ones. When these new cases are used over time to learn from, a feedback loop is created. This feedback loop can exacerbate discrimination through selection bias.

Example: Suppose that, according to the data, claims from a certain country are more often fraudulent. Claims from that country will then receive a higher risk indication and will be investigated more often. A more intensive search for fraud is in itself a way to find more fraud there. When the algorithm is retrained, this country will therefore get an even stronger focus. This repeats itself: A self-reinforcing feedback loop has emerged.

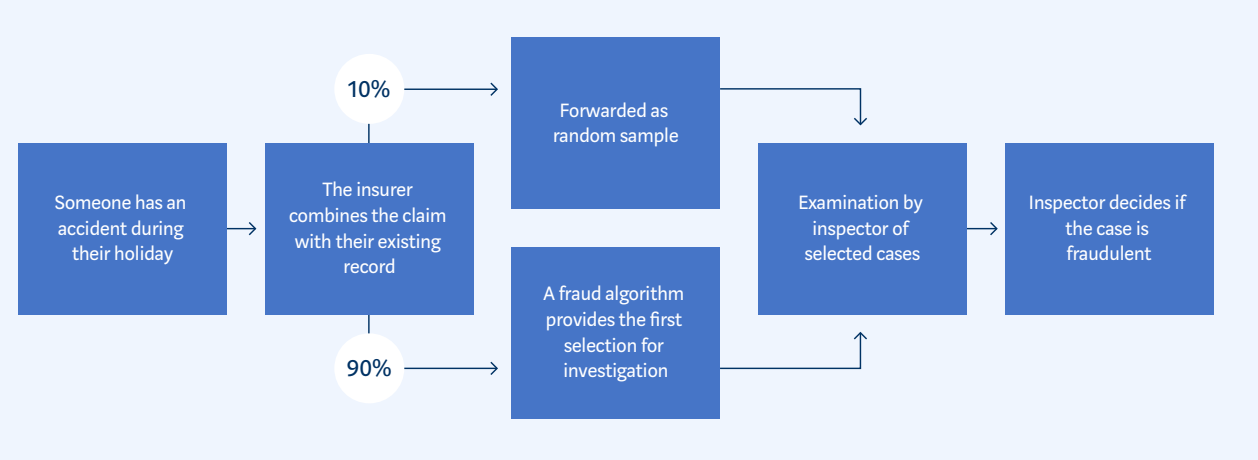
Having sufficient and correct data does not guarantee non-discrimination. Characteristics for the intended distinction are often related by proxies to the prohibited grounds for distinction.

Example: Suppose that migration background has an influence on the country where people go on holiday. And with that also indirectly affect the estimation of the algorithm of the chance of fraud. Migration background is not observed. Assuming that migration background is a prohibited ground for making distinctions, this algorithm may discriminate unintentionally. Even when the data used to train the algorithm is a perfect reflection of reality.

Overreliance on algorithms is an important risk. There is interaction between the human assessor and the outcome of an algorithm. Instead of checking the outcome, people tend to quickly assume the outcome for truth: "the computer will know." This phenomenon is also called automation bias. As a result, it seems as if there is a human intervention as safeguard, when in reality it is limitedly effective.

Overreliance on algorithms gives room for discrimination. As previously mentioned, there can be discrimination in the predictions of fraud algorithms in various ways. Human intervention serves, among other things, as a safeguard against discrimination. If human evaluators over-rely on algorithms, discriminatory predictions can be adopted. It is essential that the human assessor continues to look critically at the outcome of an algorithm.

GRAPH 4.3: AN INSURER CAN ADOPT A RANDOM SAMPLE BY SELECTING A PERCENTAGE OF THE CLAIMS AT RANDOM. THESE CLAIMS ARE THEN FORWARDED FOR INSPECTION REGARDLESS OF THE ESTIMATE BY THE ALGORITHM



4.3 Random sample

One possible measure to reduce the above mentioned risks is the random sample. When a random sample is applied, part of the cases are randomly selected to be investigated for fraud. A random sample of 10% is shown in the image (Graph 4.3).

The optimal percentage of cases that the sample should select differs per fraud algorithm. An important consideration here is that the sample can only serve as a reference if sufficient cases are included.

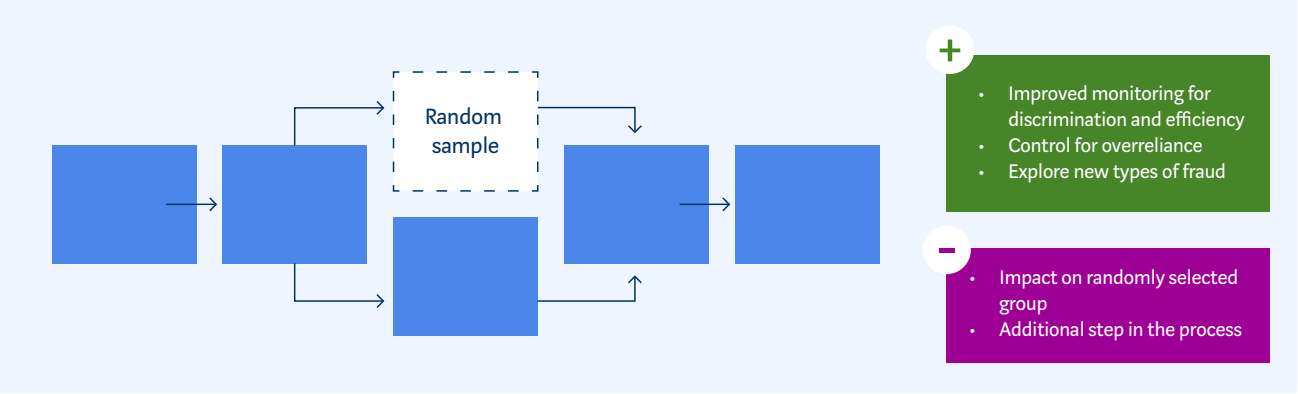
Using the random sample as a reference, part of the risks of discrimination can be monitored. For example, by using the sample as a reference, it can be tracked whether a group is not disproportionately represented by the algorithm. The importance of random selection for obtaining representative data has previously been described by the EU Agency for Fundamental Rights (FRA). Again, here is an example of travel insurance.

Example: Suppose that the fraud algorithm of the travel insurer selects, for 95% of its cases, claims with nationality Y. A random sample is used, and of the fraud found within the sample only 15% of people appear to have nationality Y. It seems here that the algorithm places a disproportionate emphasis on checking people with nationality Y.

Random sampling can reduce the risk of automation bias. When a sample provides part of the selection for fraud investigation, not all investigations are based on a high-risk signal. This allows the process to be set up in such a way that the inspector does not know whether a case to be investigated has been selected randomly. As a result, the inspector can no longer rely blindly on the algorithmic results, but is encouraged to be critical.

The group to which the random sample applies varies with context. In some applications, a fraud indication will not mean selection, but rather an exclusion. Blocking an online order is an example of this. In such a case, there is a different process overview. A sample can be used here through randomly passing on high-risk cases. There are also applications where the impact on the randomly selected individuals is significant. Think of home visits as part of a fraud investigation. A balance of interests is central to this: banking on a sample does not directly mean that you can search someone's house.

GRAPH 4.4: THE DECISION TO USE A RANDOM SAMPLE DEPENDS ON THE IMPACT ON THE PROCESS AS A WHOLE



In addition to risk reduction, a random sample contributes to measuring efficiency and exploring new types of fraud. The reference provided by a sample can be seen as a basis for comparing the performance of the algorithm. In addition, the element of arbitrariness will ensure that unknown and new forms of fraud also occur in the data over time. A new version of the algorithm can then learn from this.

4.4 Worth considering

The decision to use a random sample depends on the impact on the process as a whole. The technique touches on different parts of the process and can therefore not be weighed against a stand-alone risk. In order to make the decision transparent, a system without a sample can be compared to a system with a sample (see Graph 4.4).

In many cases, the random sample contributes to a more responsible use of profiling and selecting AI systems. When using fraud algorithms, the random sample is therefore worth considering.

The use of a random sample relates to the AI Act and the GDPR. A system around a fraud algorithm must comply with legislation (with or without a sample). A profiling or selecting AI system processes personal data and thus the GDPR applies. In addition, the AI Act will enter into force in the coming years. Within the AI Act, it is relevant whether an AI system classifies as high-risk. For example, for a fraud algorithm for essential government benefits and services, this will be the case. Providers of high-risk systems are required to identify and mitigate reasonably foreseeable risks. Potential discrimination is such a risk. The random sample can be used as a control tool.

Certain methods of combating discrimination depend on the processing of special personal data. For example, to measure whether a group with a certain origin is treated differently, information about race or ethnicity will have to be processed. These sensitive data, known as 'special personal data', are given additional protection in the GDPR. The processing of special personal data is prohibited, unless there is an exception. The AI Act may provide for an exception for the processing of special personal data to detect or counter discrimination for high-risk AI systems but under strict conditions.

Box 4.1

Fundamental rights risks in deployment of AI systems

The use of AI systems can lead directly and indirectly to a violation of fundamental rights. This risk affects both very simple algorithms such as decision trees and complex systems,for example, those based on neural networks.

A known risk in AI systems can be seen in the right to non-discrimination (Article 21 of the EU Charter of fundamental rights), but there are also risks to other fundamental rights. Non-discrimination issues often stand out because in many cases AI systems are used to make a distinction. In addition, it must be explicitly examined whether the distinction can be legally justified. But there is also the fundamental right to fair and just working conditions (Article 31 of the Charter). This can be put under pressure by algorithmic management. For example, when this impairs healthy, safe or dignified working conditions. Another example is the protection of personal data (Article 8 of the Charter), a fundamental right that ensures that data are processed fairly with the consent of the data subject or on the basis of a legal framework. AI systems work on the basis of large amounts of data which often includes personal data. These personal data should be processed lawfully and in line with this fundamental right, including in the case of an AI system. The AP, as an independent authority, monitors to see compliance with these rules.

Two other relevant fundamental rights are freedom of information and the right to good administration. Both fundamental rights are reflected in this report and are relevant in the context of AI systems. The fundamental right that affects media freedom and pluralism is important in the deployment and use of AI in the online information provision (see Chapter 2). The right to good administration ensures that matters are dealt with impartially, fairly and within a reasonable time by public institutions and bodies. The use of simple algorithms and AI systems in the public domain has, in recent years, led to risks and incidents that are difficult to reconcile with this right.

