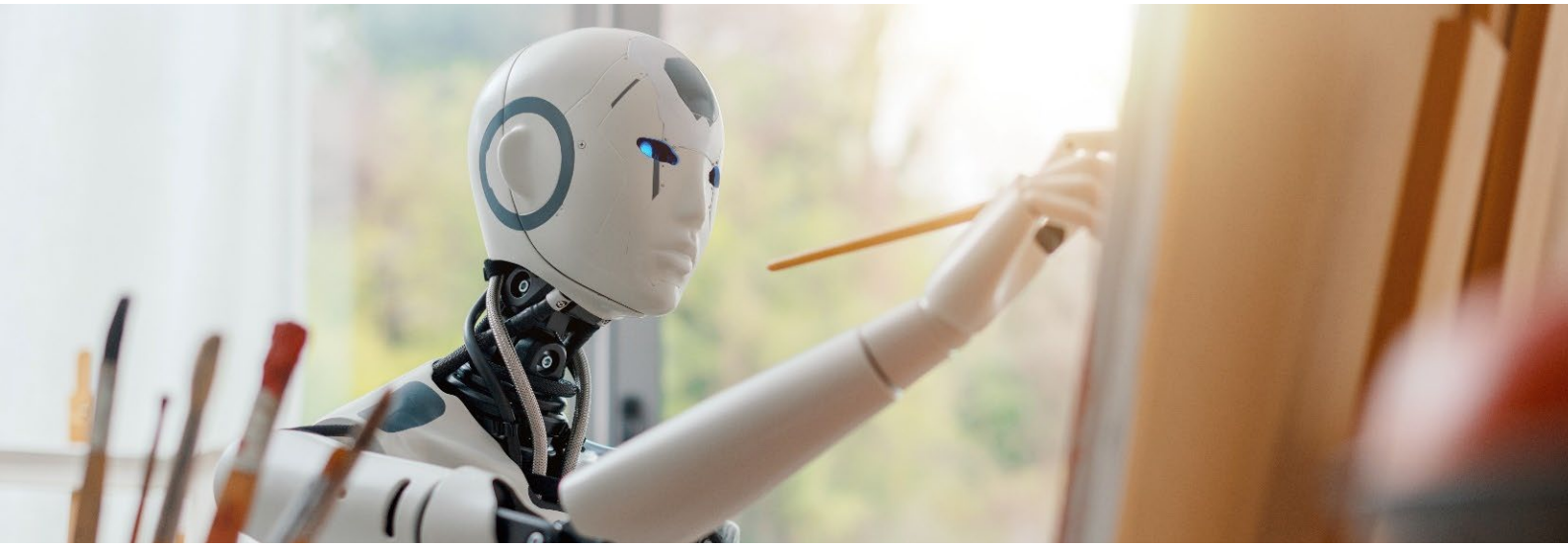
Rathenau Instituut

Generative AI



Rathenau Scan

Introduction
Few technologies provoke as much discussion as generative artificial intelligence (GAI). Systems such as ChatGPT and Bard have taken us a step towards computers that can perform countless tasks, but how far artificial intelligence (AI) extends is unclear. Some experts think that computers are becoming so powerful that they threaten the survival of humanity. Others think this is exaggerated, or point to important short-term risks such as prejudices and incorrect output.

This scan takes stock of the situation: what is GAI, what is it currently capable of, and what may it be capable of in the future? What opportunities, risks to public values, and policy options are associated with it? The scan – which is intended for policy-makers and politicians – was carried out at the request of the Dutch Ministry of the Interior and Kingdom Relations, based on a short-term study involving a review of the relevant literature, workshops, and interviews.

Contents

Summary

Why publish a Rathenau scan of generative AI?
The term “generative AI” (GAI) refers to AI systems that can create content automatically, at the request of a user. You can ask such a system to produce a summary, for example, or create a picture in the style of Van Gogh. Since the launch of ChatGPT in November 2022, millions of users worldwide have been experimenting with this technology, and it is already impacting society while expectations of what it will bring are high. The present scan provides an overview of the possibilities and risks associated with GAI, and potential policy actions.

Is generative AI something new?
Generative AI builds on existing AI technologies and is a subset of learning AI systems. At the same time, generative AI systems have a number of distinctive features:

- first, they are significantly better at language than other AI systems;
- second, they can work effectively with different “modalities”, such as image, sound, video and speech, and even such things as protein structures and chemical compounds;
- third, generative AI systems receive general training, which provides the basis for all kinds of specific applications.

Subsequently, GAI systems can perform many different tasks, unlike many other AI systems that fall into the category of “narrow AI”, and are trained for just one specific task.

What can you do with generative AI?
In the present scan, we distinguish four roles that GAI systems can fulfil. A GAI system can be deployed as:

1. a learning tool: for example to look up information or to act as a source of information when doing one’s homework;
2. a production tool: the system creates something at the behest of a user. Many people are already experimenting with this in the workplace;
3. a solver of complex problems: for example in science, with GAI systems helping fold protein structures, for example to support developing new types of medication;
4. to create an experience: some users find it enjoyable or fascinating to interact with GAI systems, which can take on the role of a companion. For example, someone has already created a chatbot that imitated a deceased loved one.

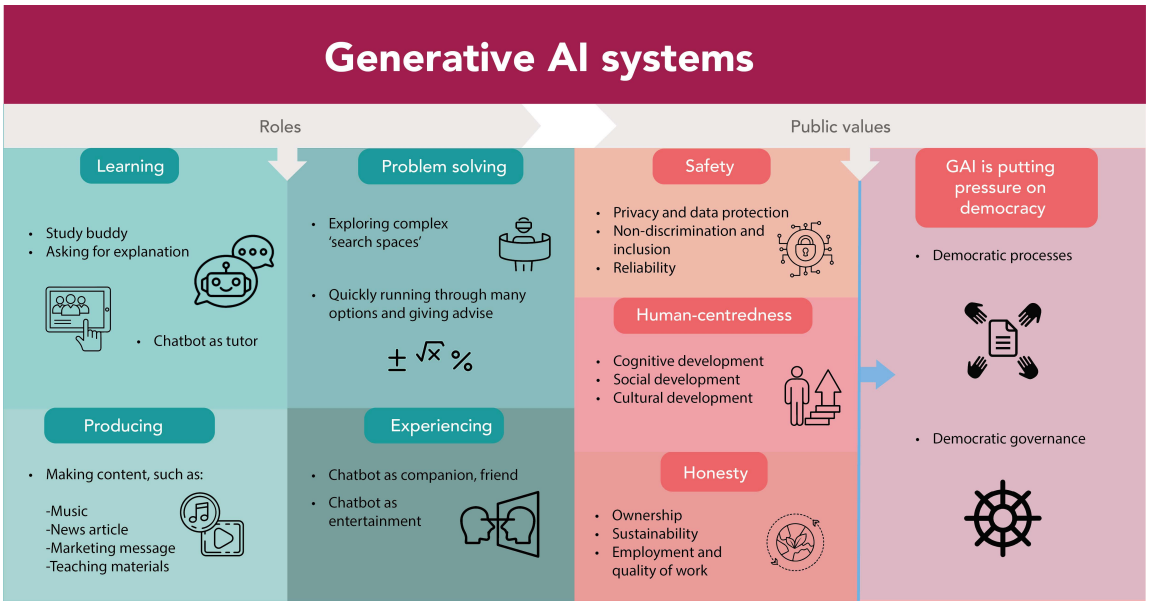
Despite these possibilities, the technology has its limitations. Generative AI systems are based on statistics, and thus calculate the most likely answer. This may lead to incorrect answers or discriminatory content. The underlying algorithms are also so complex that people can only understand how they function to a limited extent – and that includes those who have developed them. As a result, the technology is not yet good enough to be deployed in important processes, such as medical diagnostics.

What is at stake with the rise of GAI?
Generative AI involves numerous risks that can put public values under pressure. In this scan, we have grouped those risks according to three themes. First, there are concerns about the **safety** of GAI systems: they can violate users' privacy, express prejudices, and provide false information. Moreover, they are so complex that developers and external parties cannot fully understand how they work, making it difficult to prevent risks, whether now or in the future.

Second, there is the question of how **human-centred** the systems are: what will they mean for our cognitive, social, and cultural development? Will chatbots encourage creativity? Will we unlearn social skills if we frequently interact with a GAI system? Will we genuinely process grief through chatbots that imitate our deceased loved ones? In short: what does it mean to be human in a world of robots?

Third, there are concerns if the distribution of benefits and burdens is **equal and just**: who benefits from GAI systems? Who bears the costs, for example in term of who's job, and nature of that job, will be affected? How do we protect the work of the creative professions? Which jobs are going to change, and how do we ensure decent work? And how do we deal with the environmental impact?

Finally, we identify a central feature, namely the impact of GAI on our **democracy**. GAI can hamper democratic processes, such as public debate and political decision-making, and because of the increasing power of a few tech companies, it can affect the ability to exert democratic control over digital technology in many domains of society.



Source: Rathenau Instituut

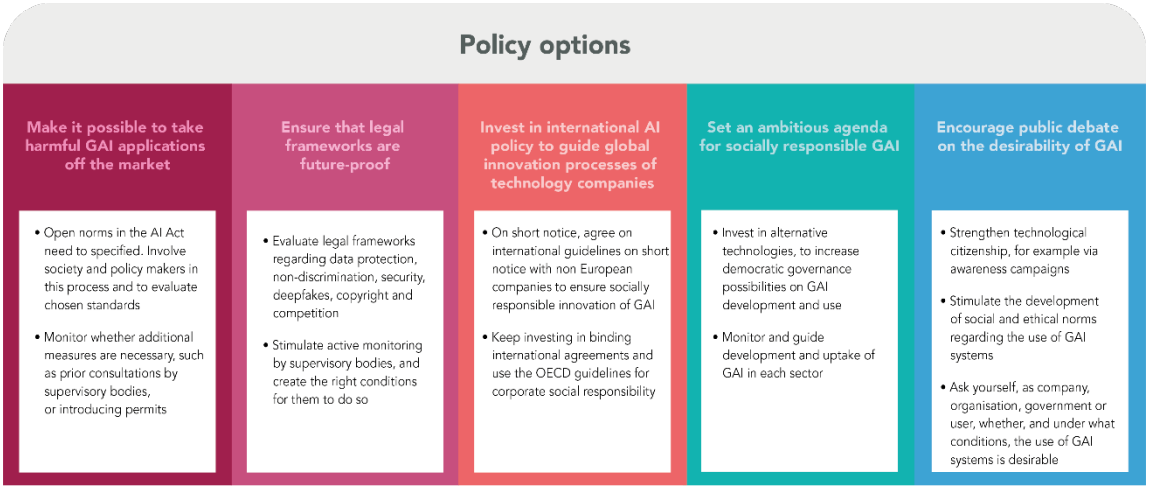
What needs to be done?

In the present scan, the Rathenau Instituut concludes that generative AI amplifies risks within digital society and also introduces new risks. In recent years, policy-makers at national, European, and international level have been working to steer AI in the right direction, with the EU's forthcoming AI Act as important policy instrument. It is unclear, however, how the open norms set out in that legislation regarding respect for human rights will be given shape in actual practice. When, for example, has the risk of discrimination been reduced to an acceptable level? And for whom is it then acceptable? It is also open to question whether other legal frameworks and policies adequately address the risks associated with generative AI.

The key question is therefore whether the policy efforts that are being made are actually sufficient. There is a real possibility that current and proposed policies may be unable to cope with the impact of generative AI systems, for example as regards non-discrimination, security, disinformation, competition, and exploitation of workers. It is therefore imperative that the Dutch government sets out a strategy for improving society's grip on this technology. Doing so should start by evaluating Dutch and EU policies and testing where they need to be strengthened. It is also important to provide maximum support for regulatory oversight, to make arrangements with developers, and to warn society about the risks of GAI. Globally, those risks are indeed being taken seriously; every individual and every institution in the Netherlands must do the same.

The Rathenau Instituut formulates five courses of policy action for the government:

1. Make it possible to take harmful GAI applications off the market;
2. Ensure that legal frameworks are future-proof;
3. Invest in international AI policy to guide global innovation processes of technology companies;
4. Set an ambitious agenda for socially responsible GAI;
5. Encourage public debate on the desirability of GAI.



Source: Rathenau Instituut

1. What is generative AI?

1.1. Introduction

The term “generative GAI” refers to artificial intelligence systems that can create content automatically, at the request of a user. That includes, for example, texts such as a job application letter, programming code or a school essay, images such as paintings or photographs, videos, and sounds such as the voices of various people. The user does not need to master software coding, and can interact with the generative system through human language (“prompts”). The development of generative AI systems has accelerated rapidly because a new algorithmic method has made it possible to process data far more efficiently, calculating with many more variables. This has led to complex algorithmic models that can perform impressive language tasks, also referred to as “large language models”.[1] We explain this technological breakthrough below (see Section 1.2). The present scan focuses on generative AI systems (referred to below as “GAI systems”) because they can perform a multitude of tasks and will therefore have a significant impact on society.

The most familiar GAI system is Chat-GPT, a chatbot designed by the US company OpenAI. Soon after its launch in November 2022, Chat-GPT3.5 was already being utilised by millions of users worldwide. There are also other generative systems, such as the image generator DALL-E (also from OpenAI), Microsoft's Co-Pilot, and Google's Bard chatbot.

1.2. How does it work?

As noted above, GAI is based on large language models (LLMs). These software models calculate which is the most likely next word in a sentence, for example: “The Pope believes in ...” [God]. The most likely missing word can also be calculated: “I'm eating a ..... with cheese” [sandwich]. New models can predict not only words, but also sound fragments in a music sequence, or images in a composite made up of pixels. These models can therefore handle multiple “modalities”, and can also combine text, image, and sound. They are therefore also referred to as large multimodal models (LMMs). ChatGPT4 can generate both text and images.

To detect and predict patterns in images, language and sound, the models need extensive training. This is done using existing machine learning technologies, such as neural networks.[2] A new development has emerged within machine learning: certain algorithmic models, “transformers”, allow far more text to be analysed. This helps to process the context of a word or phrase more effectively: after all, an individual

---

1 Language tasks are language-related tasks such as summarising, answering questions, translating, and the like.

2 Generative AI can be viewed as a subset within AI, and specifically a subset within machine learning and deep learning AI technologies. AI can be described as a system that exhibits intelligent behaviour by analysing their environment and taking action – with a certain degree of autonomy – to achieve specific goals; see European Commission High Level Expert Group on AI, 2019.

sentence only acquires meaning once you have also read the entire page.[3] These new models represent a crucial advance on previous AI systems.[4]

Development of a GAI system comprises roughly five phases (see Figure 1).

1. Data collection

First, large quantities of training data are collected, for example the inconceivable quantity of texts, photos and videos that are available on line, conversations on social media, and all the books and scientific literature that can be digitised. Much of this data is publicly available, but it may also comprise privately owned datasets that are purchased.[5]

2. Data curation

The data can then be filtered, for example by anonymising it, removing duplicates, or removing specific words. Here, algorithms and human work both play a role.

3. Training

In the third phase, training takes place, also referred to as “pre-training”.[6] The language model attempts to detect many different types of patterns in the vast database. In doing so, it uses billions of parameters, or variables that can have different values.[7] Both the datasets and the number of parameters are huge: GPT3 used 570 GB of data and 175 billion parameters.[8] The quantity may be even bigger: it is estimated that GPT4 was trained with 1.75 trillion (10^{12}) to a quintillion (10^{18}) parameters.