Fundamental rights risks and violations in the deployment of AI systems still go unnoticed too often. Even where specific legislation does not address all aspects of new technology, or where specific legislation does not yet exist, fundamental rights will always have to be respected and protected. To address this in more detail, the AP will present a factsheet on fundamental rights risks in the deployment of AI systems later this year.

5. Policies and regulations

QUICKLY TO THIS SUBJECT

More and more attention is being paid worldwide to the regulation of AI and algorithms. The entry into force of the European AI Act on 1 August 2024 is a milestone. In doing so, some provisions will already enter into force on 1 February 2025. For example, provisions for prohibited AI applications and for AI literacy within organisations. The long transition period (until August 2030) before existing high-risk AI systems within the government have to meet all requirements is a matter of concern. Another point of attention is that the product standards are completed under high time pressure. These are decisive for the actual effectiveness and practicability of the requirements of the AI Act. In the meantime, supervisors in the Netherlands are preparing for new supervisory tasks under the AI Act.

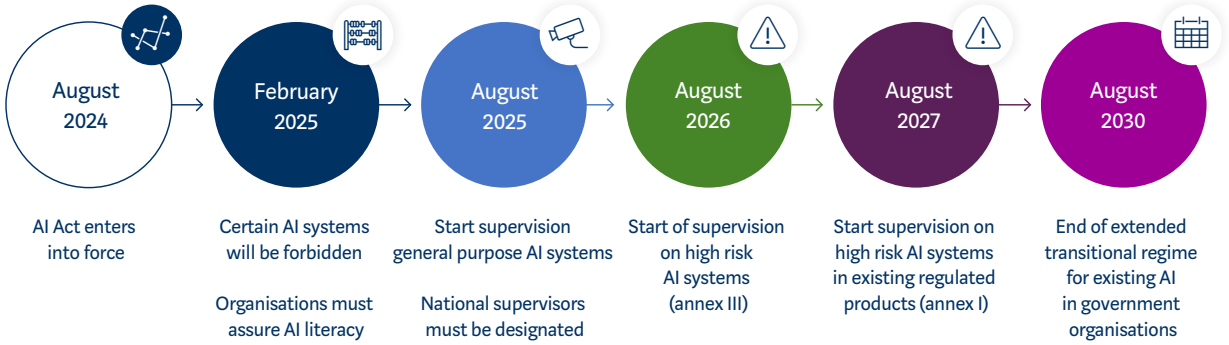
The broad coalition agreement of May 2024 provides good starting points for further work at national level to adequately manage the risks posed by AI systems. It is important that the filling of the algorithm register remains a priority and becomes an obligation, with attention to sufficient scope. The work on algorithm frameworks provides organisations with a governance tool for AI systems and this is important. At the same time, attention must be paid to frameworks that are too non-committal or that consciously or unconsciously give room for insufficiently precise or measurable standards, which sometimes lag behind or conflict with scientific insights. From the point of view of the AP, the appointment of a new cabinet is a moment to review the national AI strategy.

5.1 AI Act enters into force

On 1 August 2024, the AI Act will enter into force and begin the transitional period before the AI Act will be fully applicable. For companies, governments and other organisations that work with AI systems, this is the time to start preparations to make systems and organisations compliant (see Box 5.1). Now that the requirements have been established, the different parts of the Regulation will enter into force soon, but in gradual steps. The first major component to which this applies is the provisions on banned AI applications on the European market. This section shall enter into force in February 2025. Broadly speaking, requirements from other components will also enter into force between February 2025 and August 2027 (see Graph 5.1).

A great deal remains to be regulated and clarified, both at national and European level, in order to ensure that these requirements are met in a workable manner. Many concepts and procedures in the AI Act require further explanation or interpretation. It should also be made clear how developers can meet the requirements for high-risk AI systems by means of European standards. Furthermore, oversight of AI systems at national level needs to be set up quickly and with sufficient resources.

GRAPH 5.1: IN THE COMING YEARS, THE AI ACT WILL ENTER INTO FORCE IN STAGES



It is also positive that the requirements for AI literacy within organisations will be applied quickly. Organisations that develop or use AI systems must ensure that there is sufficient knowledge of AI among the employees who work with the AI systems. The level of knowledge must be in line with the experience and training and knowledge of the staff, but also with the context in which the AI systems are used and how the systems' user groups can be affected. This is a general provision, applicable to all AI systems and will apply as early as February 2025. This provision is also an important step towards the obligation for users of AI systems to ensure adequate human oversight.

Less positive are the over-extended deadlines given to public organisations to comply with the AI Act. Most high-risk systems that are of will be t in operation before February 2026 will only be required to comply with the AI Act if there is a significant change in the design of the system after February 2026. AI systems that are subsequently placed on the market will have to comply with the requirements immediately. AI systems intended for public sector organisations shall comply with the requirements only by February 2030. The AP is concerned about the duration of the period for AI systems which are intended for government organisations. The transitional regime also creates a perverse incentive for companies and the government to make significant changes before February 2026 and not after that date. Practice has shown, especially with existing AI systems – and underlying algorithms – that fundamental rights and values may be at stake. At the same time, the AP stresses that a higher level of ambition is possible and that this can be enshrined in national legislation. An executable transition plan for existing high-risk AI systems consists of (i) mapping them first, (ii) having public organisations draw up a plan on how to make self-developed high-risk systems compliant and (iii) prescribing when – earlier than February 2030 – these systems must comply with the AI Act. In addition, governments should only purchase systems that already comply as much as possible with the rules in the AI Act.

5.2 Prohibited AI

Some AI applications will be banned as of February 2025, but the precise scope of these prohibitions requires further clarification and explanation. The bans are an important part of the AI Act, as there will soon be systems whose placing on the market or putting into service must be fully restricted. This will apply to systems that pose an unacceptable risk because they for example limit people's free choice too much, exploit people or manipulate people. Only when a practical and more concrete interpretation of the AI systems which are included in the prohibitions is quickly made clear, can European supervisors intervene or not intervene. This provides certainty to the market and society. The AP therefore welcomes the clarification that the European Commission will provide with guidelines. A first guideline will specify the definition of an AI system. A second guideline will provide more clarity on the implementation of the prohibited AI applications.

The AP will do a call for input this year to explore the concrete implementation of the prohibited AI applications. Stakeholders will have the possibility to deliver input on two prohibited product applications included in the AI Act. The aim is to collect knowledge and practical questions from organisations, in order to ultimately enable adequate supervision and legal certainty. This also provides a basis for further policy-making with other European supervisors, for example in the context of the guidelines to be developed.

5.3 AI standards

European product standards are essential for compliance with the AI Act. Such standards provide AI developers with guidance on the requirements of the Regulation. In the standardization process, however, the time pressure is high. Also, the results in the form of product standards are not yet freely accessible. Standardisation organisations CEN and CENELEC are developing standards to further specify the requirements of the AI Act. When organisations work according to these standards, it is assumed that their high-risk systems meet the requirements set out in the AI Act. In practice, the standards will therefore play a major role in demonstrating compliance and assessing conformity.

However, the AP is concerned about the speed with which the standards must be delivered. Standardisation organisations have only three years from the European Commission's standardisation request to develop standards. However, drafting technical product standards usually takes a long time. Moreover, delivering standards for the AI Act is even more complex, given the broad focus on both health and safety and fundamental rights. Providers and users of AI systems should therefore take into account that the standards may be available at the same time as, or only after, the entry into force of the provisions on high-risk applications. Policy makers must work on a scenario in which organisations must comply with the product requirements of the regulation before they can use the standards.

It remains a point of contention that these product standards will in principle only become accessible after payment. Especially since these are standards that should contribute to the protection of fundamental rights and values. Because the standards are not generally accessible, they are less likely to have an impact on the general AI literacy that is necessary throughout the organisation and society. It also creates an additional threshold for the general public to scrutinise an important elaboration of the AI Act. At the same time, it should be recognised that it is a well-established practice that product standards – and the underlying standardisation process – are financed in this way. If policy makers choose to make the product standards publicly available, a suitable solution must therefore be found for this.

Box 5.1

Start preparing for the AI Act

Companies, governments and other organisations that use or develop AI would do well to prepare for compliance with the act. We recommend that providers and users of AI systems immediately start with an implementation plan and set up the internal risk management. A first step is to identify the systems they develop or use and whether they fall within the definition of 'AI system' in the Regulation. Subsequently, they usually have to assess whether these systems fall into one of the following risk groups:

1. **Prohibited AI.** These systems must be withdrawn from the market and their use must be stopped. The provisions on prohibited AI will already be in force from February 2025. There is also a good chance that these systems are already violating the law, such as legislation in the field of equal treatment, privacy or employment legislation.

2. **High-risk AI.** These systems must comply with requirements such as risk management, the quality of the data used, technical documentation and registration, transparency and human oversight. Governments or performers of public tasks may be subject to additional requirements, such as carrying out a 'fundamental rights impact assessment'.

3. **Low-risk AI.** Systems intended to engage with individuals or that generate content, such as deepfakes, are subject to transparency obligations. If these systems are offered or used, people should be informed about them.

Users of AI systems can already assess whether their AI systems comply or will comply with the AI Act. They can inform providers to what extent the AI used already meets the requirements. When purchasing an AI system, it is important to check the purchasing conditions and, for example, also pay close attention to what happens to data that the AI system processes and how the rights to that data are regulated. Work is also ongoing on standard contractual clauses for the procurement of AI in the public sector from the European *public buyers community*.[130]

Roles and responsibilities in the AI Act may overlap and shift. Both providers and users must comply with certain obligations. Organisations that develop and use AI systems themselves essentially fulfil both roles and must therefore comply with all obligations. The roles can also shift. Is a purchased system modified or used for another purpose? Then an organisation can become a provider of a system, which must then, for example, adhere to the rules for high-risk systems. We therefore recommend that users of AI systems keep a close eye on whether (and if so, how) they themselves contribute to the development of the AI systems they use.

Developers of AI systems can go to the Dutch regulatory sandbox. The AP, the National Inspectorate for Digital Infrastructure (RDI) and the Ministry of Economic Affairs are working on the preparation of the sandbox. Providers will be able to go here in the course of 2026 to solve compliance questions about the AI Act. Until then, the supervisors will test the functioning of the sandbox and its processes in pilot schemes.

5.4 Supervision of the AI Act in the Netherlands

The ongoing preparations in the Netherlands for the supervision of the AI Act deserve continued attention. In May, the AP, together with the RDI, issued an opinion to the ministries of EZK and BZK on how to properly regulate the supervision of the AI Act. [131] [132] This advice is the result of a collaboration of more than ten different supervisors, colleges and inspections. The AP has played a coordinating role in the drafting of this opinion. In the opinion, the supervisors elaborated on how the supervision of the AI Act should be organised and who should be designated as so-called market surveillance authorities.

The AI Act provides an opportunity to strengthen oversight of AI. However, this requires sufficient capacity and the right framework conditions. First of all, it must quickly become clear which authorities will carry out the different parts of the supervision. Then, sufficient budget and staff must be made available in time to all supervisors involved, so that they can start their tasks (such as information and enforcement) on time. Last of all, AI is a system technology that is not only regulated by the AI Act. It is therefore important that the supervision of the AI Act and the existing supervision of existing legislation reinforce and complement each other. For example, in the supervisory relationship in the field of consumer protection, data, education and employment. Setting up supervision of AI is a task in which the RDI also cooperates with UNESCO on behalf of the Netherlands and Europe (see Box 5.2).

5.5 International developments in AI regulation and policy

The need to regulate AI is also strongly felt outside the EU. Wordwide initiatives are being taken, however with differing approaches. Various countries are working hard on legislation, standards and principles for the regulation of AI However, there is a risk of fragmentation due to lack of coordination. Globally, there are different responds to the developments around AI. This depends on the specific challenges faced by countries and on underlying values that are prioritised. For example, the focus in the global south is mainly on the opportunities that AI creates for the national economy, whereas in Europe and the US, for example, policies and regulations are rather focussed on the risks of further developed AI.[133] This is also in line with the differences in risk perception between citizens from these regions, as discussed in Chapter 1. In the EU, individual rights have an explicit position in internet policy and regulation. The US currently shares the European ambition to ensure that AI is trustworthy, safe and contributes to the protection of human rights. The visions differ in the trade-off between stimulating innovation on the one hand and safeguarding social values on the other.[134] Another angle for AI regulation is visible in China, where the national interest is emphatically paramount.[135]

International developments increasingly provide a basis for cooperation in the development of AI systems. In China, a more comprehensive law on AI is on the legislative agenda for the first time even though the regulation of AI in China has been more patchy so far. Regulation does not target AI as a whole, rather its purpose is to get a grip on specific systems or products, such as recommender systems.[136] In addition, in Brazil, for example, there is a new law that seeks to regulate the development of AI in a way similar to the approach taken by the EU in the AI Act.[137] [138] It is visible that concepts such as reliability, security and human rights are being given more attention in national regulatory initiatives worldwide. At the same time, careful attention should continue to be paid to the differences in challenges, perspectives and regulatory initiatives. Fragmentation in national approaches can lead to tensions when AI is used across borders.

Inclusive international standards and norms can contribute to the necessary harmonisation. International norms and standards are important for the safe development and deployment of AI on a global level, both by private parties and by national governments. Inclusive international standards can reduce the gap in policy and regulation and contribute to harmonised implementations. International organisations such as the OECD, UNESCO and the G20 play an important role in harmonising standards and regulations.[139]

The OECD published an update of the OECD's AI principles in early 2024. The use of these principles by OECD member countries (currently 38 countries) provides a basis for global interoperability between countries. The principles guide the trustworthy development of AI and provide recommendations for government policy and strategy for AI. In 2019, the OECD member states signed these principles. The OECD has updated the principles this year to reflect new developments around AI systems, in particular the emergence of foundation models and generative AI. The updated version addresses challenges to privacy, intellectual property rights, security and information integrity in the context of misinfor-

mation and disinformation. It also underlines the importance of responsible business conduct and good cooperation in policy and governance.[140]

A first global resolution on AI was adopted at the United Nations in March 2024.[141] This non-binding UN resolution has been tabled by 122 countries, including China. The resolution calls on states to guarantee human rights and to ensure the reliable development and deployment of AI through regulation and governance. The resolution also calls for closing the digital divide, so that countries in which the development of AI is less advanced can also benefit from the opportunities that AI brings.[142]

A global institute for AI governance can be an important step towards strong common standards. Although not formally binding, the UN resolution reflects an international consensus on the standards to be followed in the development and use of AI. Previous, similar initiatives have shown that such agreements provide guidance for the drafting of national policies and regulations.[143] Nevertheless, it should not be assumed that such standards will reduce fragmentation in national strategies and regulatory initiatives. In addition to national governance, an international governance framework should therefore be developed.

To explore this, a special advisory body has been set up within the UN. The High-Level Advisory Body on Artificial Intelligence recently delivered the interim report 'Governing AI for Humanity'.[144] It calls for the strengthening of international governance. The advisory body emphasises an inclusive global approach to achieving harmonisation. This is partly achieved by setting up a global AI governance framework, which covers the following functions: (i) identifying and monitoring AI developments, (ii) building consensus on international standards and (iii) monitoring systemic vulnerabilities to global stability. A global institute, in which various stakeholders are involved, can keep an eye on global developments through these activities. And these insights can be used to create inclusive international norms and standards. It is also important to note that the supervision of AI should not be neglected. In response to this interim report, the AP, as coordinating algorithm supervisor, therefore wrote and published a discussion paper in which the importance of national and international supervision in the governance of AI is also raised.[145][146] After all, supervisory authorities are able to monitor developments and risks at an early stage. Together, they can contribute to the development of guidelines and standards for the responsible use of AI.

5.6 National developments in AI regulation and policy

The coalition agreement for the formation of the current Dutch cabinet contains important provisions on algorithms and AI, which can strengthen current policies. In the agreement, it was agreed that there would be a scientific standard for the use of models and algorithms. This standard sets out the requirements that these are public and exemplary, with a clear instruction for what these models and algorithms are, for what they may be used for and for what they are not intended to be used for.

The AP welcomes these requirements for the use of models and algorithms. However, the requirements should be considered in conjunction with the provisions of the AI Act. The AI Act contains similar requirements. The Regulation requires high-risk systems to meet quality requirements, taking into account the purpose of use and the generally recognised state of the art.

One of the requirements of the AI Act is that AI systems must be accurate in the purpose of their use. This prevents arbitrariness by algorithms and AI systems, which is one of the objectives for which the AP is committed to in the coordinating algorithm task. The AI Act will lead to the development of benchmarks and measurement methods for assessing AI systems for accuracy and robustness. For example, in cooperation with metrology and benchmarking authorities (see Article 15 of the AI Act).

The AI Act also states that high-risk AI systems should be designed in such a way that their functioning is sufficiently transparent. This enables organisations using such a system to interpret the output of the system and use it appropriately. The instructions for use for a high-risk AI system should also address the purpose and level of accuracy (see Article 13 of the AI Act).

The aforementioned coalition agreement sees the benefits of the use of AI by the government, but also acts on the risks. There is an ambition to strengthen knowledge of digitalisation within the government. At the same time, it was agreed in the coalition agreement that special conditions are attached to the use of AI by the government to ensure safety, privacy and legal protection. The AP sees this as a positive ambition. The most important tool to manage the use of AI by the government is the most proactive possible preparation for, and implementation by government organisations of, the requirements of the AI Act. The importance of (i) the broad definition of AI system (see Key Messages in this ARR) and (ii) a level of ambition for the government that exceeds the lower limit in the extended transition period for existing AI systems (see the section 'AI Act enters into force' in this chapter). In addition, the knowledge requirements for AI literacy (as of 1 February 2025) provide an opportunity to increase the level of knowledge about AI (and therefore digitalisation in a broader sense).

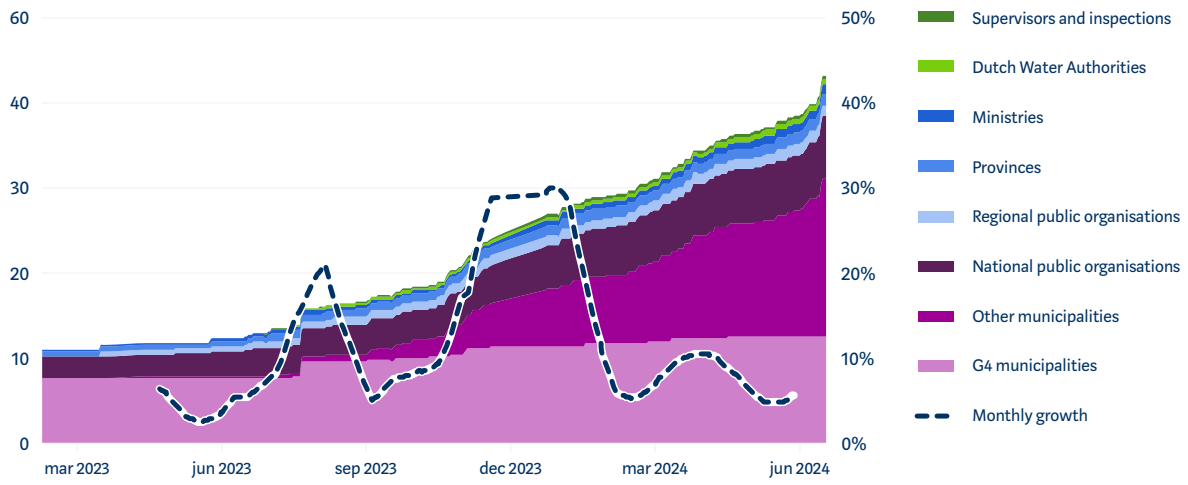
On the basis of the coalition agreement, the new government must also make society more resilient against disinformation and deepfakes. This is in line with the observations made by the AP based on its coordinating algorithm oversight on the impact of AI on information provision in democracy (see Chapter 2 of this ARR) and the emergence of deepfakes, which are easier to produce on a large scale by generative AI (see Chapters 1 and 2 of this ARR). The agreements to strengthen the approach to digital threats are also in line with the risk picture that follows from this ARR. The observation is that malicious parties can use (generative) AI in a disruptive way (see Chapter 1 of this ARR).