4. Fine-tuning

After being trained, the model is adjusted and refined further in the fourth phase so as to make it suitable for particular tasks (called ‘fine-tuning’). If it is to be deployed in the field of medicine, for example, a GAI system can be trained with medical terms and literature. The system can also undergo specific training to reduce the likelihood of it generating racist or offensive statements, as was done with GPT4. This is also referred to as “alignment”, and involves using humans to assess and align the output of the model concerning hateful, offensive, discriminatory, or otherwise illegal or unwanted content. A system can also be further refined with instructions (“prompt engineering”). For example, a programmer can use reasoning steps to show how a model can arrive

---

3 A transformer does even more: it involves a "new" attention mechanism that indicates which context information is important. Moreover, input can be processed in parallel instead of word by word – a major limitation of the neural networks previously used for language. See Vaswani et al., 2017

4 Brown et al., 2020; Kaplan et al., 2020

5 The current trend is for the amount of freely available internet data to decrease. This is partly because various platforms are shielding their data to a greater extent. Reddit and X (formerly Twitter), for example, have taken measures to prevent data from being “scraped” from their platforms. The relative value of internet data may also decrease as the internet begins to contain more AI-generated content. See *inter alia* Vipra & Myers West, 2023

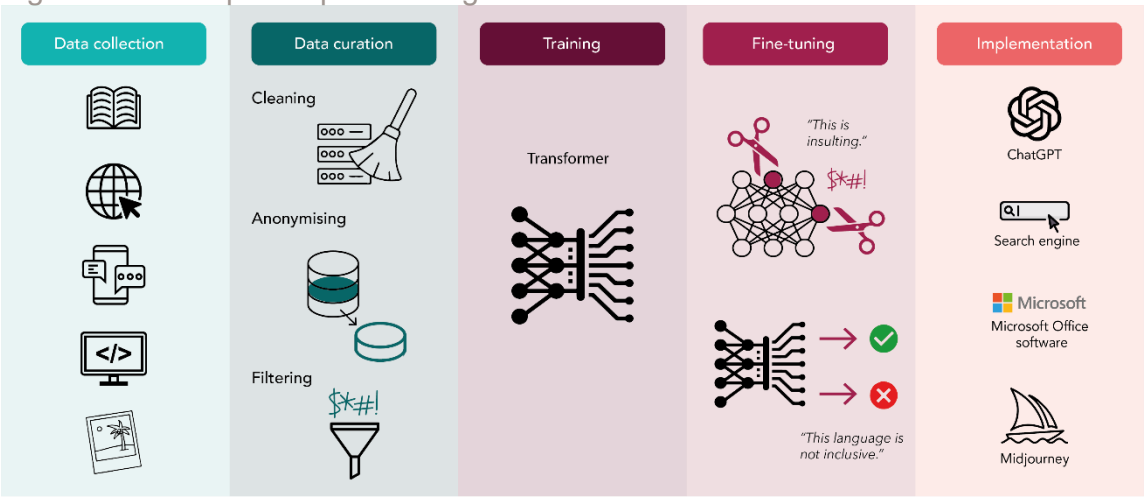
6 Now that these terms have been explained, the abbreviation “ChatGPT” becomes clear: Generative Pre-trained Transformer.

7 “Parameters” means the number of variables the model has been trained with, for example weightings in neural networks or coefficients in linear regression.

8 OpenAI has released several variants of GPT3, which are also referred to as the GPT3 “family”. We refer here to the largest model in the GPT3 family.

at the desired answer. A more specific application is thus constructed on a general foundation, which is why large language models are also referred to as “foundation models”.[9] This method differs significantly from “narrow AI” systems, which are trained for a specific task on the basis of a specific dataset.

Figure 1: Development phases of generative AI[10]



Source: Rathenau Instituut

5. Implementation

Finally, the model is applied in practice. This can take various forms: as a chatbot or as an image generator (Midjourney), or incorporated into a search engine. In Section 2, we consider possible applications. In the case of a chatbot, users can once again generate different results with prompts, and interact productively with the generative AI. For example, they can brainstorm with a chatbot about the title of a book, with the AI generating options based on certain instructions. The answer depends partly on how the instructions are given. The user can also use reasoning steps to try to teach the AI system something so as to get the right answer.[11]

1.3. Who develops it?

Since 2018, several technology giants have developed large language models by building on their AI know-how and products. They include OpenAI, Google, Meta and Microsoft in the US, and Baidu in China. They have created models focused not only on language but also for other modalities, such as CodeX (OpenAI), protein structures (such as Deepmind's AlphaFold), and robotics (such as Google's PaLM-E).[12] The tech companies – either themselves or through partnerships – often have both the necessary infrastructure (the computing resources), and the data to train the models, the models

9 Bommasani et al., 2022

10 Figure modified based on Bandi et al., 2023; Bommasani et al., 2022; Zhao et al., 2023.

11 One can, for example, ask a maths question that a chatbot does not know how to answer properly. If it is given the roadmap for solving the sum together with the question, the chatbot will learn how to arrive at the correct answer in future.

12 The big tech companies have each developed multiple models and modalities – we mention only a few examples here. The models work in a similar manner; patterns can be discovered in anything that can be expressed in a row of characters. Based on those patterns, predictions can be made about possible combinations of characters.

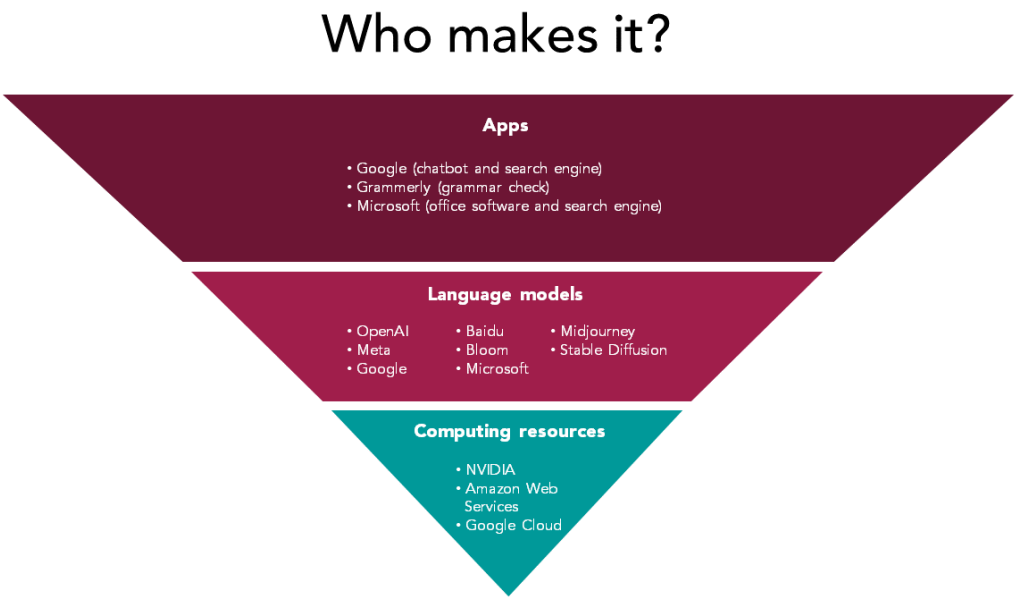
themselves, and the end-user software programs. Microsoft, for instance, has invested billions in OpenAI, is supplying the supercomputers for OpenAI together with NVIDIA, and has announced that it will integrate GPT into its office software.

Because training the models requires a great deal of computing power, hardware and data, it has so far been mainly big tech companies that have been capable of training a language model. Depending on the size of the model, that takes between a few days and several months. Such computing power costs money. Estimates of just how much it costs vary, and they depend very much on the exact model and hardware used. If someone were to wish to train a language model from scratch, with similar computing power as in the models of the big tech companies, the cost will quickly run towards USD 100 million.[13] If the hardware is scaled up even further, these costs can be significantly higher. On top of training, there are the operational costs. Keeping ChatGPT online costs an estimated USD 700,000 a day.[14] This kind of computing power also consumes a great deal of energy and water. Researchers and companies are working on more efficient ways to develop generative AI systems (see Section 3).

Several systems differ in the degree of openness, allowing companies, researchers and private end-users to get to work with language models developed by others. This is possible, for example, by means of a licence for the source code.[15] In the summer of 2023, Meta released Llama 2, whose source code is accessible subject to various restrictions. Start-ups have thus emerged that provide specific services based on existing generative AI systems, for example Jasper and Grammarly. However, one can only speak of truly open development if the source code, datasets, and other training information can be viewed by anyone. An example of a successful open system is called BLOOM, which was created by a collective made up of scientists and developers.

13 Simon, Julien, n.d.
14 Mok, 2023
15 For an overview of degrees of openness, see Solaiman, 2023.

Figure 2: Simplified overview of technology components and actors



Source: Rathenau Instituut

1.4. How good is generative AI?

The results produced by generative AI are impressive, and are a huge step forward in creating high-quality texts. Compared to older AI methods, a variety of tasks can be performed much better.[16] The latest models for example can answer questions much more substantively and correctly than, say, the AI voice assistants Alexa and Google Assistant. Generative AI recognises language better, and can perform tasks that could previously not be automated, such as rewriting a text in the style of a well-known author, or explaining a physics principle. But the most innovative feature of generative AI is the number of different tasks that it can perform. Once the foundation model has been developed, it can be tailored to a specific task with relatively little effort. Moreover, the systems can be linked to other programs, such as search engines, stock market valuations and websites, providing the systems with up-to-date information and allowing them to perform even more tasks, such as booking an airline ticket or buying and selling shares.[17] Developers are integrating more and more functions into their services, such as Meta AI in WhatsApp or voice and speech in ChatGPT.[18]

Notwithstanding their impressive performance, today's generative AI systems also have limitations that relate to their underlying properties.

16 For an overview of progress on different tasks, and a comparison with how humans score (both experts and laypersons), see Rudolph et al., 2023

17 Lorenz et al., 2023; Zhao et al., 2023. Previous models, including GPT3, were unable to do this, and could not provide answers about events that took place after the training phase. GPT3 was trained on data from the period up to the end of 2021.

18 Meta, 2023; OpenAI, 2023

Generative AI is based on data and statistics

Generative AI is essentially based on data and statistics, i.e. probability and likelihood. When creating output, the systems are in fact making an informed guess. This statistical approach can lead to “invented” answers, with the system reporting an event that has not taken place (also referred to as “hallucinating”). When Google's Bard was introduced in the Netherlands, for example, it stated that in the impending elections Hugo de Jonge (a *CDA* minister) would be the lead candidate for the *VVD* party.[19] Systems based on data and statistics can also incorporate biases, which can lead to discriminatory content. Users and developers report errors in the systems on a daily basis. Developers attempt to avoid such errors by curating datasets, utilising human feedback, and applying other “alignment” techniques.[20] So far, however, such data curation cannot guarantee that no more errors are made.

Moreover, there is a further risk: if a system is trained during new training sessions with the erroneous data generated previously, the model’s performance may deteriorate, leading the system into a downward spiral of erroneous data and faulty training.[21] This is termed “model collapse”. Synthetic data can play a role in this. Normally speaking, data refers to something in reality, such as someone's name or place of birth. However, GAI systems can also create, or synthesise, data that does not refer to specific individuals or events. If GAI is then trained again using that data, the representation of reality may deteriorate further.

There is a great deal of debate about the abilities of generative AI systems to correctly represent the world in their calculations, and the differences between human understanding and the operation of algorithms. People think in concepts, such as “an aeroplane” or “a sandwich”, and can grasp causal connections and apply them in new situations. It is questionable to what extent a generative AI system can represent concepts and causes statistically at all.[22] Some experts argue that this problem is unsolvable, but not everyone agrees. The systems can do things that were previously thought impossible, and breakthroughs are regularly taking place that are not yet well understood by researchers. At the moment, not enough is known about exactly what the models can and cannot do. Existing methods are unsuitable for testing the capabilities of generative AI systems, such as abstract reasoning. This is also the object of research.[23]

Language models are complex and therefore difficult to interpret

Developers and researchers have only a limited understanding of the precise algorithmic operation of language models.[24] This means that the models' performance

---

19 Quekel & Hoijtink, 2023

20 The Anthropic company produces AI systems that provide – based on a standards framework, “a constitution” – their own feedback on the results from GAI systems. The company calls this “constitutional AI”. See Anthropic, 2022; Bai et al., 2022.

21 Wong, 2023

22 Bender et al., 2021; Bender & Koller, 2020; Floridi, 2023; Gurnee & Tegmark, 2023

23 Bandi et al., 2023; Kaddour et al., 2023; Zhao et al., 2023

24 Bandi et al., 2023; Bommasani et al., 2022; Bowman, 2023

and results are of limited predictability, which can lead to reliability, safety and security risks. The fact that developers and researchers themselves have only a limited understanding of the language models also means that they sometimes need to revise previous assumptions about the properties of those models. Researchers at Stanford University raised doubts early in 2023 as to whether LLMs can suddenly learn something new, as was previously thought.[25] It became clear by applying different measurement methods that the models did not learn suddenly, but rather step by step. This is a highly relevant discovery, given that the lofty expectations regarding the possibilities and risks of generative AI are bound up with the idea that the models can learn new tasks with little or no additional training. Many scientists therefore consider this kind of research into the underlying workings of the models to be essential.

Box 1 Artificial general intelligence?

The development of generative AI is viewed by some experts as a step towards “artificial general intelligence” (AGI), i.e. systems that can perform so many tasks so well, and reason so structurally, that they possess a very high degree of autonomy. The danger inherent in this may be that these systems will start performing large-scale operations based on goals that developers and users do not have in mind: the “alignment” problem. A number of tech companies and experts are therefore warning of future existential risks of generative AI for humanity. Other experts criticise this view. They consider the development towards AGI to be speculative and feel, for example, that this view distracts from the already present risks of generative AI, namely providing unreliable answers, spreading disinformation, biases, and lack of transparency.

Regardless of what the models can or will be able to do, it is clear that we are no longer dealing with “narrow AI”, in other words an AI system trained for – and good at – only a single specific task. The wide variety of tasks that GAI systems can perform is only expected to increase. This is already having an impact – both positive and negative – in the economy and in society, and the range of issues facing society is increasing accordingly.

1.5. Conclusion

Despite their limitations, generative AI systems deliver impressive performance, and the full potential of the technology has yet to be realised. Based on even more complex statistical pattern recognition and the increase in computing power, the systems will

[25] This property is also known as “emergence”: an algorithm suddenly acquires a skill that it previously did not have. See Schaeffer et al., 2023.

probably become better at a variety of tasks; be able to combine more modalities such as language, text, and video; and be linked to other systems and programs more frequently. This also makes it harder to define where generative AI ends or begins. What exactly AI systems will and won't be able to do in the future is wide open to question, especially given that during development of the systems, human assessment of the data and output is indispensable. There is good reason why expectations are high, but the technology also presents some serious problems that can potentially impede progress. In the following section, we examine the applications and possibilities that can be expected in various domains of society.

2. What are the expectations as regards generative AI?

2.1. Introduction

This section looks at the ways in which GAI technology can be utilised in various domains of society. Scientists, companies, employees, and individuals are experimenting with GAI systems in many fields. In this scan, we highlight some of those fields: education and science, defence and cybersecurity, the labour market, and healthcare. Our analysis is based partly on real-world applications, but it consists mainly of substantiated assessments by researchers, companies, and journalists. As GAI systems are relatively new, there are also still relatively few empirical scientific studies of real-world applications. Most articles attempt to predict what will be possible or assess the ethical issues arising from possible applications.[26] The focus in the present section is on the possibilities offered by GAI; we discuss the social and ethical risks in Section 3. We do however mention possible doubts about the effectiveness of GAI. This section concludes by looking at four roles that GAI can fulfil: as a learning tool, a production tool, a problem solver, and as a tool for social interaction, entertainment, i.e. an experience tool.

2.2. Possibilities in various domains of society

Education and science

The scientific literature lists dozens of tasks that GAI can perform in education, for both pupils and students, and for teachers. Examples include making summaries, writing teaching materials and planning lessons, as well as customising teaching materials, for example for a specific learning style or disability. The systems can also help to assess the work produced by pupils and students, or to construct tests. A chatbot can also provide learning support as a real-time source of information, a study buddy, or a source of ideas and suggestions.[27] The performance of generative AI systems is such that they can pass exams for a variety of school subjects or university courses. ChatGPT, for instance, passed a law exam and a medicine exam.[28] In the future, multimodal GAI models will be able to further improve this experience and assistance.