In the short term, it is important that algorithm registration remains a priority. Registration is essential in order to bring transparency to citizens and to gain insight into their own algorithm use. Registration is also a good basis for risk management. Over the past six months, the national Algorithm Register for public organisations has been further filled, to more than 400 algorithms (end of June 2024). It is striking that at the end of 2023 there was a peak in the growth of the number of registrations. The national Algorithm Register is a good way to respond to the call for registration. Keeping enough pace in the growth of this register is important to achieve full registration. The growth in recent months has mainly come from registrations by municipal organisations. It is striking that regulators and inspections, water boards, ministries, provinces and regional government organisations have so far placed few algorithms in the register (see Graph 5.2).

The government has agreed with the House of Representatives to register all high-risk AI systems within the central government by the end of 2025 at the latest. The AP would like to see a broader scope and clarity about the consequences if organisations do not meet this deadline. The AP remains in favour of registration by government organisations being made mandatory quickly. The Algorithm Register could also provide insight into the organisational control of AI and algorithms, and transition paths to support algorithm registration and AI control, for the government organisations that register AI systems. The AP also stresses that the scope of the Algorithm Register must be broad enough. Deciding whether or not a particular AI system (or algorithm) is a high-risk application can be difficult. That is why it is important that political bodies and third parties can also assess this trade-off. Algorithm registration provides the basis for this. Registration obligations should therefore not be limited to algorithms for which it is already certain (in advance) that they are at high risk.

An additional point of attention is the extent to which algorithms and AI systems in the quaternary sector are in sight and registered. Organisations in health care, education, public housing and public transport contribute to essential services. However, the scope of the current Algorithm Register does not extend to these organisations, which are remote from the government. At the same time, AI systems are increasingly part of how these organisations provide their services. Due to the lack of registers, it is currently difficult to gain insight into how these organisations use algorithms and AI and how risk management is doing.

GRAPH 5.2: THE NUMBER OF REGISTERED SYSTEMS IN THE ALGORITHM REGISTER FOR PUBLIC ORGANISATIONS HAS BEEN GROWING IN THE PAST PERIOD, MAINLY THANKS TO MUNICIPALITIES, THE GROWTH RATE IS DECREASING AFTER A PEAK AROUND YEAR-END 2023



Explanation: *development of registered algorithms on algoritmes.overheid.nl from 13 March 2023 until 30 Juni 2024*

Earlier this year, the government published a government-wide vision on generative AI. A positive aspect of the vision is that it highlights and recognises the risks involved in the deployment of generative AI. Privacy and copyright risks are specifically identified, as is the potential market power of big tech companies. The publication of this vision strengthens public debate and can give an impetus to future policy. Points of attention that the AP sees are that the technology is in full development and that the interpretation of this is limited. In addition, there are indirect consequences that are only taken into account to a limited extent, such as algorithm distortion. More attention is also needed for society-wide education. The way in which generative AI systems meet data protection requirements remains a major concern for the AP, as previously highlighted.[147]

The government is also working on a new algorithm framework. The design of the algorithm framework seems supportive. According to the government, the framework aims to provide a practical overview of 'the main existing standards [...] and measures that can help with that'. In addition, the algorithmic framework includes 'Guidelines and measures that are not mandatory, but serve as guidance for safeguarding fundamental values. The requirements of the AI Act will still be included'. Points of attention in such a support framework are possible non-committal and how organisations implement open standards (see also the section 'Frames, standards and tools for the deployment of AI' in this chapter). That is why it is important that there are also requirements for the build-up (or enforceable build-up) of AI knowledge, AI governance and AI strategies within government organisations. This also explicitly requires investments in personnel, IT and education, with associated financial resources and managerial responsibility within organisations. This will make it possible to gradually improve the control of the risks of algorithms and AI.

In the AP's view, the appointment of a new cabinet is also a moment to review the national AI strategy. This has been discussed earlier in Chapter 1 of this ARR. The current national AI strategy is the Strategic Action Plan for AI (SAPAI) of October 2019. Due to the turbulent development of AI technology over the past four years, it is appropriate to review and re-establish the strategy. This is necessary in order to respond to new challenges and the further social transition that needs to be made in the coming years.

A national AI advisory board can also play a role in this, as policy makers are currently exploring. The AP sees room for a multi-stakeholder approach, in which the advisory board brings together knowledge from science, supervision, policy, practice sectors and also the citizens perspective.

5.7 Frameworks, standards and tools for the deployment of AI

There is a risk of a proliferation of sub-frameworks (including review standards and implementation tools) in the absence of sufficiently precise, complete and measurable standards for AI systems. This entails two risks. First, given its partial nature, the broad range of frameworks can provide fake certainty. For example, if (i) an implementation tool gives a lot of freedom of interpretation, (ii) criteria are difficult (objectively) to measure or weigh and (iii) there are no competence requirements for the users of the framework. Secondly, the frameworks can be used selectively to legitimise the deployment and performance of an AI system. For example, when a certain framework focuses on a certain type of criterion and/or compliance from a certain angle (this does not detract from the usefulness and necessity of this framework for the intended purpose). It is not the intention that the outcome is applied more broadly to other domains or angles from which a system must also be assessed.

At worst, organisations can make the situation appear rosier than it is, which can lead to a form of AI ethics washing. For example, when an organisation commits itself on paper to an ethical risk management of AI and provides procedural evidence for this, but in practice does not sufficiently follow up with actual measures to manage the risks for a long time.

A recent report by an international think tank points out that there are serious shortcomings in many frameworks and instruments. In December 2023, the World Privacy Forum published a report on 'AI supporting governance tools', the umbrella term for guidance, assessment frameworks, frameworks and similar tools.[148] A study of nearly 20 of these instruments shows that nearly 40 percent refer to measurement methods that, according to scientific literature, are inappropriate or irrelevant when measuring AI systems. One example is prescribing the 80% rule for assessing the bias of an AI system, while this is a measure that is unsuitable for many applications. It is also striking that there are major differences in the form of these types of instruments. Sometimes a 'framework' is limited to practical guidance, possibly supplemented by a questionnaire for a self-assessment. In other cases, the framework also includes a technical framework, including software, scores and scales to assess the outcome, including thresholds to determine whether an AI system complies with that framework.

Box 5.2

Unesco strengthens AI supervision in the Netherlands and the European Union

Bij: National Inspectorate for Digital Infrastructure

The Dutch Authority for Digital Infrastructure and other European supervisors face a joint task: effective supervision of AI. This task presents supervisors with some challenges. AI transcends both physical and digital boundaries and therefore requires a coordinated approach. Supervisors should also take into account both existing and new legislation, such as the AI Act. In addition, there is still little clarity in the form of, for example, guidance or best practices. And perhaps the most important challenge: not all supervisors currently have sufficient experience and knowledge of AI supervision.

The Dutch Authority for Digital Infrastructure (RDI) is working with UNESCO to develop the capabilities of the European regulators to monitor AI. In light of these challenges, the RDI, as chair of the European and Dutch working groups of AI supervisors, requested support from the European Commission on behalf of other European supervisors. The European Commission therefore called on UNESCO to assist supervisors in the challenges of effective supervision of AI. The collaboration complements existing activities, such as those of the Dutch and European working group of AI supervisors.

The mission to UNESCO is: "Provide support to the RDI and members of the Dutch and European working groups of AI supervisors to strengthen their supervisory capabilities in line with the AI Act and other relevant legislation."

Cooperation between RDI and UNESCO has a number of concrete objectives: First, a baseline measurement based on a comprehensive report on the current practices of AI supervision in Europe and beyond. AI systems are not limited to specific domains and national borders inside or outside the EU. Therefore, a broad scope for developments is necessary. Second, the development and discussion of case studies on AI supervision with the members of the European and Dutch working groups chaired by the RDI. Third, drafting and disseminating a set of best practices for dealing with specific AI oversight issues. Fourth, explore approaches and options, and present them to relevant stakeholders within AI oversight. A fifth objective is to train supervisors, based, among other things, on the best practices developed.

In addition to improved capacity, the cooperation provides more uniform supervision. UNESCO and the RDI focus primarily on inspectors who have to supervise AI systems. However, because different supervisors learn the same lessons, they are also taught the same supervisory approach. This results in more uniform supervision by the various national and European authorities.

The collaboration will lead to tangible results in the short term. A first report will be delivered in mid-2024 and the other objectives will follow in stages. The goal is that the project will be completed by the end of 2025.

This Box was written by the National Inspectorate for Digital Infrastructure, which supervises the availability, continuity and reliability of the digital infrastructure in the Netherlands.

Appendix: What makes managing AI risks so complex?

From impact assessments, ethical standards and implementation frameworks to evaluation frameworks, fairness metrics and transparency obligations: all safeguards that contribute to the protection of fundamental rights and values when deploying AI systems. But how do these concepts relate to the entire life cycle of AI systems?

There is no silver bullet for managing the risks of AI systems. Responsible use of AI within organisations, or the provision of AI systems to private end-users, requires interaction and coherence between risk control measures in the development, deployment and evaluation phase of an AI system. But responsible deployment or offering also affects the overarching culture, ethics, knowledge and governance that are required for this at the level of the organisations that develop and/or deploy AI systems. This is precisely why it is difficult for organisations to get an overview of the many frameworks they are provided with and the corresponding perspectives and accents (see also Chapter 5).

The outline in this appendix gives an overview of the coherence in the risk management of a simple AI system. The sketch provides a non-exhaustive overview of building blocks for the risk management of a simple AI system that is developed and deployed within the same organisation (see infographic). For example, by a government agency. In a way, this is the simplest situation that can occur. The control framework becomes more complex once multiple organisations are involved and multiple AI models (algorithms) are interrelated in a process based on an AI system. In general, it can be said that the required risk management framework must be a form of customisation for each individual AI system, depending on the objective, autonomy and context in which the AI system is deployed.

The foundation of the control of an AI system is provided by (i) behavior and culture and (ii) governance of the organisation involved in the AI system. For example, the importance of ethical awareness, diversity and due diligence by individuals involved in the development and deployment of AI systems is often discussed. These are organisation-wide issues and the same applies to the AI governance of an organisation. Good governance provides clarity about who is finally responsible for the deployment of AI within the organisation and creates frameworks for how control and knowledge of AI take shape within the organisation. An example of an overarching organisation-wide requirement can be found in Article 4 of the AI Act. This article requires organisations which are deploying AI systems to ensure an adequate level of AI literacy among their staff.

At organisational level, the first step is to determine the purpose of (the possible exploration of) the deployment of an AI system and to make a balanced decision to this effect. The sharp determination of the explicit or implicit purpose of the AI system provides the basis for the assessment framework: What should the AI system be used for? It also allows for a trade-off: What are the benefits of the AI system? With what certainty can these benefits be achieved? And how do these benefits outweigh the disadvantages and risks, and the uncertainty about whether these disadvantages and risks actually occur – including associated control costs? Proportionality also plays a role in this. It is also crucial that it is predetermined that there is an intention to deploy or develop an AI system. In the case of public organisations, democratic legitimacy also plays a role here. Case studies show that this is still often insufficient – see chapters 1 and 3, but also the first edition of the ARR (summer 2023).

The continuous risk management of an AI system within an organisation then requires a continuous cycle of (i) development and implementation, (ii) deployment and (iii) evaluation. Within each part of the cycle, there are different building blocks, each contributing in their own way to the management of AI risks throughout the entire life cycle. For example, predetermining the fairness standards to be used during the further development and implementation phase helps to assess whether an AI system meets the standards at the end of the evaluation phase. Equally, ensuring registration and transparency of an AI system allows information about the AI system to be available to the public during the deployment phase. This contributes to the ability to report (and process) incidents with AI systems.

During the further development and implementation phase, extensive testing takes place and the conditions are set up that must enable the actual deployment of the model AI system. The diagram gives a representation of some building blocks that can be thought of. These are also partly reflected in requirements from the AI Act, but also other laws and regulations. Before the system can be deployed, requirements must be met in terms of quality control (see, inter alia, Article 17 of the AI Act) and risk management – in fact, the cycle of risk identification, assessment, evaluation and control measures (see, inter alia, Article 9 of the AI Act). When implementing an AI system, a first building block is a fundamental rights impact assessment, which may be mandatory for those responsible for use (see, inter alia, Article 27 of the AI Act). This assessment consists of identifying, weighing and addressing fundamental rights risks. Such an assessment can also be partially published, or partially published, for example in a register. Related to this is the Data Protection Impact Assessment (DPIA), which identifies privacy risks in advance and enables organisations to take measures to reduce them.

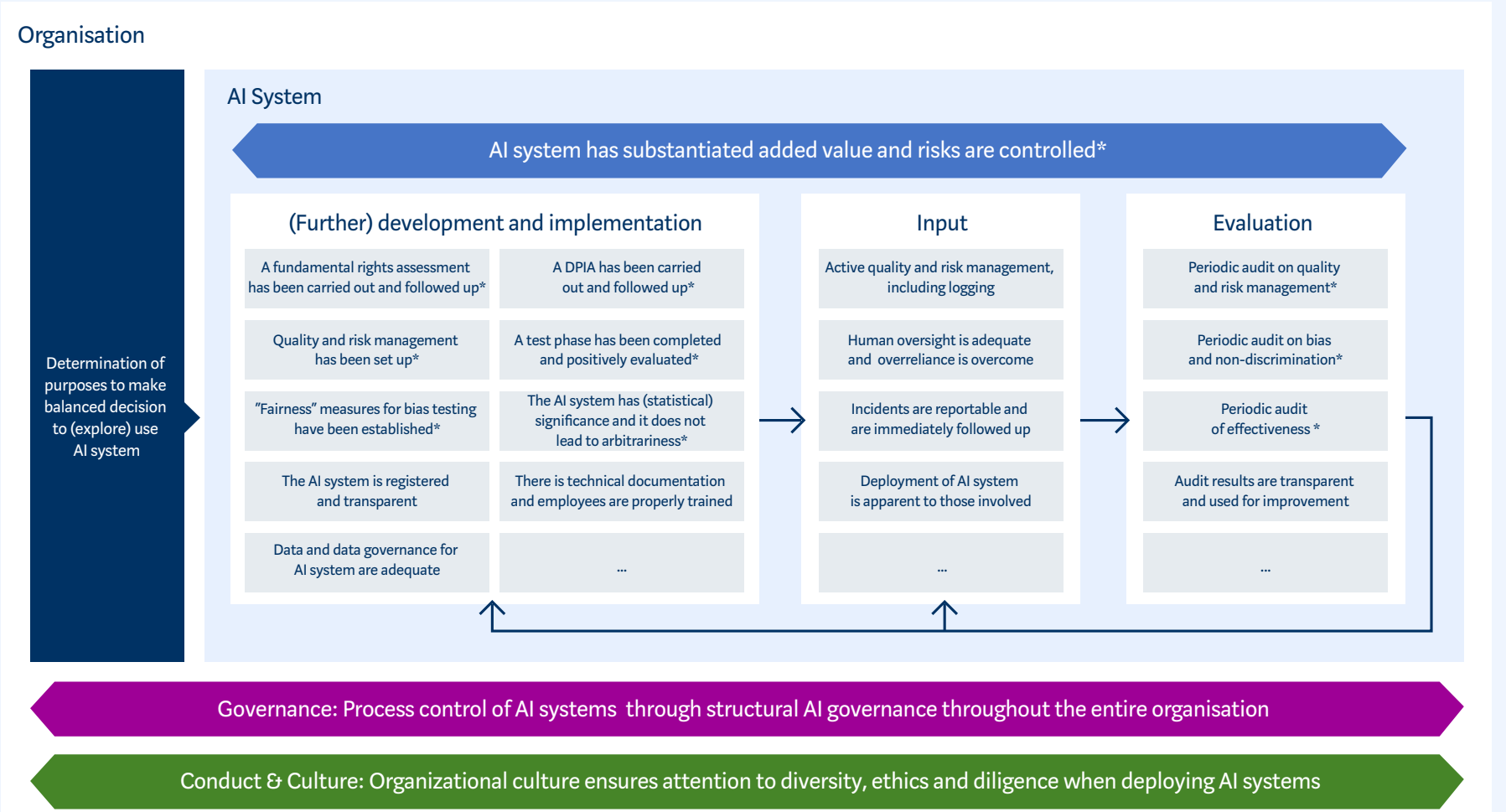
Attention should also be paid to testing, documentation and registration in the development of AI. Part of the further development is going through a test phase on the basis of predetermined assessment criteria and reliability thresholds that are appropriate for the intended purpose (see, inter alia, Article 9 of the AI Act). Determining fairness metrics is linked to this (see, inter alia, Article 10 of the AI Act), but at the same time the AI system must be accurate in relation to the purpose determination and not lead to arbitrariness (see, inter alia, Article 15 of the AI Act). Registration of an AI system placed on the market contributes to transparency and traceability of a system and is for AI systems with a high-risk element of the conformity procedure to be followed when placing such systems on the market (see, inter alia, Article 49 AI Act). Similarly, technical documentation for use must be drawn up (see, inter alia, Article 11 and Article 13 of the AI Act), which must also guide use when implementing an AI system. Furthermore, the usability of an AI system depends on data quality, for which data governance must also be in order (see, inter alia, Article 10 of the AI Act).

During the deployment phase, users and third parties should be able to quickly identify and correct hazards and incidents with an AI system. A first building block for this is logging of the events during the deployment of the AI system, so that the cause of incidents can be determined (cf. Article 12 of the AI Act that requires developers to build in logging capabilities). Excessive trust in systems can limit automated decision-making. Should an incident occur, a user or third party must be able to report it so that recovery is possible (cf. Article 72 AI Act). Transparency and explainability requirements help to make the involvement of the AI system visible to users or third parties (see, inter alia, Articles 50 and 86 AI Act, but also the GDPR in the area of automated decision-making).

The periodic review of the AI system shall ensure that the system is efficient and qualitatively sound and that fundamental rights risks are mitigated. Evaluation is part of the control cycle prescribed in the risk management framework for AI developers (see, inter alia, Article 9 of the AI Act). This is also part of the post-commencement monitoring system for the AI system. To that end, the developer of the AI system shall operate on the basis of a monitoring plan. It is crucial to record the evaluation in such a way that necessary corrective or preventive measures are laid down. These measures are then part of the further development of the AI system.

Aligning AI systems with fundamental rights and public values requires actions throughout the lifecycle of an AI system

Schematic representation of building blocks that contribute to control of a simple AI system developed and deployed within one organization.



* Specialist and/or external support supports legitimacy

Explanation of this report

This report is about systems and applications of algorithms and artificial intelligence (AI) that can have an impact on people and society.

This is the third edition of the ARR, which is published biannually. The content is based on the knowledge obtained through the AP's monitoring network. Such as desk analysis and interviews with more than one hundred relevant national and international organisations. However, developments are moving fast and the view is still incomplete on many fronts. With this in mind, the AP nevertheless tries to form the best possible picture of current risks and developments in control measuresand to link policy recommendations to this in a constructive way. Nevertheless, errors or omissions in this ARR are possible.

AI systems automate, at their core, actions and decisions that people previously made. Or that were not possible in this way before. Simply put, we are talking about algorithms and AI. This ranges from relatively simple applications, in which a single algorithm functions on the basis of static decision rules, to very complex applications of machine learning or neural networks. The risk analysis in this report makes no distinction based on the technical functioning of algorithms and AI. This is in line with the policy consensus on the meaning of the term 'AI system' (see Box 'AI system as a broad definition').

The AI & Algorithmic Risk Report Netherlands (ARR) describes trends and developments in risks. These are risks in the use of algorithms and AI that can affect individuals, groups of persons or society as a whole. In the end, it can also disrupt society. The AP prepares the ARR to make stakeholders – private and public organisations, politicians, policy makers and the public – aware of these risks in a timely manner so that they can take action. There are two caveats in the description of trends and developments in risks. First, the use of algorithms and AI not only entails risks, but can also make positive contributions, also to strengthen fundamental values and fundamental rights. The supervision focuses on the elimination of risks and elimination of said risks. Secondly, the focus in this periodic report is on trends and developments. This means that emphasis is placed on the analysis, in addition to structural risks.