Frequently cited possibilities for using generative AI in education involve time-saving and efficiency, improved teaching materials, better learning outcomes, and boosting student motivation. Because chatbots can quickly personalise material – for example as accessible text or in a different language – bots also offer opportunities for facilitating inclusion. At the same time, however, there are doubts. Studies point, for instance, to the limited reliability of language models, meaning that teachers and pupils cannot rely entirely on the output, and lose time checking. Moreover, GAI systems are not didactically trained, and not everyone will be able to work with them.[29]

---

26 Sohail et al., 2023
27 Farrokhnia et al., 2023; Jeon & Lee, 2023; Lo, 2023; Sabzalieva & Valentini, 2023
28 Choi et al., 2023; Kung et al., 2023; Lo, 2023. The language models are not equally good in every discipline; see, for example, Lo, 2023.
29 Blodgett & Madaio, 2021; Jeon & Lee, 2023; Lodge et al., 2023; Malinka et al., 2023; Rahman & Watanobe, 2023

Besides education, GAI also offers possibilities in the domain of science. Students and researchers can utilise GAI to conduct literature searches and acquire ideas.[30] A number of scientific journals have already published articles that list ChatGPT as co-author. In 2022, Meta developed a special model to assist researchers, Galactica. Among other things, the system was trained on scientific books and articles, encyclopaedias, and on-line learning materials. It would be able to summarise articles, help solve mathematical problems, and write scientific texts. However, Meta had to take Galactica off line within just a few days because it displayed many incorrect results and biases.

There are high expectations about the ability of GAI systems to solve complex search problems.[31] Such models are trained on specific data, including medical images and texts, protein structures, and mathematical problems.[32] For example, the “search space” when searching for medication consists of about 10^{23} to 10^{63} molecular structures that can be explored.[33] Medical AI models, such as Deepmind's Alphafold, can be used to run through the search space much faster. In chemistry, the hope is that models can assist in the search for synthetic molecules and materials. This does however require sufficient high-quality data.[34]

Defence and cybersecurity

Applications of generative AI systems are being explored in at least two security domains: defence and cybersecurity. In the field of defence, the United States Department of Defense has announced a *Generative AI Taskforce*, which, among other things, will investigate applications for gathering intelligence and for improving administrative processes.[35] Application on the battlefield is a lot more difficult to put into practice but is already being researched. GAI systems could be developed, for example, that enhance strategic decision-making,[36] such as GAI that calculates which plan of attack is likely to be successful, or which plan involves risks that are too high. Decision-making on the interpretation of the laws of war in specific situations could also be supported by GAI. Finally, GAI systems may be able to provide advice on logistics.

Military application of GAI is also related to the discussion on autonomous weapons , such as drones. Scientists, states, and the UN Secretary-General have warned that there must always be meaningful human control as regards to automatic decision-making by systems.[37]

---

30 Grünebaum et al., 2023; Harrer, 2023; Koncz, 2023; Kung et al., 2023; Qi et al., 2023
31 Acosta et al., 2022; Lang et al., 2023; Nógrádi et al., 2023
32 Callaway, 2022; Vogt, 2023
33 Wang et al., 2023. For example, Insilico Medicine uses GAI for “target identification”, which has produced a (potential) drug to treat idiopathic pulmonary fibrosis. See Field, 2023; Philippidis, 2023.
34 Nature editorial, 2023
35 U.S. Department of Defense, 2023
36 Baughman, 2023
37 United Nations, 2023

GAI also offers possibilities in the field of cybersecurity. A GAI system can be instructed, for example, to evaluate a cybersecurity system: where are the weaknesses that malicious parties can take advantage of?[38] Systems such as Microsoft's Security Co-Pilot can also support cybersecurity professionals in their work, answering questions and helping respond quickly to incidents.[39] The GAI system would then be trained based on the user's data and IT environment, and applies the advice accordingly.

In the case of both defence and cybersecurity, GAI applications may prove disappointing. For instance, accuracy is essential in gathering intelligence or ensuring that an IT environment is secure, and is literally a matter of life and death when making decisions on the battlefield. One must always be able to trust that a GAI system will not make mistakes, or can be sabotaged by the enemy. A great deal will therefore depend on the quality of the GAI system.

The labour market

Since the introduction of ChatGPT, there has been a great deal of interest in the potential impact of generative AI on the labour market.[40] On the face of it, what generative AI promises is impressive. It can create new content such as texts, images, audio, or a combination of these, and can be used to produce news articles, promotional literature, summaries, recipes, computer code, or even entire music tracks.[41] A recent study showed that GAI is already being used to provide customer service staff with real-time suggestions during calls and possible answers to questions.[42] The study also found that particularly less experienced employees benefited, because the system takes the most productive and successful employees as reference.

GAI applications are expected to lead to substantial increases in productivity and efficiency.[43] The OECD refers,[44] for example, to “transformative” effects of generative AI, partly because the systems can be deployed in so many sectors: from healthcare to the courts, from industry to banking. GAI is also expected to break down existing language barriers and boost international trade. The broad applicability of GAI technology means that basically any profession can be affected. Initial expectations are, nevertheless, that unlike previous waves of automation, the impact will be mainly on white collar jobs, such as knowledge workers and managers.[45] Follow-up studies will need to determine whether these expectations actually materialise.

---

38 Al-Hawawreh et al., 2023
39 Jakkal, 2023
40 Chohan, 2023; Gmyrek et al., 2023; Knight, 2023; Villasenor & West, 2023
41 Early in 2023, for instance, an AI-generated number by Drake and The Weeknd suddenly appeared. It quickly went viral but was soon taken off line by streaming services and online platforms; see Coscarelli, 2023. See also Bronzwaer, 2023.
42 Brynjolfsson et al., 2023
43 Alshurafat, 2023; Cardon et al., 2023; Noy & Zhang, 2023
44 Lorenz et al., 2023
45 Chui et al., 2023; Gmyrek et al., 2023; Gownder & O’Grady, 2023. During previous waves of automation, routine tasks were particularly susceptible to automation. Generative AI is also capable of performing non-routine tasks.

The most pressing questions relate to what extent GAI systems will take over people's jobs, to what extent people will interact with the systems, and to what extent the technology will make new jobs possible. It makes a difference whether a GAI system is only used to provide linguistic advice when someone is writing a text, or that person can only check the work of the GAI system. Jobs consist of bundles of tasks, and technology will often take over only part of those tasks.[46] Researchers from the University of Pennsylvania and OpenAI (the company behind ChatGPT) have calculated that for 80% of jobs, at least 10% of the tasks involved could be automated by means of generative AI.[47] Workers will thus save time which they can devote to other things, like acquiring new skills. Partly for this reason, the International Labour Organisation (ILO) expects GAI systems will mainly support and augment the work done by humans.[48]

Discussions of the application of GAI in the cultural sector offer examples of how bundles of tasks within a job may change. Although GAI has the potential for being used to generate art works completely independently, it is conceivable that the technology will mainly be used during the phase in which the raw work is created. The artist would then play an important role at the start of a project, when ideas are developed with or without the aid of GAI.[49] Also at the end of the project, when the work needs to be completed and "polished", the artist can still play an important role.[50]

A new understanding of what art is, and of who can call themself an artist, may emerge, as well as a new kind of artist who will benefit from the possibilities offered by generative AI. The ease with which art can be created with chatbots could also lead to a huge volume of new work.[51] New artistic genres may arise; digitalisation has already helped to make new musical genres possible, like Trance and Drum 'n Bass. Artists who use AI for their work are likely to become more skilled at writing prompts and might invent new ways to leave their markon a work.[52] It is also conceivable that the increase in art made with the aid of AI will bring about a resurgence of crafts and handmade work.[53]

Healthcare

According to the Hospitals AI Monitor for 2023 and the Amsterdam University Medical Center (Amsterdam UMC), LLMs offer possibilities for medical practice.[54] Firstly, GAI can be used for information management, and for administrative and writing tasks, thus reducing the administrative burden on healthcare professionals and giving them more

46 Rathenau Instituut, 2015; Went et al., 2015
47 Eloundou et al., 2023
48 Gmyrek et al., 2023
49 Epstein et al., 2023
50 Hugenholtz & Quintais, 2021
51 Epstein et al., 2023
52 Epstein et al., 2023
53 Hugenholtz & Quintais, 2021
54 Janssen et al., 2023; Sparnaaij et al., 2023

time to devote to patients.[55] LLMs can document, transcribe, summarise, classify, and check conversations and information,[56] meaning that they can be used for efficiently analysing electronic healthcare records and other data archives.[57] For example, a doctor may ask a system to retrieve measurements from a surgery report, select suitable participants for clinical trials, or check treatment records in order to identify dangerous drug interactions.[58] GAI can also be called on to assist, for example for scheduling appointments, managing drug intake, producing informative content (such as creating websites, patient information leaflets, and instructional videos), and simplifying medical jargon.[59]

The hope is that GAI systems can also be used to prevent cases of misdiagnosis and to improve treatment. This could, for example, help primary-care physicians to make diagnoses without having to refer the patient to a specialist.[60] The question here is of course whether this would be a responsible approach, given the errors that GAI systems aremaking.

GAI can be combined with other technologies, such as wearable sensors and sensors that allow remote monitoring of a patient. It can also be combined with Unlearn.AI's “digital twins” (medical digital representations), which use GAI models to predict how individual patients' health may change in the future.[61] Combinations are also possible with Brain Computer Interfaces (BCIs). Models can decode brain activity into text faster and more accurately than previous technologies. This enables people who cannot speak because of paralysis due to amyotrophic lateral sclerosis (ALS), for example, to communicate more effectively.[62] An AI thought decoder was recently linked to a Brain-Spine Interface, allowing a paralysed man to walk to some extent.[63] Researchers have also suggested that adding GAI software to BCI decoders could produce a system that could convert dreams (brain activity) into art, sound or video, rather than just text.[64]

Besides professionals, lay people can also use GAI chatbots for medical questions or mental support, for example, with a prompt such as “What is the best nutrition plan for a diabetes patient with high blood pressure?”[65] Research even suggests that chatbots respond more sympathetically and comprehensively than doctors.[66] This is confirmed by

---

55 Harrer, 2023; Van Buchem et al., 2021
56 Marr, 2023; Van Buchem et al., 2021
57 Harrer, 2023; Sweeney, 2021
58 Janssen et al., 2023; Marr, 2023
59 Harrer, 2023; Koncz, 2023; Loh, 2023; Marr, 2023
60 The research presented in Raso et al., 2018 concerns AI in general. For studies specifically about GAI and LLMs, see Acosta et al., 2022; Bell et al., 2023; Janssen et al., 2023; Lang et al., 2023; Nógrádi et al., 2023. Based on a study involving prompts of neurological symptoms, Nógrádi et al. suggest that ChatGPT makes diagnoses with a higher probability of correctness than a primary-care physician. Acosta et al. discuss the possibilities of multimodal medical LLMs. Lang et al. discuss the possibilities that GAI may offer over previous AI for reliable medical imaging.
61 Marr, 2023; Unlearn.AI, n.d.
62 Ravindran, 2023; Tang et al., 2023; Whang, 2023a; Willett et al., 2023
63 Lorach et al., 2023; Whang, 2023b
64 Kelsey, 2023a, 2023b
65 Harrer, 2023
66 Ayers et al., 2023; Korteweg, 2023

the co-founder of the Koko mental health app, who added GPT-3 to the app alongside advice from medical professionals. Messages were, however, rated worse once people discovered that a machine was involved.[67] A comparative study of various chatbots suggests that there is still little evidence that they can provide clear benefits.[68]

2.3. Reflection: the four roles of generative AI

The above overview shows that GAI can take on various roles, often simultaneously. The first is that of a learning tool. This role is readily apparent in education: students can use GAI chat systems to look up information and can engage with GAI, enabling them to quickly place a lot of information in context. In everyday life, many people use digital search engines, and GAI can improve their user-friendliness, which is why Google and Microsoft have integrated GAI into their search programs. The question, however, is to what extent instructors and students can rely on the quality of the systems, and exactly what skills students acquire through their use.

GAI is also used as a production tool: the user wants GAI to create something. In education, students want to have a paper or a summary, while teachers use GAI to produce teaching materials for them. The cultural sector is experimenting with GAI that composes music, writes literature, or paints paintings. Here, humans can cooperate with GAI: for example, a writer can first create a text himself and then ask the system to convert it into a different style or a different language. Someone who wants to produce music can give extensive feedback on the tunes that GAI produces. GAI's production capability is rapid and scalable. Once a GAI system can make something of a certain quality, it can endlessly vary it and perform new tasks. Here, the promise of GAI is that it can deliver high-quality products quickly and cheaply. The question is, however, whether that high quality can indeed be delivered, and also whether workers and society in fact want GAI to take over certain tasks.

The third role is that of problem solver. The hope is that certain complex issues can be solved faster with the aid of GAI. Examples include a general on the battlefield who has to decide what his artillery should target and who can acquire advice through evaluation by a GAI system. Another example of GAI as problem solver is the use of GAI in the development of new drugs and biochemical structures. In cases like these, GAI will attempt to answer questions that people cannot answer, or only with difficulty. If GAI is to play this role, then the answers must of course be sufficiently correct. Nobody can put their trust in a “hallucinatory” system – certainly not in the operating theatre or on the battlefield. Moreover, the point is that people need to be able to identify how the GAI system arrived at certain results, which can be difficult due to the complexity of the calculations involved.

Finally, GAI is also being used as an experience tool. Some users find it fascinating to communicate with a system that has been trained on immense data sets and is able to talk back to them in an understandable and even friendly manner. The Replika chat

---

67 Ingram, 2023
68 Pandey & Sharma, 2023

service, for instance, is designed to be your digital friend. A chat system could become a companion who you consult regularly, as depicted in the film *Her*. Here, the question is whether the experiences we have with GAI systems enrich us as individuals or in fact add but little to our lives, or are even unhealthy for us.

2.4. Conclusion

GAI is a technology that will have an impact in many areas of society, and will be used for a wide range of tasks. It is a new technology and it often remains to be seen how effective and workable it will be in actual practice, also given the various technical weaknesses of GAI. GAI can accomplish complex language tasks with impressive speed and precision, but it is uncertain whether the technology will be reliable enough to support decision-making in hospitals or on the battlefield. Nevertheless, GAI will affect the whole of society and therefore many public values. That is the topic of the next section of this scan.

3. What public values are at stake?

This section provides an analysis of the risks of generative AI concerning the protection of public values. The analysis is based on a study of scientific and "grey" literature.[69] We have grouped the risks into three themes: safety, human-centeredness, and equal and just distribution of benefits and burdens. For each theme, we also specify the public values that are at stake.

By public values, we mean what is considered important in a society and for which systematic protection is deemed necessary.[70] The way we have grouped them is based on various surveys of public values.[71] We have not limited ourselves to these lists, however, given that new issues may also emerge from an analysis of the social impact of technology.
Finally, at the end of this section we reflect on the significance of GAI for democracy.

3.1. Safety

GAI systems can harm people and put pressure on public values in a number of ways: people's privacy and data protection rights can be violated, and the systems can be discriminatory and unreliable in various ways. Together, they put pressure on the safety and trustworthiness of GAI systems.