The ARR does not contain any predictions. With the current knowledge and available information, the AP wants to provide a compact and understandable picture of the current risks of the use of algorithms and AI and the challenges in managing these risks. Where possible, the AP makes proposals for policies that can counteract risks. This should not be seen as concrete guidance. The analyses and recommendations in the ARR provide organisations and policy makers with insights to reduce the risk of undesirable effects when using algorithms. The ARR can also be used to better understand algorithms and AI and to strengthen dialogue on opportunities and risks of algorithms in society.

The ARR remains a work in progress and can contain errors. The Netherlands is a global leader in working on careful control of algorithms and AI, so that its deployment is at the service of people and society. The design of the coordinating AI and algorithm oversight at the AP and the periodic system analyses in this ARR are examples of this. This new task started last year and is under construction. The first edition of the ARR (summer 2023) focused on the work of the DCA.

Get in touch with us. Your comments on the ARR and suggestions are welcome. You can send an email to dca@autoriteitpersoonsgegevens.nl

1 OECD. (March 2024). Explanatory Memorandum on the Updated OECD Definition of an AI system. OECD Artificial Intelligence Papers, No. 8. https://www.oecd-ilibrary.org/docserver/623da898-en.pdf

2 Financial Times. (6 June 2024). 'Most exciting moment' since birth of WiFi: chipmakers hail arrival of AI PCs. https://www.ft.com/content/6a546ad6-ae03-4c2d-92f5-c8efdd4bba3b

3 BCG. (21 September 2023). How People Can Create—and Destroy—Value with Generative AI [news item]. https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai

4 Google. (30 May 2024). AI Overviews: About last week. [news item]. https://blog.google/products/search/ai-overviews-update-may-2024/

5 Microsoft. (7 June 2024). Update on the Recall preview feature for Copilot+ PCs [news item]. https://blogs.windows.com/windowsexperience/2024/06/07/update-on-the-recall-preview-feature-for-copilot-pcs/

6 Open AI. (19 May 2024). How the voices for ChatGPT were chosen [news item]. https://openai.com/index/how-the-voices-for-chatgpt-were-chosen/

7 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2 August 2023). Attention is all you need. https://arxiv.org/abs/1706.03762

8 Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton. J., Kelton, F., Miller, J., Simens, M., Askell, A., Welinder P., Christiano, P., Leike, J., Lowe, R. OpenAI. (4 March 2022). Training language models to follow instructions with human feedback https://arxiv.org/pdf/2203.02155

9 AI Safety Institute (United Kingdom). (April 2024). Fourth Progress Report. https://www.aisi.gov.uk/work/fourth-progress-report.

10 AI Seoul Summit. (21 May 2024). Seoul Declaration for safe, innovative and inclusive AI by participants attending the leaders' session of the AI Seoul Summit. https://www.president.go.kr/download/664ca1113f0e7.

11 AI Safety Institute (United Kingdom). (20 May 2024). Advanced AI evaluations at AISI: May update. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

12 National AI Advisory Committee. (May 2024). Finding & Recommendations: AI Safety. https://ai.gov/wp-content/uploads/2024/06/FINDINGS-RECOMMENDATIONS_AI-Safety.pdf.

13 U.S. AI Safety Institute. (21 May 2024). The US AI Safety Institute: Vision, Mission, and Strategic Goals. https://www.nist.gov/system/files/documents/2024/05/21/AISI-vision-21May2024.pdf.

14 Dutch Data Protection Authority and National Inspectorate for Digital Infrastructure. (11 June 2024). Second (interim) advice on the supervisory structure of the AI Act. https://www.autoriteitpersoonsgegevens.nl/system/files?file=202406/20240516%20AI%20Act%20tweede%20tussenadvies.pdf.

15 OECD. (2023). Using AI to support people with disability in the labour market. https://read.oecd.org/10.1787/008b32b7-en.

16 UWV. (5 December 2023). Research shows: inclusive technology works. https://www.uwv.nl/nl/kennis-en-cijfers/uwv-als-kennisorganisatie/onderzoek-toont-aan-inclusieve-technologie-werkt.

17 EU Funding & Tenders Portal. (2023). Innovative and Inclusive Democratic Spaces for Deliberation and Participation (iDEM). https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/org-details/999999999/project/101132431/program/43108390/details.

18 MinBZK. (January 2024). Government-wide vision Generative AI. p. 41 Overheidsbrede visie Generatieve AI

19 TNO. (4 April 2024). Generative AI in Dutch healthcare. p. 7-8. TNO2024 R10662 Generatieve AI in de Nederlandse zorg getekend | Publicatie | Gegevensuitwisseling in de zorg

20 DNB, AFM. (2024). The impact of AI on the financial sector and supervision. p. 23-25 (*IA PDF AI Rapport (dnb.nl)

21 DNB, AFM. (2024). The impact of AI on the financial sector and supervision. p. 26 *IA PDF AI Rapport (dnb.nl)

22 Axis. (2024). Enhancing Safety and Security through Remote Surveillance. Enhancing Safety and Security through Remote Surveillance. Enhancing Safety and Security through Remote Surveillance | Axis Communications

23 Mobiliteit.nl. (7 June 2024). NS tests smart camera that recognises and captures potential violence. https://www.mobiliteit.nl/ov/2024/06/07/ns-test-in-amsterdam-met-slimme-camera-om-potentieel-geweld-te-vast-te-leggen/?gdpr=deny&gdpr=deny

24 Adams, R., Adeleke, F., Florido, A., de Magalhães Santos, L. G., Grossman, N., Junck, L., & Stone, K. (2024). Global Index on Responsible AI 2024 (1st Edition). South Africa: Global Center on AI Governance. The Global Index on Responsible AI (global-index.ai)

25 Court of Audit. (15 May 2024). Accountability study results 2023 Ministry of Infrastructure and Water Management Resultaten verantwoordingsonderzoek 2023 Ministerie van Infrastructuur en Waterstaat | Rapport | Algemene Rekenkamer

26 Second Chamber of the States-General. (26 February 2024). Report of the Parliamentary Committee of Inquiry

into the Fight against Fraud and Services: 'State powers were blind to people and law' https://www.tweedekamer.nl/nieuws/persberichten/rapport-parlementaire-enquete-commissie-fraudebestrijding-en-dienstverlening

[27] The Alliance. The housing fraud algorithm. https://www.de-alliantie.nl/over-de-alliantie/wat-we-doen/innovatie/innovaties/toekomstgerichte-organisatie/woonfraude/

[28] Berg, J. van den, Zwaan, I. de. (19 March 2024). Unrest in primary schools about the results of the new flow test: 'This is simply not correct'. De Volkskrant. https://www.volkskrant.nl/binnenland/onrust-op-basis-scholen-over-uitkomsten-nieuwe-doorstroom-toets-dit-klopt-gewoon-niet~b2bebdb4/.

[29] Ministry of Education, Culture and Science (3 June 2014). Decree-Law PO. Official Gazette 2014, 209. https://zoek.officielebekendmakingen.nl/stb-2014-209.html.

[30] The PO Council. (9 April 2024). PO-Raad delves deeper into the figures relating to flow-through tests and organises masterclass tests. https://www.poraad.nl/po-raad-duikt-dieper-in-de-cijfers-rondom-doorstroom-toets-en-organiseert-masterclass-toetsing.

[31] Minister for Primary and Secondary Education. (17 April 2024). Primary education; Government letter; First pass-through test. https://zoek.officielebekendmakingen.nl/kst-31293-729.html.

[32] Algorithm register of the Municipality of Amsterdam (2024). Research worthiness: Smart maintenance check https://algoritmeregister.amsterdam.nl/ai-system/onderzoekswaardigheid-slimme-check-levensonderhoud/1086/

[33] The Parool. (14 February 2024). Opinion: 'Research bias of both algorithm and civil servant' https://www.parool.nl/columns-opinie/opinie-onderzoek-vooringenomenheid-van-zowel-algoritme-als-ambtenaar~bd69aa5e/

[34] For the full documentation made available by the municipality of Amsterdam, see the 'Overview of Processed Data and Features' under 'Non-discrimination' in the description of the algorithm 'Research worthiness: Smart maintenance check' (https://algoritmeregister.amsterdam.nl).

[35] Dutch Data Protection Authority. (18 December 2023). Reporting AI- & algorithm risks Netherlands (RAN) - winter 2023 (https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-najaar-2023

[36] Government of the Netherlands. (8 October 2019). Strategic Action Plan on Artificial Intelligence [policy paper]. Strategisch Actieplan voor Artificiële Intelligentie | Beleidsnota | Rijksoverheid.nl

[37] TNO Vector. (4 June 2024). The economic value of strategic autonomy. De economische waarde van strategische autonomie - TNO Vector

[38] ACM FAccT conference. (2024). List accepted papers. https://facctconference.org/2024/acceptedpapers

[39] Dutch Data Protection Authority and National Inspectorate for Digital Infrastructure. (11 June 2024). Second (interim) advice on the supervisory structure of the AI Act. https://www.autoriteitpersoonsgegevens.nl/system/files?file=2024-06/20240516%20AI%20Act%20tweede%20tussenadvies.pdf

[40] Dutch Data Protection Authority. (18 December 2023). Reporting AI- & algorithm risks Netherlands (RAN) - winter 2023 (https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-najaar-2023

[41] Commissariat for the Media. (June 2024). Digital News Report Netherlands 2024. 2031086-CvdM-DigitalNewsReport-2024_def.pdf

[42] Wired. (23 January 2024). The Biden Deepfake Robocall Is Only the Beginning. https://www.wired.com/story/biden-robocall-deepfake-danger/

[43] The New York Times. (June 2024). A Small Army Combating a Flood of Deepfakes in India's Election. [news article]. https://www.nytimes.com/2024/06/01/world/asia/india-election-deepfakes.html

[44] NRC. (30 May 2024). 'Alla yes on Rafah' goes viral, but there is also criticism: 'This is an AI photo, so your eyes don't have to look at Rafah'. https://www.nrc.nl/nieuws/2024/05/30/is-het-protestbeeld-all-eyes-on-gaza-leunstoelactivisme-zo-kweek-je-bewustwording-bij-jongeren-a4200501?t=1717404010