Privacy and data protection

The training and use of generative AI systems can lead to violations of a person's privacy and data protection rights. Training data utilised by GAI models may contain personal information about individuals or disclose their personal information as output. The EU's General Data Protection Regulation (GDPR) delineates the conditions under which personal data may be processed. A number of European data protection authorities have therefore sought clarification from OpenAI, and are investigating whether the processing involved aligns with the GDPR. The Italian data protection authority has already temporarily banned ChatGPT pending such an investigation. These investigations may require developers to modify their models. Making such modifications can be challenging: OpenAI has already indicated, for example, that it is not currently possible to delete or correct personal data at the request of the concerned individual.[72]

---

69 This involved searching for articles and review studies in scientific databases on social and ethical risks of generative AI and/or large language models using the following search terms: ethical, moral, societal, social risks, implications, challenges, impacts of generative AI and LLM.

70 Here we follow the interpretation of studies in the field of public administration; see for example Bozeman, 2007; Bruijn & Dicke, 2006; Nabatchi, 2018; Riemens et al., 2021. Multiple public values exist and they are dynamic rather than static. Some public values are closely related to human rights, or are already codified within legal frameworks, such as privacy, equal treatment, ownership (right to property), and employment (right to work, and good working conditions).

71 A passage from the Work Agenda [*Werkagenda*], referring to the government coalition agreement: "We have a duty to protect fundamental rights and public values (security, democracy, self-determination, non-discrimination, participation, privacy, and inclusiveness) and the task of creating a level economic playing field with fair competition, consumer protection, and broad social cooperation." Rathenau Instituut studies of digitalisation also highlight values such as health, privacy, human dignity, truthfulness, good work. See for example Rathenau Instituut, 2017, 2019, 2020a, 2020b, 2020c, 2020d.

72 See Norwegian Consumer Council, 2023. OpenAI's privacy policy stating this is available at https://openai.com/policies/privacy-policy.

A GAI system can also derive personal information from interactions with the user. The way someone writes and communicates can provide information about that person's political preferences or health.[73] The GDPR provides additional protection for such intimate data through “special categories of personal data” but this classification does not encompass all intimate information that models can gather, for example such as someone’s emotions.

Moreover, when a GAI system combines multiple data sources, it can derive intimate information from that combined data. One study showed, for example, that a model could "see through walls" by combining camera images of people in a room with Wi-Fi signals from the same individuals electronic devices.[74] After training, the detection of Wi-Fi signals was sufficient for the model to accurately represent in detail the number of people in the room, and their body posture.

GAI models can also easily mimic someone's voice, face or style of work, in such a way that the imitation is indistinguishable from reality. Such a “digital clone” can harm a person, both psychologically and through identity theft. This is a significant risk as it has become easy to create such clones. In 2021, some two minutes of audio material were needed to accurately clone a voice. Now, even fragments lasting a few seconds can achieve the same result. The issue at hand is how to protect one’s personal image in such a digital world.

It is expected that in future, GAI systems will be able to process even more intimate data. Neuroscience researchers are working on interpreting fMRI scans[75] and have succeeded in displaying video images or a photograph of a subject's thoughts based on a brain scan. These studies are still at an early stage of development; currently, for example, they only work for the particular subject on whom the model was trained. Nevertheless, this development raises the question of the extent to which a person's mental privacy and freedom of thought can be protected in the future. Discussion is taking place worldwide on what such “neurorights” might look like. Some countries have already incorporated such rights into their legislation.[76] Another route is to look at existing and forthcomingregulations, such as the GDPR and the proposed AI Act.

Non-discrimination and inclusion

Since the launch of ChatGPT and other generative AI systems, there have been numerous examples of unfair, stigmatising, insulting, or otherwise non-inclusive results being generated. There are various reasons for this. GAI systems are trained with data, and reflect the biases present within that data.[77] In addition, the training data often lacks sufficient representation of specific groups. For example, there is typically more medical

---

73 Kaddour et al., 2023; Solaiman et al., 2023a; Weidinger et al., 2022
74 Geng et al., 2022
75 Chen et al., 2023; Takagi & Nishimoto, 2022
76 La Moncloa, 2021; UNESCO International Bioethics Committee, 2022
77 Abid et al., 2021; Bender et al., 2021; Kaddour et al., 2023; Sohail et al., 2023; Solaiman et al., 2023a; Weidinger et al., 2022

data available on men than on women, and there are relatively few internet pages for smaller language communities.

That imbalance can lead to inaccurate, unjust and derogatory content, and to unequal treatment of individuals or groups. It is precisely those who have historically been the most marginalised who are at greatest risk here. Bias can also lead to physical harm, for example if such systems are deployed in a healthcare setting.[78] Ultimately, biased GAI systems can perpetuate discriminatory social norms, in such a way that those norms will be difficult to alter in the event of mass adoption of the technology.[79] Moreover, the models may include political preferences. German researchers showed, for example, that ChatGPT contained a "pro-environmental, left-libertarian" political orientation. Other researchers found a rightist-authoritarian orientation in ChatGPT4.[80] Precisely because GAI systems can be used for all kinds of tasks, bias can recur in all kinds of different contexts. Developers therefore enlist individuals, often in low-wage countries, to assess the output of the models. These individuals are required to review unpleasant content for minimal compensation.[81]

It is open to question whether one can eliminate, or greatly reduce, the risk of discrimination in these large models. Researchers note that the current technologies for careful data curation are only possible with smaller data sets.[82] The hope is that new technologies will in time become available that will enable data curation even for the vast amount of training data needed for GAI systems. For the present, however, GAI models cannot be made bias-free.[83] Therefore, it is not possible to guarantee that generative AI systems do not discriminate.

Reliability

GAI systems can be unreliable in various ways. On one hand, this may be the result of defects in the system itself, such as containing and providing incorrect information. On the other hand, there are risks associated with the system's usage, for instance if malicious parties use correct information of the GAI system t to cause harm.
Firstly, systems containing and providing incorrect information can cause harm and unsafe situations. It may lead for example to an incorrect medical diagnosis, incorrect advice on medication usage, or program code containing vulnerabilities. Unsafe situations can also arise when people start rely too heavily on the advice provided by a system (similar to stories of individuals blindly following their car's navigation system, thus manoeuvring themselves into dangerous traffic situations). Users themselves can also start spreading the incorrect information (misinformation) that GAI systems

78 Incorrect medical advice may be given, for example. See for example Rathenau Instituut, 2023a. Experts also note the risk of over-reliance on the output from the systems and confirmation bias, i.e. the system may confirm what people already think.
79 This is also referred to as "value lock"; with mass adoption it can become more difficult to change cultural views. See Bender et al., 2021; Weidinger et al., 2022.
80 Feng et al., 2023; Hartmann et al., 2023
81 Hao & Seetharaman, 2023; Perigo, 2023
82 Bender et al., 2021; Mittelstadt et al., 2023; Wachter et al., 2021
83 It is questionable anyway whether bias can be completely eliminated. For an explanation, see Rathenau Instituut, 2022c.

produce – if they believe the it. Another effect is that authentic content is also called into question.[84]

Besides the risk of the systems themselves containing erroneous information, it is possible that malicious parties may deliberately use GAI systems to cause harm,[85] for example by having them write malware, or seeking advice on how to make bombs, dangerous chemicals, or other weapons. Developers are attempting to counter this by teaching the system not to provide such information. In practice, however, users often turn out to be able to bypass the safety barriers that have been installed. It is therefore conceivable that information about dangerous and unwanted content will start circulating. One aspect of this problem is the spread of disinformation and deepfakes, many examples of which are currently online. Malicious parties may start producing convincing fake messages and videos, contaminating public debate and allowing people to be subjected to personal attacks. Porn videos are already being created with other people's faces digitally inserted.[86] GAI systems can be used to tarnish someone's reputation and cause them personal harm. However, dissemination of disinformation and deepfakes can also be politically motivated and can influence democratic debate.

Thirdly, the limited understanding among developers, researchers, and users as to how the models actually work can lead to safety and security risks. It is unclear, and unpredictable, how the systems “behave” in certain circumstances and what vulnerabilities this produces.[87] This discussion also reiterates the aforementioned issue of “alignment”, i.e. can we trust GAI systems to respect public values, legislation, and regulations when performing their tasks? There are worrying examples of systems that are willing to lie to people in order to accomplish a task. A GAI system managed to ask a human to perform visual tasks by tricking them; it did so by claiming to be someone with a visual impairment.[88] There are other such examples. This explains some parties' fears that the long-term security risk of GAI systems is incalculable, namely existential.

The relative secretiveness of private developers about the development of their language models means that there is limited public understanding of the capabilities of those models, and how they evolve.[89] Checking developers' claims and determining the

---

85 See, for example, Doorenbosch, 2023; Europol Innovation Lab, 2023; Gupta et al., 2023; Weidinger et al., 2022; Yamin et al., 2021.

85 See, for example, Doorenbosch, 2023; Europol Innovation Lab, 2023; Gupta et al., 2023; Weidinger et al., 2022; Yamin et al., 2021.

86 For example, various Dutch media personalities have been the victim of deepfake pornography; see Van de Ven, 2023. Unknown individuals can also be affected. In Spain, AI-generated nude photos of dozens of girls were recently circulated on social media. NOS, 2023.

87 Ananthaswamy, 2023; Anderljung et al., 2023; Bianchi & Hovy, 2021; Bommasani et al., 2022; Bowman, 2023

88 This concerns a “CAPTCHA” task, used by websites to distinguish robots from humans. The model asks a person to perform a visual CAPTCHA task. If that person asks why (are you perhaps a robot?), the model lies by asserting that it is a human with a visual impairment. See Edwards, 2023.

89 Currently, researchers and developers have insufficient benchmarks/methods for precisely measuring the capabilities/output of models. See Bommasani et al., 2022; Kaddour et al., 2023; Urbina et al., 2023; Zhao et al., 2023.

security risks their systems pose is therefore possible to only a limited extent. Finally, understanding how a system works is crucial for explainability and accountability. It needs to be possible for individuals to make informed choices, to give or receive redress in the event of errors, and to assign responsibility for preventing errors. This certainly applies in public sectors, including the use of GAI by public authorities, in healthcare, or by the courts.[90]

3.2. Human-centredness

GAI systems influence how people develop and live together. The risks involved impact public values such as human dignity, autonomy, health and pluralism, but they cannot be properly categorised under a single public value. We therefore discuss the risks on the basis of three aspects of human development: cognitive, social, and cultural.

Cognitive development

We noted in Section 2 that generative AI can be used as a learning tool. There are concerns in that regard that this may lead to a deterioration in users' skills, i.e. “deskilling”. Discussion of this issue focuses on higher cognitive skills, such as creativity, critical reflection, and learning skills.[91] GAI systems can frustrate this learning process, for instance if students have the system produce essays for them, or commence every brainstorming process with the ideas provided by the GAI system. This might be less problematic if individuals already possess these skills, as a GAI system can effectively refine them. However, if a GAI application replaces the tasks necessary for people to learn, how can they still learn to master a complex task?[92] Acquiring a higher skill sometimes requires carrying out much simpler work first. Perhaps the best way to fine-tune an essay at a higher level is to have already written essays of your own from start to finish.[93]

Social development

GAI can lead to intimate interaction between humans and chatbots. It is well known that people can become attached to inanimate objects, such as stuffed toys or cars, and therefore also computers. This phenomenon is called anthropomorphism and can also occur when an object or system does not mimic human traits.

In recent years, developers have set themselves the goal of mimicking human and social behaviour as accurately as possible,[94] and with the advent of the language abilities of GAI systems, a new, major step has been taken – especially now that lifelike voices and images are being added.[95] There are already bots that one can teach to

---

90 European Commission High Level Expert Group on AI, 2019; OECD.AI Policy Observatory, n.d.; Rudin, 2019; UNESCO, 2022

91 Blodgett & Madaio, 2021; Farrokhnia et al., 2023; Kasneci et al., 2023; Lodge et al., 2023

92 Malinka et al., 2023

93 At the same time, the advent of GAI will also lead to new skills, specifically when dealing with the system: how can the user get as much out of it as possible?

94 Véliz, 2023

95 OpenAI, 2023

behave like a deceased loved one[96] and chatbots such as Replika that can be set up as a friend. Chatbots – which unlike humans are always available – respond instantly and have endless patience. There is therefore good reason why Replika advertises with the message: *"Replika is for anyone who wants a friend with no judgment, drama, or social anxiety involved."*[97]

Although anthropomorphism can have positive aspects, there are concerns about how this phenomenon affects social development and social interaction. Researchers have been studying this issue for some time. What if virtual encounters are so addictive that people lose the need for interaction with others? Will they unlearn how to deal with other people if they become used to the features of a GAI system?[98] What are the psychological consequences of relying on a chatbot for emotional support? Would it be beneficial for the grieving process if someone creates a chatbot that imitates their deceased loved one?

It is important to find answers to these questions, and to ensure that human interaction is not undermined by our dealings with computers.[99] Extra caution seems warranted in the case of vulnerable groups such as children, knowing that they are the specific target of a number of chatbots. Snapchat's chatbot has already offered to actually meet up with children in the real world.[100]

Cultural development

The emergence of GAI systems may also affect cultural developments: bots will be involved in, or will create, cultural expressions. This raises the question of whether in the future people's creative abilities will be sufficiently drawn upon and able to develop. If every painting starts with a preliminary version generated by a GAI system, it will impact the artistic abilities of artists. If you can produce countless images and texts just by pressing a button, will there still be the time and space for unconstrained and unexpected creative processes? The output can reinforce cultural and artistic norms because the datasets contain historical examples. This mechanism can put pressure on the pluriformity of cultural expression. The question is then how *new* the output of the systems actually is. Moreover, ideas and styles that receive attention on social media will be widely copied, and certain cultural views and choices may thus become overrepresented in datasets.[101]

The developments outlined above raise the question of what it means to be human in a world of robots. What human activities do we not want to outsource? And what human skills do we not want to lose? Because GAI systems have a much better command of language than previous AI applications, they can engage, entertain, and entice us. That makes it attractive to utilise this technology more and more, and to constantly surround

---

96 Fagone, 2021
97 Apple, n.d.
98 Turkle, 2015
99 Danaher, 2019, 2020
100 NOS Nieuws, 2023. This has since been altered.
101 Epstein et al., 2023

ourselves with it. But do all these individual choices lead to a society in which we can thrive?

3.3. Justice

Generative AI systems can also influence who benefits or loses from its development and use. Generative AI systems can take over work from humans, thus influencing the economy and who has a job, and also the quality of that work. We already noted that powerful tech companies can strengthen their position further through GAI systems. There are also issues regarding the environmental impact of the training and use of these systems. In this subsection, we discuss four public values that can be negatively impacted by GAI systems: ownership, sustainability, employment, and quality of work. These aspects all have to do with the equal and just distribution of the benefits and burdens of this technology.

Ownership

An important question raised by GAI systems is whether developers are legitimately using the data with which they train the language models. That data in fact consists of countless contributions by artists, writers, researchers, translators, and programmers that have been digitised over the years. Without those creative works, language models could not operate.

Copyright is protected as a fundamental right under the right to property, and protects people's creations – such as texts, films, and software – as a means of recognising, rewarding, and encouraging the creation of art.[102] “Makers” are seriously concerned about their future due to the massive use of creative work in the models. Meanwhile, a number of lawsuits have been filed against developers of GAI systems. The courts will need to clarify exactly how copyright, and the statutory exceptions for text and data mining, apply. There is also the issue of how to deal with output from GAI systems that imitate existing work but deviate from it only slightly, thus not infringing copyright. Is this desirable?

Another element of ownership involves who ultimately owns and controls the technology. There are concerns about the major role played by just a handful of big technology companies, such as Meta, Google, and Microsoft. We discuss these concerns in Section 3.4.