[45] Waag Futurelab, Dutch AI Coalition. (30 April 2024). A social Research Agenda for AI. Eindrapport-Een-maatschappelijke-onderzoeksagenda-voor-AI.pdf (waag.org)

[46] Commissariat for the Media. (June 2024). Digital News Report Netherlands 2024. 2031086-CvdM-DigitalNewsReport-2024_def.pdf

[47] Dutch Data Protection Authority. (18 December 2023). Reporting AI- & algorithm risks Netherlands (RAN) - winter 2023 (https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-najaar-2023

[48] Government of the Netherlands. (23 December 2022). Letter to the House of Representatives on an effective approach to disinformation. [chamber]. https://open.overheid.nl/repository/ronl-d3369562e78345a02126dc-d644ae9e6edc1a5b12/1/pdf/kamerbrief-over-rijksbrede-strategie-effectieve-aanpak-van-desinformatie.pdf

[49] NCTV. (25 June 2024). Main report Trend analysis National Security 2024. https://www.nctv.nl/binaries/nctv/documenten/rapporten/2024/06/25/hoofdrapport-trendanalyse-nationale-veiligheid-2024/Hoofdrapport+Trendanalyse+Nationale+Veiligheid+2024.pdf

[50] Rathenau. (13 October 2022). Digital threats to democracy. p. 21. https://www.rathenau.nl/sites/default/files/2020-10/RAPPORT_Digitale_dreigingen_voor_de_democratie_Rathenau_Instituut.pdf

[51] NOS (3 May 2024). Chatbots advised: spread disinformation and sow fear about EU elections. [News article]. Chatbots adviseerden: verspreid desinformatie en zaai angst over EU-verkiezingen (nos.nl)

[52] NRC. (31 May 2024). How Google is changing its search engine, and with it the entire internet. [news article]. Hoe Google zijn zoekmachine op de schop gooit, en daarmee het hele internet - NRC

[53] De Volkskrant. (24 May 2024). New AI search service Google: Put glue on your pizza and put chlorine gas in your washing machine. [news article]. Nieuwe AI-zoekdienst Google: doe lijm op je pizza en stop chloorgas in je wasmachine | de Volkskrant

[54] De Groene Amsterdammer. (5 June 2024). Accountability when researching AI books on Bol. Verantwoording bij het onderzoek naar AI-boeken op Bol – De Groene Amsterdammer

[55] The Verge. (6 February 2024). Meta says you better disclose your AI fakes or it might just pull them. [news article]. Meta to label AI-generated images on Facebook, Instagram, Threads - The Verge

[56] Content Autheniticity Initiative. (2024). Restoring trust and transparency in the age of AI. Content Authenticity Initiative

[57] BBC. (4 March 2024). Haiti violence: Haiti gangs demand PM resign after mass jailbreak. [news article]. Haiti violence: Haiti gangs demand PM resign after mass jailbreak (bbc.com)

[58] Arnold, M., Goldschmitt, M., Rigotti, T.(21 June 2023). Dealing with information overload: a comprehensive review. Dealing with information overload: a comprehensive review - PMC (nih.gov)

[59] Forbes advisor. (4 June 2024). Top website statistics for 2024. Top Website Statistics For 2024 – Forbes Advisor

[60] The Brussels Times. (15 February 2024). TikTok overtakes Google as most popular search engine among Gen Z [news article]. TikTok overtakes Google as most popular search engine among Gen Z (brusselstimes.com)

[61] The European Parliament. (8 November 2023). REPORT on addictive design of online services and consumer protection in the EU single market. REPORT on addictive design of online services and consumer protection in the EU single market | A9-0340/2023 | European Parliament (europa.eu)

[62] Charter of Fundamental Rights of the European Union. Artikel 11 - De vrijheid van meningsuiting en van informatie | European Union Agency for Fundamental Rights (europa.eu)

[63] Funk, A., Shahbaz, A., Vesteinsson, K. (2023) Freedom House. Freedom on the net 2023. The repressive power of artificial intelligence. p.16 https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[64] The Economic Times of India. (22 February 2024). X disagrees with govt blocking orders for certain posts and accounts. [news article]. https://economictimes.indiatimes.com/tech/technology/x-complies-with-govt-blocking-orders-for-certain-posts-accounts/article-show/107904325.cms

[65] DW (17 April 2024). X blocks posts in India after election commission order. [news article]. https://www.dw.com/en/x-blocks-posts-in-india-after-election-commission-order/a-68846469

[66] Funk, A., Shahbaz, A., Vesteinsson, K. (2023) Freedom House. Freedom on the net 2023. The repressive power of artificial intelligence. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[67] ACM. Trustworthy flaggers. Betrouwbare flaggers | ACM.nl

[68] The Media Authority. (June 2024). Between Bits and Principles: How AI challenges the core values of media policy. AI-Verkenning-CvdM-Tussen-Bits-en-Principes.pdf

[69] AIVD. (2023) 2022 annual report. https://www.aivd.nl/onderwerpen/jaarverslagen/jaarverslag-2022

[70] Defend democracy. (19 April 2024). Automated anarchy: the rising tide of bad bots on the internet. https://defenddemocracy.eu/automated-anarchy-the-rising-tide-of-bad-bots-on-the-internet/

[71] The Guardian. (19 July 2023). Disinformation reimagined: How AI could erode democracy in the 2024 U.S. elections. [news article]. Disinformation reimagined: how AI could erode democracy in the 2024 US elections | US elections 2024 | The Guardian

[72] The Rathenau Institute. (13 October 2020). Digital threats to democracy, p. 55 Digitale dreigingen voor de democratie | Rathenau Instituut

[73] Funk, A., Shahbaz, A., Vesteinsson, K. (2023) Freedom House. Freedom on the net 2023. The repressive power of artificial intelligence. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[74] UCL, University of Kent. (2 February 2024). Safter scrolling: How algorithms popularise and gamify online hate and misogyny for young people. Safer-scrolling.pdf (ascl.org.uk)

[75] Funk, A., Shahbaz, A., Vesteinsson, K. (2023) Freedom House. Freedom on the net 2023. The repressive power of artificial intelligence. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[76] NRC. (19 March 2024). Also known Dutch victim deepfake

porno: violation of sexual privacy. [news article]. Ook bekende Nederlanders slachtoffer deepfake porno: schending seksuele privacy - NRC

77 The Media Authority. (June 2024). Digital News Report Netherlands 2024 2031086-CvdM-DigitalNewsReport-2024_def.pdf

78 The Media Authority. (June 2024). Digital News Report Netherlands 2024. Digital News Report Nederland 2024 2031086-CvdM-DigitalNewsReport-2024_def.pdf

79 Wagner, M. (2024). Affective polarization in Europe. European Political Science Review, 1–15. doi:10.1017/S1755773923000383 Affective polarization in Europe | European Political Science Review | Cambridge Core

80 Mediawijheid.nl. What is media literacy Wat is mediawijsheid? - Mediawijsheid.nl

81 MinBZK. (7 June 2024). State-wide strategy for the effective tackling disinformation and announcing new actions. Brief - Voortgangsbrief Rijksbrede strategie voor de effectieve aanpak van desinformatie en aankondiging nieuwe acties (overheid.nl)

82 Government of the Netherlands. Digital literacy in school. Digitale geletterdheid op school | Digitalisering in het onderwijs | Rijksoverheid.nl

83 SLO. (6 March 2024). Concept Core Objectives Learning Area Digital Literacy + Explanatory Document. Conceptkerndoelen leergebied digitale geletterdheid + toelichtingsdocument - SLO

84 Official Journal of the European Union. (26 April 2024). Mededeling van de Commissie — Richtsnoeren van de Commissie voor aanbieders van zeer grote online-platforms en zeer grote onlinezoekmachines inzake de beperking van systeemrisico's voor verkiezingsprocessen overeenkomstig artikel 35, lid 3, van Verordening (EU) 2022/2065 (europa.eu)

85 European Commission. (24 April 2024). Commission stress tests platforms' election readiness under the Digital Services Act. [news article]. Commission stress tests platforms' election readiness under the Digital Services Act | Shaping Europe's digital future (europa.eu)

86 European Commission. (30 April 2024). Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act. [Press Release]. Commission opens formal proceedings under DSA (europa.eu)

87 TNO. (June 2024). Quickscan AI in Public Services (TNO 2024 R11005). https://publications.tno.nl/publication/34642601/SASNc3ZW/TNO-2024-R11005.pdf.

88 See TNO. (June 2024).

89 Algorithm register. 2024. Automated document check and face comparator (Municipality of's Hertogenbosch). https://algoritmes.overheid.nl/nl/algoritme/38325188.

90 Algorithm register. (2024). Chatbot Guus (AI Version) (Municipal Goes). https://algoritmes.overheid.nl/nl/algoritme/15943226.

91 Algorithm register. (2024). Early warning (Municipality of Amsterdam). https://algoritmes.overheid.nl/nl/algoritme/66453169.

92 For example, RTL News calculated in January 2020 that at that time 25% of the Dutch people already lived in a municipality that uses scan cars. RTL news. 16 January 2020. Scanning cars are generating millions (and growing) in municipalities. Scanauto's leveren gemeentes miljoenen op (en het worden er steeds meer) | RTL Nieuws | RTL.nl

93 Dutch Data Protection Authority. (2024). 2023 annual report. https://www.autoriteitpersoonsgegevens.nl/documenten/ap-jaarverslag-2023.

94 Dutch Association for Councillors. Tasks of the City Council. Taken gemeenteraad | Nederlandse Vereniging voor Raadsleden

95 Gemeentewet, art. 182(1) wetten.nl - Regeling - Gemeentewet - BWBR0005416 (overheid.nl)

96 General Administrative Law Act 9(22-27).

97 Council for Public Administration and Council for Culture. 2020. Local media: Not to be missed. p. 22. Adviesrapport Lokale media: niet te missen | Publicatie | Raad voor het Openbaar Bestuur (raadopenbaarbestuur.nl)

98 See TNO. (June 2024).

99 Hooghiemstra & Partners. (June 2021). How municipalities decide on algorithms & human rights. p. 7 and 16. https://publicaties.mensenrechten.nl/publicatie/60dd2c7b98d7821c6468363e.