Sustainability

In the first section of this scan, we discussed the computing power required by big tech companies to train large language models. That computing power has an environmental impact, in terms of energy consumption, CO_2 emissions, and water consumption. Scarce resources are also needed to build the hardware that is used. Researchers are working on more energy-efficient solutions.[103] At the same time, the digital transition –

---

[102] Visser, 2023
[103] Lorenz et al., 2023; Patterson et al., 2021; Vipra & Myers West, 2023

including artificial intelligence and large language models – is viewed as essential in order to achieve the sustainable transition (the “twin transition”).[104]

There are currently no standardised methods for measuring the environmental impact of AI (and generative AI); different methods measure different aspects.[105] For example, one study estimated that integrating language models into search engines could lead to 4-5 times more computing power being needed per search, thus leading to a significant increase in energy consumption and CO_2 emissions.[106] Over the last year, several developers have reported about the energy use of their model’s training, ranging from 433 Mwh of electricity for BLOOM, up to 1.287 Mwh for GPT-3.[107]. A limited number of studies so far have looked into the energy demands of the inference phase: once the model is deployed, it learns from new data.[108] Next to energy demands, the development and use of GAI also has a broader environmental impact, for example regarding e-waste or cooling water. One study estimated, for example, that each chat conversation with some 20-50 replies requires about 500 ml of cooling water – roughly a bottle of water per session.[109]

Because of rising costs, researchers and companies are working on alternatives that are more efficient and therefore more economical, such as using less computer memory by means of (Q)LoRa, more efficient algorithms, new technologies such as quantum computing, smaller models, and more high-quality data.[110] Technologies such as (Q)LoRa already make it possible to further refine raw models with consumer hardware, thus making alteration of existing models more accessible and economical for more parties.

At the same time, the expectation is that technological solutions alone will not be sufficient to realise a more sustainable IT infrastructure..[111] Growth of the digital infrastructure also depends on the demands of users, society in general, and politicians – and the choices made in this regard. It therefore remains a challenge to make the digital economy sustainable.

Employment and quality of work

In Section 2, we noted that GAI can be used as a production tool. This can make workers more productive, but it can also mean that work is automated. A great deal of public debate on this matter focuses on potential job losses due to GAI. There are alarming reports, especially from the tech sector itself, about the impact GAI could have on the labour market. For instance, the CEO of OpenAI has warned that artificial

104 European Commission, 2020; Muench et al., 2022
105 Lorenz et al., 2023; OECD, 2023b; Rathenau Instituut, 2022a; Solaiman et al., 2023b; de Vries 2023
106 Stokel-Walker, 2023
107 De Vries 2023
108 Verdecchia et al., 2023
109 Li et al., 2023
110 Ananthaswamy, 2023; Dettmers et al., 2023; Vipra & Myers West, 2023
111 Rathenau Instituut, 2022a; De Vries 2023

intelligence will inevitably lead to job losses, probably already within the coming decade.[112]

History suggests that technological developments do eliminate jobs, but up to now more new jobs have been created, and jobs have mainly changed due to technology.[113] Past concerns about mass unemployment have not materialised, but the new, and changing, job did require new skills from the workforce. Through training and government policy, it has been possible to meet this changing demand for skills and to provide protection for vulnerable groups, i.e. those who lost their jobs and for whom retraining was difficult. During previous technological revolutions, it was specifically routine tasks that proved susceptible to automation. Since the advent of AI, it has become apparent that various non-routine tasks can also be automated.

In recent years, there have been studies on the impact of automation and AI on the labour market. These point to potential effects such as job and wage polarisation, short-term unemployment, or a more unequal distribution of income or wealth.[114] The OECD reported an expected growth in the number of jobs that may be affected by AI from 14% in 2019 to 27% in 2022. This may increase with the advent of GAI.[115]

A study by the International Labour Organisation (ILO) of the United Nations calculates that the highest level of automation will be among office jobs and knowledge workers.[116] According to the ILO, this can be explained by the fact that, unlike previous technologies, GAI is also able to perform non-routine cognitive tasks. The ILO nevertheless concludes that in most cases, it will be only part of the package of tasks that will be automated, meaning that technology is likely to support and augment workers rather than replace them. Researchers also expect GAI to take over tasks from knowledge workers, but that new tasks will emerge.[117]

Besides the quantity of work, the quality of work is also part of the discussion on the automation of work. Is the work done by humans also *good* work? These concerns also arise in connection with the advent of GAI. For instance, partial automation of their work could worsen the position and working conditions of workers, for example by making their work less challenging.

---

112 Andersen, 2023; Felsenthal & Perrigo, 2023. OpenAI previously conducted its own study of the potential impact on the labour market. Eloundou et al., 2023.

113 Jobs consist of packages of tasks, with technology taking over part of the package and the work done by humans changing. Rathenau Instituut, 2015; Went et al., 2015

114 Job or wage polarisation means that wages rise at the bottom and top of the labour market, putting pressure on the middle segment. This phenomenon has been apparent for some time. See van den Berge & ter Weel, 2015. Frictional unemployment involves a temporary rise in unemployment due to the mismatch between the skills required and available within the labour force. See Koning, 2013. Inequality in income can increase as the share of wages in national income falls as income derived from capital grows. See Autor, 2022. See also Brynjolfsson, 2022.

115 OECD, 2023a. In 2019, the OECD estimated this rate at around 14%. OECD, 2019.

116 Gmyrek et al., 2023

117 Autor, 2022

The quality of work is also under pressure due to the working conditions of those who assess the data and output of the models. It is known that processing hateful, discriminatory, or otherwise illegal or unwanted content can cause psychological harm, for example because people are repeatedly exposed to obscene or shocking utterances.[118] Such work is also often carried out under poor working conditions, by workers with little labour protection and in a precarious situation, for example refugees. This leads to harmful practices such as the workers being underpaid or refused payment. In addition, this work is only documented to a limited extent, which adds to the lack of transparency about how language models work.

The OECD warns of the above risks and therefore calls on countries to improve support for low-income workers, to invest in safety measures and responsible use of GAI in the workplace, and to invest in new skills.[119]

3.4. Reflection: GAI puts democratic society under pressure

A central feature emerges from the overview of public values and risks, namely that development and use of GAI puts democratic society under pressure in two ways. First, there are concerns about public debate being influenced through the spread of disinformation and misinformation, and the consequences this can have for social and political trust and for communication between individuals and elected representatives. GAI can undermine democratic processes. Second, the growing power of a few tech companies in an increasing number of areas of society may impede democratic governance of digital technology.

Democratic processes

There have long been concerns about the increase in online disinformation, and GAI is exacerbating those concerns. The news and public debate are important sources of information for people to form their opinions on societal topics and political issues. It is therefore important that everyone has access to reliable and truthful information. Unfortunately, GAI systems make it possible to generate false and misleading information on a large scale. For example, the US news watchdog NewsGuard identified in August 2023 37 websites using chatbots to copy and edit articles from media outlets such as CNN, *The New York Times*, and Reuters.[120] These sites use software that can find, rewrite, and publish articles without human intervention. In March 2024, NewsGuard identified 739 generated AI news websites worldwide that run with little or no human oversight and produced false claims.[121]

There is also a risk that people will trust the information GAI systems give too much. The systems can then have a major impact on the opinions and world view of the user. They may also play a powerful intermediary role in elections by influencing what information a user consumes. That effect may reinforced when GAI is utilised for “hyper-personalisation”, which means that users are only shown news that interests

---

118 Hao & Seetharaman, 2023; Norwegian Consumer Council, 2023; Perigo, 2023
119 OECD, 2023b
120 Brewster et al., 2023
121 Sadeghi et al. 2024

them or confirms their world view. This may result not only in the absence of shared truth or agreement on facts, but also of shared experiences.

Finally, GAI could weaken the ability of elected representatives to respond to public concerns.[122] GAI systems can be abused, for example, in democratic processes such as public consultations, by making submissions on a large scale in an attempt to influence the outcome of the consultations.[123] Malicious individuals can also use social media or parliamentarians' mailboxes to convey a distorted picture of public opinion.

Democratic governance of digital technology

The second risk has to do with the growing position of power of just a few tech companies. A significant number of language models and applications are now owned by tech companies such as Microsoft, Google, and Meta. It is they who decide who has access to the models, under what conditions and at what price, how they are trained, with what data, and what content is prioritised or in fact filtered out. In other words, it is the tech companies that largely determine the values that the GAI application takes into account.

Through their social media platforms, these companies also influence how news and political ideas spread, and how public debate takes place. Concerns about the power of tech companies are not therefore limited to their market power but also relate to their influence on public domains such as education, healthcare, journalism, and the rule of law. Within those domains, public institutions are becoming increasingly dependent on the services of just a few technology companies. This places a strain on the shared ability – with individuals, civil-society organisations, public institutions, and businesses – to determine how digital technology is developed and applied within society.[124]

A new issue in this regard is the influence of a few tech companies on the development of scientific knowledge. Training advanced language models requires so much computing power that it makes the associated price tag unaffordable for many academic institutions, thus making them dependent on the tech companies' models. Moreover, it is difficult to verify the knowledge claims of the tech companies, given that public knowledge institutions generally have limited access to the underlying training data, while there is also a lack of transparency as to how the technology giants' language models actually work. Finally, a number of researchers have long warned that too much emphasis on data and statistics can impoverish science, for instance when correlations are equated with the creation of scientific knowledge and associated methods.[125] Therefore, there is an ongoing debate as to what extent, and under what conditions, scientists should use large language models.[126] Some argue that socially responsible science should only use models that are open to clear interpretation.

---

122 Kreps & Kriner, 2023
123 Bell et al., 2023
124 Kerssens & van Dijck, 2023; Nemitz, 2018; Passchier, 2021; Sharon, 2016; van Dijck et al., 2018
125 Anderson, 2008; Haggart, 2023
126 Bender et al., 2021; Birhane et al., 2023; Rudin, 2019

3.5. Conclusion

The previous section showed that GAI offers countless possibilities, but that it is not yet good enough to be used for critical processes in healthcare or defence. This section has shown that GAI technology is associated with a large number of risks in relation to public values. Some of these issues are not new within the digital society, for example privacy, discrimination, disinformation, sustainability, and the power of tech companies. GAI reinforces and complicates existing problems.

For example, preventing discrimination in learning systems was already not easy, and currently impossible with GAI. The environmental impact of the digital infrastructure was already a concern, and GAI is now making it significantly more so. Countering the spread of disinformation was already difficult, and GAI technology makes producing it significantly easier. Learning AI systems were already a black box, and GAI increases this complexity only further. Technology companies were already powerful, and with GAI, they can significantly strengthen their economic position and the role they play in providing society with information. One must also bear in mind that these language models can form the basis of many other applications.

Some of the risks of GAI are new; they played little or no role in digitalisation and AI policies in the past four years. During the term of office of the last government, there was little discussion of copyright law, for example, or the impact of automation in the workplace and the transition of work due to technological changes. And although the powerful position of technology companies was already on the political agenda, their influence within science is an additional concern. Academic and public knowledge institutions generally have only limited access to the underlying data and source code of the models. Developers' claims – for example regarding operation, new features, safety and other risks – are therefore difficult to verify.

The existing and emerging risks of GAI add up, all in all, to a worrying picture, one that affects democracy and is to an extent already manifesting itself. Users report misinformation, bias, and deepfakes on a daily basis. It is therefore necessary to confront the risks posed by GAI. In the next section, we will explore how policy-makers, politicians, businesses, civil-society organisations, and individuals can do so.

4. What policy options to take?

In the previous section, we noted that GAI technology amplifies existing risks within the digital society and introduces new ones. The companies developing GAI applications still seem to be following the motto “move fast and break things”, i.e. introducing applications and then look to see exactly what happens with them. That approach is a risky one. Without mitigation, the adverse effects of GAI applications may come to dominate.

To an extent, mitigation policies have already been put in place. Over the past decade, a great deal of work has been done to steer digitalisation in the right direction, and AI in particular. At international level, ethical principles have been formulated for socially responsible applications.[127] With existing and new legislation, the EU is seeking to entrench these principles in legislation; examples of such legislation include the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), the Digital Markets Act (DMA), and the forthcoming AI Act. In the Netherlands, public values and societal challenges have also become increasingly central to digitalisation policy.[128]

The key question is whether these efforts will be sufficient. There is a real possibility that current and proposed policies may be unable to cope with the impact of GAI systems, for example as regards non-discrimination, safety, competition, disinformation, and worker exploitation. It is therefore imperative that the Dutch government formulate a strategy to strengthen society's grip on this technology. That starts by thoroughly evaluating and making timely amendments to Dutch and European policies, from legislation to financial and communication mechanisms.[129] Given the broad range of public values at stake, this is a task that is both complex and urgent.[130] It is incumbent on the government to take up this task, and to shape the development and use of this technology responsibly together with the business community and civil-society organisations.

The government has already indicated that it is considering to set up a “rapid response team”, i.e. a group of experts – from within and beyond the public authorities – to advise on generative AI. It has also indicated that it wishes to consider having a team that "not only advises, but can also intervene rapidly...". It is worth exploring how existing expertise can be utilised. At the same time, the intention to enable "intervention" raises the question of what exactly that means, and how the work of the team will relate to that of supervisory bodies and to democratic accountability for policy.

Given the wide range of relevant policies, it was not possible to undertake a comprehensive policy analysis within the scope of this scan. Based however on a

---

127 European Commission High Level Expert Group on AI, 2019; OECD.AI Policy Observatory, n.d.; UNESCO, 2022

128 Ministerie van Economische Zaken en Klimaat, 2019; Rijksoverheid, 2022b

129 See the mechanisms referred to in the “Policy Compass” [*beleidskompas*] (Kenniscentrum voor beleid en regelgeving, n.d.), and the publication *Stand van digitaal nederland* (Rathenau Instituut, 2021).

130 One can build on the studies already carried out by advisory boards and knowledge institutions on the broader governance of AI and digitalisation. See Cyber Security Raad, 2021; European Parliamentary Research Service et al., 2021; Onderwijsraad, 2022; Raad voor de leefomgeving en infrastructuur, 2021; Raad voor het Openbaar Bestuur, 2021; Rathenau Instituut, 2021, 2022a, 2022b; Wetenschappelijke Raad voor het Regeringsbeleid, 2021.

review of the relevant literature, interviews with experts, and a workshop with policy officials, the Rathenau Instituut proposes five options for action on the part of the Dutch government:

1. Make it possible to take harmful GAI applications off the market;
2. Ensure that legal frameworks are future-proof;
3. Invest in international AI policy to guide global innovation processes of technology companies;
4. Set an ambitious agenda for socially responsible GAI;
5. Encourage public debate on the desirability of GAI.

4.1. Make it possible to take harmful GAI applications off the market

It is conceivable that a GAI application does so much harm that it needs to be taken off the market. Society must then have the option to do so. The question is whether there is at present an effective legal mechanism for doing so. Specifically, the upcoming AI Act, which final text is currently being drawn up, is being considered as regards this point. We will discuss what is known at the time of writing.

It is likely that the AI Act will require that all “foundation models” be permitted to enter the market only once certain aspects have been evaluated and the development process has been documented. If a GAI application can also be categorised as a “high-risk AI system”, the developer must subject it to a “conformity assessment”. Such an assessment also forms part of the EU’s rules on product safety. Before the GAI system can be released, a check must be performed to ensure that it satisfies a wide range of requirements regarding data quality, cybersecurity, and documentation.

Depending on the type of system, the developer may perform the assessment themself, or an independent third party must do so. Exactly which LLMs and GAI applications the latter will apply to, remains to be seen.[131]

Unlike the safety rules for other products, the AI conformity assessment also comprises protection of human rights. For example, the documentation must specify the risks to fundamental rights, and a wide range of measures must be put in place to prevent bias in training data. In addition to conformity assessment by providers, the European Parliament's proposal includes a mandatory Fundamental Rights Impact Assessment

[131] Following the outcome of the final negotiations in December 2023, the AIA will introduce obligatory independent assessment and stricter rules for “highly capable foundation models” or “high-impact foundation models”, such as GPT-4. The criteria for such models will include, for example, the computing power and the amount of training data. Because such foundation models pose “systemic risks” (a term used in the Digital Services Act (DSA)), the legislature will centralise enforcement by a yet-to-be-established European Artificial Intelligence Board (EAIB). Besides the additional rules for highly capable foundation models, additional requirements have been proposed for AI applications based on foundation models and deployed by more than 10,000 commercial users or used by 45 million private individuals. The foregoing rules, that were introduced during the final negotiations, resemble a DSA and DMA approach. See Bertuzzi, ‘AI Act’; Bertuzzi, ‘EU Policymakers Enter the Last Mile for Artificial Intelligence Rulebook’. However, countries including Germany, France, and Italy criticised the rules for foundation models in November, stating that they would hamper start-ups too much. Discussions between the EU institutions on regulation of GAI systems are therefore still ongoing. If no consensus is reached in December, the entire piece of legislation could be jeopardised, given that the European Parliament will be dissolved for fresh EU elections next year. See Bertuzzi, 2023c.

(FRIA) for *operators* of high-risk (G)AI systems. The European Parliament reasons that applications may lead to risks to fundamental rights that could not be foreseen during the development phase.[132]

In addition to internal and external checks prior to the launch of GAI applications, there will also be subsequent enforcement and monitoring by supervisory bodies. They will have the same instruments at their disposal as under other product safety regulation. If the supervisory authority decides to investigate a conformity assessment and finds that the application fails to meet the set standards, then the application can be taken off the market. Introduction of the FRIA would make it possible to prohibit applications whose risks to fundamental rights have not been sufficiently registered and mitigated by the operator.

In theory, the AI Act could therefore provide society with the necessary set of controls. There are various problems and open questions, however. First, most providers and operators of GAI systems will not yet have the necessary expertise regarding human rights for ensuring that their products meet the set standards, and for checking that themselves. Second, having a set of controls requires there to be effective enforcement. The large number of AI providers, and the required coordination between the various regulatory domains (data protection, market regulation, product regulation, etc.) can be expected to bring about a significant increase in the enforcement burden.

Third, product safety regulation so far is based on norms that have been translated into measurable, verifiable standards, for example electromagnetic radiation levels in smartphones. The AI Act introduces human rights into product safety regulation, but the translation into measurable, verifiable norms is more difficult, as human rights requirements are open norms. This is a more fundamental problem. When, for example, has the risk of discrimination been reduced to an “acceptable” level? And acceptable to whom? There are several bias detection methods, but they are linked to different ideas of what is meant by “fair”..[133] In practice, the answer will be provided by the standards that standardisation organisations such as the ISO and the Netherlands Standardisation Forum will develop, and by the litigation that is likely to follow.

It is therefore important that politicians and policy-makers continue to evaluate whether the AI Act currently addresses the risks of GAI sufficiently. The AI Act can be strengthened, for example by introducing a system of permits, with a supervisory body issuing prior approval. There could also be a requirement to consult the supervisory body in the case of highly impactful applications, along the lines of the existing “prior consultation” provided for in the GDPR.

---

[132] Article 29a and recital 58a AI Act (EP version).
[133] Rathenau Instituut, 2022c

In the political agreement of the AI Act in December, an approach similar to DSA and DSA is adopted to address the 'systemic risks' of 'high impact foundation models'.[134] Providers of these models must evaluate and mitigate risks periodically and report serious incidents for example.[135] However, the criteria used to indicate and evaluate the capabilities of such models are still subject of research, and uncertainty remains which risks may fall out of scope.[136]

Finally, it is important that politicians and citizens continue to be involved in the standards that set out which human rights requirements GAI systems must meet – and to continue to evaluate them critically.

4.2. Ensure future-proof legal frameworks

Besides the AI Act, there are a number of other legal frameworks that require attention. Some issues involve the possible clarification or adaptation of existing frameworks. Other questions are more fundamental: does the chosen type of system still reflect the potential impact of GAI on society? Below, we summarise the main issues below that emerged from our survey.

Data protection

The General Data Protection Regulation (GDPR) applies when GAI systems process personal data, but the question remains whether intimate data will in fact be sufficiently protected. GAI systems can, for example, collect information about a person's mood ("sentiment analysis"). Generally speaking, this kind of information does not fall into the "special categories of personal data" that enjoys extra protection under the GDPR. Consideration might be given to expanding these categories to include mood, speech, and facial and body data, so as to protect individual people's data more effectively. GAI systems are expected to be able to process more and more types of sensitive data, including information from brain scans. Discussion has therefore commenced on how freedom of thought (mental privacy) can be protected.[137]

Discrimination

Discrimination is prohibited under Dutch and European law. Nevertheless, social media are full of examples of GAI systems that express prejudice and reinforce stereotypes. This can lead to unequal treatment. Developers therefore need to combat bias within their systems.[138] The forthcoming AI Act will also impose requirements on developers as regards the quality and governance of data. The question is whether these measures can significantly reduce discrimination. Researchers note that the current techniques for careful data curation are only possible with smaller data sets. This makes it difficult at present to ensure that GAI models are "bias-free". Given these technical limitations, government will need to carefully consider what anti-discrimination requirements it will

134 See footnote 132
135 Rathenau Instituut, 2022c
136 Kutterer (2024)
137 See also the discussion of neural rights in the Rathenau Scan on immersive technologies Rathenau Instituut, 2023b.
138 College voor de Rechten van de Mens, 2021; Van Bekkum & Zuiderveen Borgesius, 2023

impose, and what should be done if GAI systems are unable to meet those requirements.

Safety and security

Generative AI systems involve multiple safety and security risks. For instance, they can disseminate false or misleading information, produce unpredictable results, and be misused so as to spread disinformation or carry out cyberattacks. In recent years, the Netherlands and the EU have already invested heavily in cybersecurity, including by tightening up legislation and regulations and introducing a statutory framework for product security. The upcoming AI Act is also expected to impose requirements for the performance, interpretability, and cybersecurity aspects of systems. The question is whether that will be sufficient to prevent misuse andabuse. Moreover, the relative secretiveness of private developers means that researchers and the public have little insight into how language models are developed and what abilities they acquire. What level of insight and control do politicians consider necessary?

Disinformation

Important legal frameworks to combat disinformation include the Digital Services Act (DSA) and the forthcoming AI Act. There is debate as to whether these frameworks are sufficient, given that GAI systems make it easier to create convincing disinformation, ranging from credible emails to deepfakes.[139] For example, the Dutch Public Prosecution Service is exploring the possibility of using the Criminal Code to tackle “deep nudes”, i.e. manipulated videos in which people are involuntarily shown as naked and perform sexual acts. An alternative is to ban specific uses of deepfakes, such as in a pornographic context, or to regulate deepfakes as a high-risk application in the AI Act. Currently, the AI Act only contains transparency obligations for parties that generate deepfakes.

Copyright

Copyright, as part of the fundamental right to property, protects “makers” of works, with the basic principle being that it is only they who may duplicate their work or make it available to the public. There are issues concerning copyrighted material regarding *input* (training data) and *output* (the content created by the systems). As regards input, the training data used for language models is highly likely to contain copyrighted material. A number of lawsuits are currently pending to clarify whether the material concerned was obtained lawfully. One of the problems here is that copyright law as we know it today places a great deal of responsibility with the rightholders themselves: it is they who must impose provisos. However, rightholders cannot always easily retrace whether their work has been used to train GAI models. Does this produce an environment that encourages the creation of art and other works and sufficiently protects property?[140]

---

139 We should note here that policies to counter disinformation involve more than just judicial frameworks, for example investment in high-quality media and in raising awareness.

140 Visser, 2023

The question is whether the current legal framework sufficiently discourages developers from using copyrighted work illegally. One needs to remember that harm is quickly done, because once a model has learned certain qualities and styles, it cannot simply "unlearn" them.[141] It is unclear how copyright should relate to GAI systems that imitate existing work but deviate from it sufficiently to not be covered by copyright. Effectively, this will mean makers competing with GAI systems. Policy-makers and politicians need to ask themselves what kind of copyright they want: AI-friendly copyright or a legal framework that protects makers?

Competition law

In recent years, various big technology companies have built up a strong position in language models, but also in social media, search engines, and cloud infrastructure. Researchers wonder whether current competition law is adequately equipped to curb the economic power of these conglomerates and whether the recently introduced Digital Markets Act (DMA) can actually alter that power.[142] A more fundamental matter is that researchers indicate that technology companies not only wield economic power but can also influence social and political issues. Various public domains are becoming more dependent on the services offered by these companies. These problems may require more than creating the conditions for "fair markets" – the policy strategy followed so far.[143] A research group has therefore been established in the Netherlands to determine the extent to which this issue can be addressed from the perspective of constitutional and administrative law.[144]

Finally, the role of supervisory authorities

Scrutinising the frameworks described above will take time, and it is therefore important that existing statutory frameworks are properly complied with in the meantime. Supervisory bodies play a crucial role in this through enforcement, and it is important that they adopt an assertive stance vis-à-vis GAI. In turn, policy-makers and politicians have a role to play in creating the right conditions, namely by ensuring that supervisory bodiess have sufficient resources and expertise their expertise.[145]

4.3. Invest in international AI policy to guide global innovation processes of technology companies

The cross-border nature of generative AI and its developers makes it important for the Netherlands to promote international cooperation as part of a strategy to govern the technology in a socially responsible way. The hope is that the AI Act, as well as other EU legislation, will have an effect internationally, with bodies elsewhere taking EU

---

141 Wong, 2023

142 For example, the DMA intervenes only to a limited extent as regards the underlying sources of power, including computing power, early mover advantages, data resources, and integrated systems. In addition, the designated "gatekeepers" and "core platform services" do not include GAI services, except where GAI is integrated into search engines. The DMA includes provisions that – if declared applicable to GAI services or developers – can promote a fair market. Yasar et al., 2023.

143 Gerbrandy & Phoa, 2022; Nemitz, 2018; Passchier, 2021; Sharon & Gellert, 2023

144 Jak & Lokin, 2023

145 An EU supervisory authority working group is already making preparations for the AI Act. It is chaired by the Dutch Authority for Digital Infrastructure.

legislation as their point of departure. In this connection, the term “Brussels Effect” is sometimes used.[146] This effect could hold true for the AI Treaty currently being finalised by the Council of Europe, if private companies fall within the scope of the Treaty.[147] The treaty aims to offer people greater legal protection, for example by allocating those affected by AI decisions the mechanisms they need for objection and redress. Court rulings on generative AI can also have global effect; see the European Court of Justice's ruling in the *Schrems I* case.[148]

It is currently too early to say whether these European agreements will indeed have such an international effect. It is clear, however, that the United States and China are also pondering regulation. The US recently decided, for instance, that GAI applications utilised by government must comply with stricter conditions.[149]

But even if other countries take inspiration from EU legislation, it is important to establish international arrangements so that companies around the world can be held accountable. The Netherlands Scientific Council for Government Policy (WRR) has already recommended that the Netherlands should push forward with “AI diplomacy”.[150] In particular, voluntary codes have been drawn up over the course of the past year. The OECD has launched the *Global Partnership on AI* (GPIA), as well as the *Partnership on AI*, bringing together industry, researchers, NGOs, and media organisations. The OECD has also produced a robust framework for corporate social responsibility, addressing such issues as exploitation and environmental impact. The G7 recently agreed on a voluntary code of conduct for GAI developers, and the European Commission has indicated its intention to work with the US to establish voluntary codes of conduct for the companies behind GAI models.[151]

Binding agreements have also been initiated. UNESCO has drawn up the global *Ethics of Artificial Intelligence* framework, by which signatory countries are bound.[152] Multiple countries are also introducing legislation in the area of corporate social responsibility. In Europe, there is the *Directive on Corporate Sustainability Due Diligence*. The more that legislation is in line with the guidelines adopted by the OECD in this field, the more companies and risks are covered.

4.4. Set an ambitious agenda for socially responsible GAI

To gain control of generative AI and ensure desirable applications, it will not be enough to strengthen legal frameworks. Government will also need to actively encourage innovation towards socially responsible AI by means of other policy instruments. This will require an ambitious agenda that comprises at least two elements:

---

146 This term comes from Finnish-American law professor Anu Bradford; see Bradford, 2020.
147 EU Commission’s last-minute attempt to keep private companies in world’s first AI treaty – Euractiv
148 *Maximillian Schrems v Data Protection Commissioner*, 2015
149 Lima & Zakrzewski, 2023
150 This can cover five areas: basic research, commercial applications, regulation, ethical guidelines, and standards.
151 Zubascu, n.d.
152 Sabzalieva & Valentini, 2023

Invest in alternative technology

In Section 3, we showed that the risks associated with the powerful position of a a few big technology companies in developing GAI.[153] There are several ways to allocate a meaningful role to other parties to ensure more democratic governance of the technology becomes possible. Public authorities at both national and EU level could co-fund the development of GAI technology together with civil-society partners.[154] One example is the recently announced GPT-NL project, in which the Netherlands Organisation for Applied Scientific Research (TNO), the Netherlands Forensic Institute (NFI), and the SURF ICT Cooperative are teaming up.[155] The Ministry of Economic Affairs and Climate Policy (EZK) has made 13.5 million euros available. This model is intended to be suitable for use within academia, government and public sectors such as healthcare and education. As a funding body, the government can more easily ensure that legal and ethical standards are respected during development. As regards this point, comparisons are possible with other publicly funded technological programmes, for example CERN, which researches elementary particles, or European public-private partnerships within space programmes – in which Europe has succeeded in achieving some major ambitions.

Open source development of GAI technology is also a possibility. The German and French governments are already working with the *Nextcloud* platform. In this regard, it is important to be critical as to what extent products are actually open source. We noted above that the Llama 2 language model offers only a limited degree of openness. Government can give preference to open source in its own procurement processes, and make it clear exactly what that means in the context of GAI.[156]

It is important for government to invest in science that ensures that GAI technology can be given shape from a public values perspective. This includes by encouraging research into language models for smaller language areas and dialects, security, synthetic data, bias detection in larger datasets, and sustainable GAI technology. Basic research on the interpretability of models should also be organised. Finally, it is important to avoid technological tunnel vision by exploring whether and how other AI technologies can address the shortcomings of GAI.

Monitor and guide developments by sector

Given the broad social impact of GAI, it is important to monitor that impact in actual practice and to pre-empt unwanted effects. How does this technology alter classroom dynamics or hospital administration? What changes does that demand in the curriculum or policies of educational institutions, for example? How severe is the environmental

---

153 The EU is committed to building up "strategic digital autonomy", vis-à-vis both countries such as China and the United States and global technology giants; see European Parliament, 2022; European Union External Action, 2020. Building up public GAI alternatives would be in line with these efforts.

154 Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2023. Where the Gaia project is concerned, there are various concerns as to whether the original objective (autonomy) has not been watered down; see for example Goujard & Cerulus, 2021

155 Digitale Overheid, 2023

156 Rijksoverheid, 2022a

impact of GAI in actual practice? And what changes will there be within the labour market, for instance with regard to wage polarisation or income and wealth distribution? Each sector will need to start thinking about how GAI can really contribute to meeting the needs and achieving the aspirations of professionals. And just as importantly, if generative AI applications fail to deliver the desired results, it is essential that people identify this in time and report it to policy-makers.

4.5. Encourage public debate on the desirability of GAI

In addition to legislative and stimulating policy measures, it is crucial to initiate public debate on GAI. Users often see GAI as a convenient, innocuous technology that anyone can experiment with. But using GAI applications comes with risks.

Those risks call for technological citizenship: people need to be aware of the dangers, to be able to deal with the risks, and to participate in democratic decision-making on GAI technology.[157] Such citizenship begins with education: it is important for every individual and organisation to know what kind of technology GAI actually is, what you can and cannot expect from it, and the risks that are associated with it.