100 Hooghiemstra & Partners. (June 2021). How municipalities decide on algorithms & human rights. https://publicaties.mensenrechten.nl/publicatie/60dd2c7b98d7821c6468363e.

101 See Hooghiemstra & Partners. (June 2021). p. 19.

102 The Rathenau Institute. (September 2020). Council know with digitalisation. p. 4. https://www.rathenau.nl/nl/kennis-voor-transities/raad-weten-met-digitalisering.

103 Council for Public Administration. (2021). Send or be sent? On the legitimacy of sending with data. https://www.raadopenbaarbestuur.nl/documenten/publicaties/2021/05/25/advies-sturen-of-gestuurd-worden.

104 Rathenau. (September 2020). Getting to know the Council through digitalisation. p. 20

105 Council for Public Administration. (November 2020). Good support for strong democracy: on the support of devolved parliament. p. 20. Adviesrapport Goede ondersteuning, sterke democratie | Publicatie | Raad voor het Openbaar Bestuur (raadopenbaarbestuur.nl).

106 Court of Audit of the Amsterdam Metropolitan Area. (October 2023). Algorithms: how Amsterdam can better apply algorithms. Onderzoeksrapport-Algoritmen-DEF.pdf

(amsterdam.nl)

[107] Rotterdam Court of Auditors. (2024). Color Confess: follow-up research into algorithms. kleur bekennen (rotterdam.nl).

[108] Court of Audit The Hague. (March 2024). 2023 Annual Report of the Court of Auditors of The Hague. RIS318254_Jaarverslag_2023_Rekenkamer_Den_Haag.pdf (rekenkamerdenhaag.nl)

[109] Fund for Journalism. (June 2022). From Steam to Power: The road to professionalism. p. 29. Van Stoom naar Stroom: de weg naar professionalisering - SVDJ

[110] Hooghiemstra & Partners. June 2021. How municipalities decide on algorithms & human rights. p. 20. https://publicaties.mensenrechten.nl/publicatie/60dd2c7b98d7821c6468363e.

[111] Hooghiemstra & Partners. (June 2021). How municipalities decide on algorithms & human rights. p. 16. https://publicaties.mensenrechten.nl/publicatie/60dd2c7b98d7821c6468363e.

[112] Council for Public Administration. (November 2020). Good support for strong democracy: on the support of devolved parliament. p. 35. Adviesrapport Goede ondersteuning, sterke democratie | Publicatie | Raad voor het Openbaar Bestuur (raadopenbaarbestuur.nl).

[113] BDO (January 2024). Benchmark Dutch municipalities 2024. P. 22-24. Download nu de BDO-Benchmark Nederlandse gemeenten 2024

[114] Scientific Council for Government Policy. (2021). The task of AI. The new system technology. p. 437-442 Opgave AI. De nieuwe systeemtechnologie | Rapport | WRR

[115] MinBZK. (17 June 2024). Collective letter Digitisation June 2024. Brief – Verzamelbrief Digitalisering juni 2024 (overheid.nl)

[116] Author, G., Sarmah, D. K., & El-Hajj, M. (2024). Automobile Insurance Fraud Detection Using Data Mining: A systematic literature review. Intelligent Systems With Applications, 200340. https://doi.org/10.1016/j.iswa.2024.200340

[117] DNB AFM. (2024). The impact of AI on the financial sector and supervision. https://www.dnb.nl/nieuws-voor-de-sector/toezicht-2024/afm-en-dnb-publiceren-rapport-over-de-impact-van-ai-in-de-financiele-sector-en-het-toezicht-daarop/

[118] Meta. (3 August 2020). How Does Facebook Measure Fake Accounts? https://about.fb.com/news/2019/05/fake-accounts/

[119] Supplements Department, Ministry of Finance. (18 March 2024). Use of dual nationality. https://herstel.toeslagen.nl/gebruik-van-dubbele-nationaliteit/

[120] House of Representatives of the States-General. (26 February 2024). Report of the Parliamentary Committee of Inquiry into the Fight against Fraud and Services: 'State powers were blind to people and law'. https://www.tweedekamer.nl/nieuws/persberichten/rapport-parlementaire-enquetecommissie-fraudebestrijding-en-dienstverlening

[121] Court of The Hague. (5 February 2020). 6.87-6.94, ECLI:NL:RBDHA:2020:865. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:865

[122] DUO. (2024). Excuses for indirect discrimination in controls at the live-out grant. https://duo.nl/organisatie/pers/excuses-voor-indirecte-discriminatie-bij-controles-op-de-uitwonendenbeurs.jsp

[123] Rensen, Frank. 2 July 2024. 'With just about every tile we light, we discover discriminatory algorithms in the government', says the Dutch Data Protection Authority. De Volkskrant. https://www.volkskrant.nl/tech/bij-zo-n-beetje-elke-tegel-die-we-lichten-ontdekken-we-discriminerende-algoritmen-bij-de-overheid-zegt-de-autoriteit-persoonsgegevens~b360ed6e

[124] This differs from a technical interpretation of discrimination in which it is regularly used as a synonym for discrimination.

[125] College of Human Rights. (2022). Question 4, Questions and Answers on recruitment and selection algorithms for employers. https://www.mensenrechten.nl/themas/digitalisering/werving-en-selectie/qa-over-hr-algoritmes-voor-werkgevers

[126] College of Human Rights. (2021). Discrimination through risk profiles - A human rights assessment framework. https://publicaties.mensenrechten.nl/publicatie/61a734e65d726f72c45f9dce

[127] In a case at the VU, a plausible suspicion of discrimination has been demonstrated in AI technology that did not work well for a student with a dark skin color. According to the final verdict, there was no discrimination here. See: College of Human Rights. (17 October 2023). Student not discriminated against by exam software Proctorio, but VU should have handled the complaint more carefully. https://www.mensenrechten.nl/actueel/nieuws/2023/10/17/student-niet-gediscrimineerd-door-tentamensoftware-proctorio-maar-vu-had-de-klacht-zorgvuldiger-moeten-behandelen

[128] FRA. (2021). Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf

[129] European Commission. (22 May 2023). C(2023)3215 – Standardisation request M/593 COMMISSION IMPLEMENTING DECISION of 22.5.2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Stand-

ardisation in support of Union policy on artificial intelligence https://ec.europa.eu/growth/tools-databases/enorm/mandate/593_en

[130] European Commission. (September 2023). EU model contractual AI clauses to pilot in procurements of AI. https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/eu-model-contractual-ai-clauses-pilot-procurements-ai.

[131] Dutch Data Protection Authority and National Inspectorate for Digital Infrastructure. (11 June 2024). Second (interim) advice on the supervisory structure of the AI Act. https://www.autoriteitpersoonsgegevens.nl/system/files?file=202406/20240516%20AI%20Act%20tweede%20tussenadvies.pdf.

[132] Dutch Data Protection Authority and National Inspectorate for Digital Infrastructure. (11 June 2024). Second (interim) advice on the supervisory structure of the AI Act. https://www.autoriteitpersoonsgegevens.nl/system/files?file=202406/20240516%20AI%20Act%20tweede%20tussenadvies.pdf.

[133] Klein, E. & Patrick, S. (21 March 2024). Envisioning a Global Regime Complex to Govern Artificial Intelligence. https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en

[134] Klein, E. & Patrick, S. (21 March 2024). Envisioning a Global Regime Complex to Govern Artificial Intelligence. https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en

[135] Roberts, H., Cowls, J., Hine, E., Morley, J., Wang, V., Taddeo, M., & Floridi, L. (2022). Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. The Information Society, 39(2), 79–97. https://doi.org/10.1080/01972243.2022.2124565

[136] Ryan-Mosley, T., Heikkilä, M., Yang, Z. (5 January 2024). What's next for AI regulation in 2024? https://www.technologyreview.com/2024/01/05/1086203/whats-next-ai-regulation-2024/

[137] Brazilian proposal for AI rules. (2024). https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

[138] Schertel Mendes, L. & Kira, B. (21 December 2023). The Road to Regulation of Artificial Intelligence: the Brazilian experience. https://policyreview.info/articles/news/road-regulation-artificial-intelligence-brazilian-experience/1737

[139] Klein, E. & Patrick, S. (21 March 2024). Envisioning a Global Regime Complex to Govern Artificial Intelligence. https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en

[140] OECD. Update to the OECD I Principles. (2024). https://oecd.ai/en/ai-principles

[141] United Nationas News. (21 March 2024). General Assembly adopts landmark resolution on artificial intelligence. https://news.un.org/en/story/2024/03/1147831

[142] United Nations. (21 March 2024). General Assembly Adopts Landmark Resolution on Steering Artificial Intelligence towards Global Good, Faster Realization of Sustainable Development General Assembly Adopts Landmark Resolution on Steering Artificial Intelligence towards Global Good, Faster Realization of Sustainable Development | Meetings Coverage and Press Releases (un.org)

[143] For example, the OECD Privacy Guidelines, which set limits on the collection and use of personal data, underlie many privacy laws. Forty-two countries adopt new OECD Principles on Artificial Intelligence - OECD

[144] United Nations Ai Advisory Board. (December 2023). Interim Report: Governing AI for humanity. https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

[145] Dutch Data Protection Authority. (April 2024). Supervisory Perspective on Global AI Governance (Discussion Paper). https://www.autoriteitpersoonsgegevens.nl/en/documents/supervisory-perspective-on-global-ai-governance-discussion-paper

[146] Dutch Data Protection Authority. (April 2024). Supervisory Perspective on Global AI Governance (Discussion Paper). https://www.autoriteitpersoonsgegevens.nl/en/documents/supervisory-perspective-on-global-ai-governance-discussion-paper

[147] Dutch Data Protection Authority. (7 December 2023). Blog post: care for generative AI. https://www.autoriteitpersoonsgegevens.nl/actueel/blogpost-zorgen-om-generatieve-ai

[148] World Privacy Forum. (December 2023). New Report: Risky Analysis: Assessing and Improving AI Governance Tools. New Report: Risky Analysis: Assessing and Improving AI Governance Tools | World Privacy Forum

AUTORITEIT
PERSOONSGEGEVENS