Every organisation, every company, every researcher, and every user will then need to ask themselves under what conditions they wish to deploy GAI technology, and whether that can be done in a socially responsible manner. The AI Act has not yet been adopted, supervisory bodies are still giving shape to their role, and there are many unresolved issues, including with regard to security, discrimination, explainability, data protection, exploitation, and sustainability. There have therefore already been various calls not to utilise the technology, for instance from researchers who feel that use of GAI is incompatible with their social responsibility.[158] UNESCO has also called for GAI technology not to be used in education for children younger than 13.[159]

It is also important that debate regarding GAI addresses *all* relevant questions and issues, for example possible romantic relationships between humans and chatbots or interactions between children and chatbots. What do we consider desirable in this regard? It is illustrative to compare the advent of GAI with that of mobile phones. These devices are multifunctional and for several reasons so attractive that they have become almost inescapable. But there are now concerns about the impact of mobile phones on our physical and mental health. When you see everyone on the train peering at their phone, you realise that technology can also take away something precious from us. Now that computers can talk to us, put us at ease and excite us intellectually, it is all the more important to think carefully about how we as a society wish to spend time with and without AI systems.

Schools, hospitals, libraries, artist collectives, and civil-society organisations all face the task of promoting technological citizenship and furthering debate on GAI. Government

---

[157] For further explanation of technological citizenship, see Est, 2020.
[158] Rooij, 2023
[159] UNESCO, 2023

can facilitate this. Over the next four years, the Rathenau Instituut intends organising a dialogue programme during which members of the public will discuss the future of the digital society. GAI will be one of the topics addressed.

4.6 Conclusion

Throughout our investigation, we noted that views among experts on the possibilities and dangers of GAI technology diverge widely. Some view it as a game changer that will increase the capabilities of AI systems at lightning speed. Others point out the inherent limitations of the technology, arguing that GAI systems usher in a “wrong paradigm" of AI.[160] These divergent views are also apparent in the debate on the risks of GAI: some point to urgent and “classic” issues concerning digital technologies, including privacy, discrimination and security, others focus their attention to potential existential risks to humanity.

We draw the following conclusions. GAI technology does indeed represent a breakthrough in the ability of AI systems to perform language tasks and to combine different modalities. At the same time, just how good the applications of this technology actually are, and to what extent people can relinquish tasks to it, is still very much in question. In addition, GAI amplifies the known risks of the digital society and adds new ones, for example loss of ownership and disruption of human development. These risks do not appear to be easily resolved. It is a realistic possibility that existing and announced policies will be insufficient to mitigate these risks.

Action is therefore needed. The government and politicians must determine where policy needs to be strengthened. In the meantime, they should provide maximum support for regulatory oversight, make arrangements with developers, and warn society about the risks of GAI. Globally, those risks are indeed being taken seriously; every individual and every institution in the Netherlands must do the same.

[160] Marcus, 2020

Appendix: Policy-makers and experts consulted

Experts interviewed

1. Lambèr Royakkers, Professor of Ethics of the Digital Society at Eindhoven University of Technology
2. Pim Haselager, Professor of Artificial Intelligence at Radboud University Nijmegen
3. Haroon Sheikh, senior research fellow at the Netherlands Scientific Council for Government Policy (WRR) and Endowed Professor of Strategic Governance of Global Technologies at VU University Amsterdam
4. Anna Gerbrandy, Professor of Competition Law at Utrecht University and Crown-appointed member of the Social and Economic Council of the Netherlands (SER)
5. Naomi Appelman, PhD researcher at the Institute for Information Law, University of Amsterdam
6. Catelijne Muller, president of the Alliance for AI and member of the European Commission's High Level Expert Group on AI
7. Lokke Moerel, Professor of Global ICT Law at Tilburg University, Senior of Counsel with Morrison & Foerster, and member of the Cyber Security Council
8. Jeroen van den Hoven, university professor and Professor of Ethics and Technology at Delft University of Technology and editor-in-chief of Ethics and Information Technology

Participants in workshop with policy-makers

1. Francisca Wals, Ministry of the Interior and Kingdom Relations
2. Jasper Kars, Ministry of the Interior and Kingdom Relations
3. Haye Hazenberg, Ministry of the Interior and Kingdom Relations
4. Elja Daae, Ministry of the Interior and Kingdom Relations
5. David van Es, Ministry of the Interior and Kingdom Relations
6. Anne Thier, Ministry of the Interior and Kingdom Relations
7. Luca Kuiper, Ministry of Foreign Affairs
8. Mijntje Jansen, Ministry of Economic Affairs and Climate Policy
9. Gelijn Werner, Ministry of Economic Affairs and Climate Policy
10. Noud Louwers, Ministry of Economic Affairs and Climate Policy
11. Vincent Pot, Ministry of Education, Culture and Science
12. Cyril van der Net, Ministry of Justice and Security

Technology experts consulted (who reviewed Section 1)

1. Eric Postma, Professor of Artificial Intelligence, Tilburg University
2. Frank van Harmelen, Professor of Knowledge Representation & Reasoning, VU University Amsterdam

The project team would also like to express its deep gratitude to Joost Gerritsen for his legal fact-finding and checking of Sections 3 and 4.

Sources

Abid, A., Farooqi, M., & Zou, J. (2021). *Persistent Anti-Muslim Bias in Large Language Models* (arXiv:2101.05783). arXiv. https://doi.org/10.48550/arXiv.2101.05783

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, *28*(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Al-Hawawreh, M., Aljuhani, A., & Jararweh, Y. (2023). Chatgpt for cybersecurity: practical applications, challenges, and future directions. *Cluster Computing*, *26*(6), 3421–3436. https://doi.org/10.1007/s10586-023-04124-5

Alshurafat, H. (2023). *The usefulness and challenges of chatbots for accounting professionals: application on ChatGPT* (SSRN Scholarly Paper No. 4345921). https://doi.org/10.2139/ssrn.4345921

Ananthaswamy, A. (2023). In AI, is bigger always better? *Nature*, *615*(7951), 202–205. https://doi.org/10.1038/d41586-023-00641-w

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). *Frontier AI regulation. Managing emerging risks to public safety* (arXiv:2307.03718). arXiv. https://doi.org/10.48550/arXiv.2307.03718

Andersen, R. (2023, 24 July). Does Sam Altman know what he's creating? The OpenAI CEO's ambitious, ingenious, terrifying quest to create a new form of intelligence. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2023/09/sam-altman-openai-chatgpt-gpt-4/674764/

Anderson, C. (2008, 23 June). The end of theory. The data deluge makes the scientific method obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Anthropic. (2022). *Constitutional AI. Harmlessness from AI feedback*. https://www.anthropic.com/index/constitutional-ai-harmlessness-from-ai-feedback

Apple. (n.d.). *Replika. Virtual AI companion on the App Store.* Retrieved November 14, 2023, from https://apps.apple.com/us/app/replika-virtual-ai-companion/id1158555867

Autor, D. (2022). *The labor market impacts of technological change. From unbridled enthusiasm to qualified optimism to vast uncertainty* (Working Paper No. 30074). National Bureau of Economic Research. https://doi.org/10.3386/w30074

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and Artificial Intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*. https://doi.org/10.1001/jamainternmed.2023.1838

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI. Harmlessness from AI feedback* (arXiv:2212.08073). arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative AI. A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260. https://doi.org/10.3390/fi15080260

Baughman, J. (2023). *China's ChatGPT war*. China Aerospace Studies Institute. https://www.airuniversity.af.edu/Portals/10/CASI/documents/Research/Cyber/2023-08-21%20China's%20ChatGPT%20War.pdf

Bell, G., Burgess, J., Thomas, J., & Sadiq, S. (2023). *Rapid response information report. Generative AI*. Australia's Chief Scientist. https://www.chiefscientist.gov.au/sites/default/files/2023 06/Rapid%20Response%20Information%20Report%20-%20Generative%20AI%20v1_1.pdf

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020). Climbing towards NLU. On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bertuzzi, L. (2023a, 17 October). *AI Act. EU countries headed to tiered approach on foundation models amid broader compromise*. EURACTIV. https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-countries-headed-to-tiered-approach-on-foundation-models-amid-broader-compromise/

Bertuzzi, L. (2023b, 25 October). EU policymakers enter the last mile for Artificial Intelligence rulebook. *EURACTIV*. https://www.euractiv.com/section/artificial-intelligence/news/eu-policymakers-enter-the-last-mile-for-artificial-intelligence-rulebook/

Bertuzzi, L. (2023c, 10 November). *EU's AI Act negotiations hit the brakes over foundation models*. Www.Euractiv.Com. https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/

Bianchi, F., & Hovy, D. (2021). On the gap between adoption and understanding in NLP. *Findings of the Association for Computational Linguistics 2021*, 3895–901. https://doi.org/10.18653/v1/2021.findings-acl.340

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, *5*(5), 277–280. https://doi.org/10.1038/s42254-023-00581-4

Blodgett, S. L., & Madaio, M. (2021). *Risks of AI foundation models in education* (arXiv:2110.10024). arXiv. https://doi.org/10.48550/arXiv.2110.10024

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the opportunities and risks of foundation models* (arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258

Bowman, S. R. (2023). *Eight things to know about Large Language Models* (arXiv:2304.00612). arXiv. https://doi.org/10.48550/arXiv.2304.00612

Bozeman, B. (2007). *Public values and public interest. Counterbalancing economic individualism*. Georgetown University Press. https://www.jstor.org/stable/j.ctt2tt37c

Bradford, A. (2020). *The Brussels effect. How the European Union rules the world*. Oxford University Press. https://scholarship.law.columbia.edu/books/232

Brewster, J., Wang, M., & Palmer, C. (2023, 24 August). Plagiarism-bot? How low-quality websites are using ai to deceptively rewrite content from mainstream news outlets. *NewsGuard*. https://www.newsguardtech.com/misinformation-monitor/august-2023

Bronzwaer, S. (2023, 9 January). Van reclameteksten schrijven tot computercode checken. Waar wordt ChatGPT al voor gebruikt? *NRC*. https://www.nrc.nl/nieuws/2023/01/09/van-reclameteksten-schrijven-tot-computercode-checken-waar-wordt-chatgpt-al-voor-gebruikt-a4153682

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bruijn, H. D., & Dicke, W. (2006). Strategies for safeguarding public values in liberalized utility sectors. *Public Administration*, *84*(3), 717–735. https://doi.org/10.1111/j.1467-9299.2006.00609.x

Brynjolfsson, E. (2022, 12 January). The Turing trap. The promise & peril of human-like artificial intelligence. *Stanford Digital Economy Lab*. https://digitaleconomy.stanford.edu/news/the-turing-trap-the-promise-peril-of-human-like-artificial-intelligence/

Brynjolfsson, E., Li, D., & Raymond, L. (2023). *Generative AI at work* (SSRN Scholarly Paper No. 4426942). https://papers.ssrn.com/abstract=4426942

Callaway, E. (2022). Scientists are using AI to dream up revolutionary new proteins. *Nature*, *609*(7928), 661–662. https://doi.org/10.1038/d41586-022-02947-7

Cardon, P., Fleischmann, C., Aritz, J., Logemann, M., & Heidewald, J. (2023). The challenges and opportunities of AI-assisted Writing. Developing AI literacy for the AI age. *Business and Professional Communication Quarterly*, *86*(3), 257–95. https://doi.org/10.1177/23294906231176517

Chen, Z., Qing, J., Xiang, T., Yue, W. L., & Zhou, J. H. (2023). *Seeing beyond the brain. Conditional diffusion model with sparse masked modeling for vision decoding* (arXiv:2211.06956). arXiv. https://doi.org/10.48550/arXiv.2211.06956

Chohan, U. W. (2023). *Generative AI, ChatGPT, and the future of jobs* (SSRN Scholarly Paper No. 4411068). https://doi.org/10.2139/ssrn.4411068

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). *ChatGPT goes to law school* (SSRN Scholarly Paper No. 4335905). https://doi.org/10.2139/ssrn.4335905

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). *Economic potential of generative AI. The next productivity*

frontier. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

College voor de Rechten van de Mens. (2021). *Discriminatie door risicoprofielen. Een mensenrechtelijk toetsingskader*. https://publicaties.mensenrechten.nl/publicatie/61a734e65d726f72c45f9dce

Coscarelli, J. (2023, 19 April). An A.I. hit of fake 'Drake' and 'The Weeknd' rattles the music world. *The New York Times*. https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html

Cyber Security Raad. (2021). *Nederlandse digitale autonomie en cybersecurity. Hoe verminderen we onze digitale afhankelijkheden met behoud van een open economie?* (Beleidsnota No. 3; CSR Advies 2021). Ministerie van Justitie en Veiligheid. https://www.cybersecurityraad.nl/documenten/adviezen/2021/05/14/csr-advies-nederlandse-digitale-autonomie-en-cybersecurity---csr-advies-2021-nr.-3

Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. *AI & Society*, *34*(1), 129–136. https://doi.org/10.1007/s00146-017-0773-9

Danaher, J. (2020). Robot betrayal. A guide to the ethics of robotic deception. *Ethics and Information Technology*, *22*(2), 117–128. https://doi.org/10.1007/s10676-19-09520-3

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA. Efficient finetuning of quantized LLMs* (arXiv:2305.14314; Version 1). arXiv. http://arxiv.org/abs/2305.14314

Digitale Overheid. (2023, 6 November). *Nederland bouwt eigen open taalmodel GPT-NL*. https://www.digitaleoverheid.nl/nieuws/nederland-bouwt-eigen-open taalmodel-gpt-nl/

Doorenbosch, T. (2023, 1 September). ChatGPT opent nieuwe wereld voor scriptkiddies en meer kwaadwillenden. *AG Connect*. https://www.agconnect.nl/artikel/chatgpt-opent-nieuwe-wereld-voor-scriptkiddies-en-meer-kwaadwillenden

Edwards, B. (2023, 15 March). OpenAI checked to see whether GPT-4 could take over the world. *Ars Technica*. https://arstechnica.com/information-technology/2023/03/openai-checked-to-see-whether-gpt-4-could-take-over-the world/

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs. An early look at the labor market impact potential of Large Language Models* (arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130

Epstein, Z., Hertzmann, A., Herman, L., Mahari, R., Frank, M. R., Groh, M., Schroeder, H., Smith, A., Akten, M., Fjeld, J., Farid, H., Leach, N., Pentland, A., & Russakovsky, O. (2023). Art and the science of generative AI. A deeper dive. *Science*, *380*(6650), 1110–1111. https://doi.org/10.1126/science.adh4451

Est, R. van. (2020). Technologisch burgerschap als dé democratische uitdaging van de eenentwintigste eeuw. *Christen Democratische Verkenningen*, *2016*(3). https://www.tijdschriftcdv.nl/inhoud/tijdschrift_artikel

European Commission. (2020). *Commission Work Programme 2020*. https://eur-lex.europa.eu/legal content/EN/TXT/HTML/?uri=CELEX:52020DC0037&from=ES

European Commission High Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament. (2022). *EU strategic autonomy 2013-2023. From concept to capacity*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733589/EPRS_BRI(022)733589_EN.pdf

European Parliamentary Research Service, Scientific Foresight Unit, & Panel for the Future of Science and Technology. (2021). *Tackling deepfakes in European policy European Parliament*. (auteurs: Van Huijstee, M., Van Boheemen, P., Das, D., Nierling, L.,Jahnel, J., Karaboga & M., Fatun, M.). https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)690039

European Union External Action. (2020, 8 June). *For a united, resilient and sovereign Europe (with Thierry Breton)*. https://www.eeas.europa.eu/eeas/united-resilient-and-sovereign-europe-thierry-breton_und_en

Europol Innovation Lab. (2023). *ChatGPT. The impact of Large Language Models on Law Enforcement*. https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement#downloads

Fagone, J. (2021, 23 July). *He couldn't get over his fiancee's death. So he brought her back as an A.I. chatbot*. The San Francisco Chronicle. https://www.sfchronicle.com/projects/2021/jessica-simulation-artificial-intelligence/

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT. Implications for educational practice and research. *Innovations in Education and Teaching International*, 1–15. https://doi.org/10.1080/14703297.2023.2195846

Felsenthal, E., & Perrigo, B. (2023, 21 June). OpenAI CEO Sam Altman is pushing past doubts on Artificial Intelligence. *TIME*. https://time.com/collection/time100-companies-2023/6284870/openai-disrupters/

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). *From pretraining data to language models to downstream tasks. Tracking the trails of political biases leading to unfair NLP Models* (arXiv:2305.08283). arXiv. https://doi.org/10.48550/arXiv.2305.08283

Field, H. (2023, 29 June). The first fully A.I.-generated drug enters clinical trials in human patients. *CNBC*. https://www.cnbc.com/2023/06/29/ai-generated-drug-begins-clinical-trials-in-human-patients.html

Floridi, L. (2023). AI as agency without intelligence. On ChatGPT, Large Language Models, and other generative models. *Philosophy & Technology*, *36*(1), 15. https://doi.org/10.1007/s13347-023-00621-y

Geng, J., Huang, D., & De la Torre, F. (2022). *DensePose from WiFi* (arXiv:2301.00250). arXiv. https://doi.org/10.48550/arXiv.2301.00250

Gerbrandy, A., & Phoa, P. (2022). The Power of Big Tech Corporations as Modern Bigness and a Vocabulary for Shaping Competition Law as Counter-Power. In H. Brouwer, M. Bennett, & R. Claassen, *Wealth and Power* (1st ed., pp. 166–85). Routledge. https://doi.org/10.4324/9781003173632-11

Gmyrek, P., Berg, J., & Bescond, D. (2023). *Generative AI and jobs. A global analysis of potential effects on job quantity and quality*. International Labour Organization. https://www.ilo.org/global/publications/working-papers/WCMS_890761/lang--en/index.htm

Goujard, C., & Cerulus, L. (2021, 26 October). Inside Gaia-X. How chaos and infighting are killing Europe's grand cloud project. *POLITICO*. https://www.politico.eu/article/chaos-and-infighting-are-killing-europes-grand-cloud-project/

Gownder, J. P., & O'Grady, M. (2023). *Forrester's 2023 generative AI jobs impact forecast, US*. Forrester. https://www.forrester.com/report/forresters-2023-generative-ai-jobs-impact-forecast-us/RES179790

Grünebaum, A., Chervenak, J., Pollet, S. L., Katz, A., & Chervenak, F. A. (2023). The exciting potential for ChatGPT in obstetrics and gynecology. *American Journal of Obstetrics and Gynecology*, *228*(6), 696–705. https://doi.org/10.1016/j.ajog.2023.03.009

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT. Impact of generative AI in cybersecurity and privacy. *IEEE Access*, *11*, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

Gurnee, W., & Tegmark, M. (2023). *Language models represent space and time* (arXiv:2310.02207). arXiv. https://doi.org/10.48550/arXiv.2310.02207

Hacker, P., Engel, A., & Mauer, M. (2023). *Regulating ChatGPT and other Large Generative AI Models* (arXiv:2302.02337). arXiv. https://doi.org/10.48550/arXiv.2302.02337

Haggart, B. (2023, 23 January). ChatGPT strikes at the heart of the scientific world view. *Centre for International Governance Innovation*. https://www.cigionline.org/articles/chatgpt-strikes-at-the-heart-of-the-scientific-world-view/

Hao, K., & Seetharaman, D. (2023, 24 July). Cleaning up ChatGPT takes heavy toll on human workers. *Wall Street Journal*. https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483

Harrer, S. (2023). Attention is not all you need. The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, *90*. https://doi.org/10.1016/j.ebiom.2023.104512

Hartmann, J., Schwenzow, J., & Witte, M. (2023). *The political ideology of conversational AI. Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation* (arXiv:2301.01768). arXiv. https://doi.org/10.48550/arXiv.2301.01768

Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1682

Hugenholtz, P. B., & Quintais, J. P. (2021). Copyright and artificial creation. Does EU copyright law protect AI-assisted output? *International Review of Intellectual Property and Competition Law*, *52*, 1190–1216. https://doi.org/10.1007/s40319-021-01115-0

Ingram, D. (2023, 14 January). AI Chat used by mental health tech company in experiment on real users. *NBC News*. https://www.nbcnews.com/tech/internet/chatgpt-ai-experiment-mental-health-tech-app-koko-rcna65110

Jak, N., & Lokin, M. (2023, 1 May). Big Tech als publiek belang vraagt om publieke waarborgen. *iBestuur*. https://ibestuur.nl/artikel/big-tech-als-publiek-belang-vraagt-om-publieke-waarborgen/

Jakkal, V. (2023, 19 October). *Microsoft Security Copilot Early Access Program is now available*. Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2023/10/19/microsoft-security-copilot-early-access-program-harnessing-generative-ai-to-empower-security-teams/

Janssen, B. V., Kazemier, G., & Besselink, M. G. (2023). The use of ChatGPT and other large language models in surgical science. *BJS Open*, *7*(2), zrad032. https://doi.org/10.1093/bjsopen/zrad032

Jeon, J., & Lee, S. (2023). Large language models in education. A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-11834-1

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and applications of Large Language Models* (arXiv:2307.10169). arXiv. https://doi.org/10.48550/arXiv.2307.10169

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* (arXiv:2001.08361). arXiv. https://doi.org/10.48550/arXiv.2001.08361

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kelsey, T. (2023a). *Dream recording through non-invasive brain-machine interfaces and generative AI-assisted multimodal software* (arXiv:2304.09858). arXiv. https://doi.org/10.48550/arXiv.2304.09858

Kelsey, T. (2023b, 6 May). *Brain Interface > GenAI > Recording Dreams?* OpenAI Developer Forum. https://community.openai.com/t/brain-interface-genai-recording-dreams/194795

Kenniscentrum voor beleid en regelgeving. (n.d.). *Beleidskompas*. Retrieved November 14, 2023, from https://www.kcbr.nl/beleid-en-regelgeving-ontwikkelen/beleidskompas

Kerssens, N., & van Dijck, J. (2023). Transgressing local, national, global spheres: the blackboxed dynamics of platformization and infrastructuralization of primary

education. *Information, Communication & Society*, *0*(0), 1–17. https://doi.org/10.1080/1369118X.2023.2257293

Knight, W. (2023). The last AI boom didn't kill jobs. Feel better? *Wired*. https://www.wired.com/story/fast-forward-the-last-ai-boom-didnt-kill-jobs/

Koncz, A. (2023, 7 February). 5 AI tools that could help huild a medical practice. *The Medical Futurist*. https://medicalfuturist.com/5-ai-tools-that-could-help-build-a-medical-practice/

Koning, P. (2013). Activerend arbeidsmarktbeleid. Een beknopte handleiding. *TPEdigitaal*, *7*(2), 60–66. https://www.tpedigitaal.nl/sites/default/files/bestand/activerend_arbeidsmarktbeleid_een_beknopte_handleiding.pdf

Korteweg, N. (2023, 2 May). ChatGPT verslaat artsen bij het geven van antwoord op medische vragen. *NRC*. https://www.nrc.nl/nieuws/2023/05/02/chatgpt-verslaat-artsen-bij-het-geven-van-antwoord-op-medische-vragen-a4163538

Kreps, S., & Kriner, D. (2023, 21 March). How generative AI impacts democratic engagement. *Brookings*. https://www.brookings.edu/techstream/how-generative-ai-impacts-democratic-engagement/

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE. Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Kutterer, C. (2024) Regulating foundation models in the AI Act: from 'high' to 'systemic' risk. https://ai-regulation.com/regulating-foundation-models-in-the-ai-act-from-high-to-systemic-risk/

La Moncloa. (2021, 14 July). *The government adopts the Digital Rights Charter to articulate a reference framework to guarantee citizens rights in the new digital age*. La Moncloa. https://www.lamoncloa.gob.es/lang/en/gobierno/news/Paginas/2021/20210713_rights-charter.aspx

Lang, O., Yaya-Stupp, D., Traynis, I., Cole-Lewis, H., Bennett, C. R., Lyles, C., Lau, C., Semturs, C., Webster, D. R., Corrado, G. S., Hassidim, A., Matias, Y., Liu, Y., Hammel, N., & Babenko, B. (2023). *Using generative AI to investigate medical imagery models and datasets* (arXiv:2306.00985). arXiv. https://doi.org/10.48550/arXiv.2306.00985

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making AI less "thirsty". Uncovering and addressing the secret water footprint of AI models* (arXiv:2304.03271). arXiv. http://arxiv.org/abs/2304.03271

Lima, C., & Zakrzewski, C. (2023, 30 October). Biden signs AI executive order, the most ambitious U.S. regulation yet. *The Washington Post*. https://www.washingtonpost.com/technology/2023/10/30/biden-artificial-intelligence-executive-order/

Lo, C. K. (2023). What Is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Lodge, J. M., Thompson, K., & Corrin, L. (2023). Mapping out a research agenda for generative artificial intelligence in tertiary education. *Australasian Journal of Educational Technology*, *39*(1), 1–8. https://doi.org/10.14742/ajet.8695

Loh, E. (2023). ChatGPT and generative AI chatbots. Challenges and opportunities for science, medicine and medical leaders. *BMJ Leader*, leader. https://doi.org/10.1136/leader-2023-000797

Lorach, H., Galvez, A., Spagnolo, V., Martel, F., Karakas, S., Intering, N., Vat, M., Faivre, O., Harte, C., Komi, S., Ravier, J., Collin, T., Coquoz, L., Sakr, I., Baaklini, E., Hernandez-Charpak, S. D., Dumont, G., Buschman, R., Buse, N., … Courtine, G. (2023). Walking naturally after spinal cord injury using a brain–spine interface. *Nature*, *618*(7963), 126–133. https://doi.org/10.1038/s41586-23-06094-5

Lorenz, P., Perset, K., & Berryhill, J. (2023). *Initial policy considerations for generative artificial intelligence*. OECD. https://doi.org/10.1787/fae2d1e6-en

Malinka, K., Perešíni, M., Firc, A., Hujňák, O., & Januš, F. (2023). *On the educational impact of ChatGPT. Is Artificial Intelligence ready to obtain a university degree?* (arXiv:2303.11146). arXiv. https://doi.org/10.48550/arXiv.2303.11146

Marcus, G. (2020). *The next decade in AI. Four steps towards robust Artificial Intelligence* (arXiv:2002.06177). arXiv. https://doi.org/10.48550/arXiv.2002.06177

Marr, B. (2023, 2 March). Revolutionizing healthcare. The top 14 uses of ChatGPT in medicine and wellness. *Forbes*. https://www.forbes.com/sites/bernardmarr/2023/03/02/revolutionizing-healthcare-the-top-14-uses-of-chatgpt-in-medicine-and-wellness/

Maximillian Schrems v Data Protection Commissioner, Case C-362/14 (ECJ October 6, 2015). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62014CJ0362

Meta. (2023, 27 September). *Introducing new AI experiences across our family of apps and devices*. https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2023, 24 January). *Antwoorden op Kamervragen over Rijksbreed cloudbeleid 2022* [Brief]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/brieven/2023/01/24/antwoorden-op-kamervragen-over-rijksbreed-cloudbeleid-2022

Ministerie van Economische Zaken en Klimaat. (2019). *Strategisch Actieplan voor Artificiële Intelligentie*. https://www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie

Mittelstadt, B., Wachter, S., & Russell, C. (2023). The unfairness of fair machine learning. Levelling down and strict egalitarianism by default. *Michigan Technology Law Review*. https://ora.ox.ac.uk/objects/uuid:09debd0c-7f13-4042-a37e-76381a389362

Mok, A. (2023, 20 April). *ChatGPT could cost over $700,000 per day to operate. Microsoft is reportedly trying to make it cheaper*. Business Insider. https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4

Muench, S., Stoermer, E., Jensen, K., Asikainen, T., Salvi, M., & Scapolo, F. (2022). *Towards a green & digital future*. Publications Office of the European Union. https://doi.org/10.2760/977331

Nabatchi, T. (2018). Public values frames in administration and governance. *Perspectives on Public Management and Governance*, *1*(1), 59–72. https://doi.org/10.1093/ppmgov/gvx009

Nature editorial. (2023). For chemists, the AI revolution has yet to happen. *Nature*, *617*(7961), 438–438. https://doi.org/10.1038/d41586-023-01612-x

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A*, *376*, 1–14. https://doi.org/10.1098/rsta.2018.0089

Nógrádi, B., Polgár, T. F., Meszlényi, V., Kádár, Z., Hertelendy, P., Csáti, A., Szpisjak, L., Halmi, D., Erdélyi-Furka, B., Tóth, M., Molnár, F., Tóth, D., Bősze, Z., Klivényi, P., Siklós, L., & Patai, R. (2023). *ChatGPT M.D.. Is there any room for generative AI in neurology and other medical areas?* (SSRN Scholarly Paper No. 4372965). https://doi.org/10.2139/ssrn.4372965

Norwegian Consumer Council. (2023). *Ghost in the machine. Adressing the consumer harms of generative AI*. https://storage02.forbrukerradet.no/media/2023/06/generative-ai-rapport-2023.pdf

NOS. (2023, 24 September). *Met AI gemaakte naaktfoto's van tientallen meisjes uit Spaans dorp verspreid*. https://nos.nl/artikel/2491701-met-ai-gemaakte-naaktfoto-s-van-tientallen-meisjes-uit-spaans-dorp-verspreid

NOS Nieuws. (2023, 19 April). *Een chatbot die in het echt wil afspreken? Ook Snapchat vindt het wat ver gaan*. https://nos.nl/artikel/2472026-een-chatbot-die-in-het-echt-wil-afspreken-ook-snapchat-vindt-het-wat-ver-gaan

Noy, S., & Zhang, W. (2023). *Experimental evidence on the productivity effects of generative Artificial Intelligence* (SSRN Scholarly Paper No. 4375283). https://doi.org/10.2139/ssrn.4375283

OECD. (2019). *OECD Skills Outlook 2019. Thriving in a Digital World*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/education/oecd-skills-outlook-2019_df80bc12-en

OECD. (2023a). *OECD Employment Outlook 2023. Artificial Intelligence and the Labour Market*. https://www.oecd-ilibrary.org/employment/oecd-employment-outlook-2023_08785bba-en

OECD. (2023b). *AI Language Models. Technological, socio-economic and policy considerations.* (No. 352). OECD Publishing. https://www.oecd-ilibrary.org/docserver/13d38f92-en.pdf?expires=1682422393&id=id&accname=guest&checksum=9528D65C1CA15C4A9A93A1689F2F507E

OECD.AI Policy Observatory. (n.d.). *OECD AI Principles overview*. Retrieved November 15, 2023, from https://oecd.ai/en/ai-principles

Onderwijsraad. (2022). *Inzet van intelligente technologie* [Publicatie]. Ministerie van Onderwijs, Cultuur en Wetenschap. https://www.onderwijsraad.nl/publicaties/adviezen/2022/09/28/inzet-van-intelligente-technologie

OpenAI. (2023, 25 September). *ChatGPT can now see, hear, and speak*. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak?utm_source=substack&utm_medium=email

Pandey, S., & Sharma, S. (2023). A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. *Healthcare Analytics*, *3*, 100198. https://doi.org/10.1016/j.health.2023.100198

Passchier, R. (2021). *Artificiële intelligentie en de rechtsstaat Over verschuivende overheidsmacht, Big Tech en de noodzaak van constitutioneel onderhoud*. Boom.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon emissions and Large Neural Network training* (arXiv:2104.10350). arXiv. https://doi.org/10.48550/arXiv.2104.10350

Perigo, B. (2023, 18 January). The $2 Per hour workers who made ChatGPT safer. *TIME*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Philippidis, A. (2023, 10 February). Insilico gains FDA's first orphan drug designation for AI candidate. *Genetic Engineering and Biotechnology News*. https://www.genengnews.com/news/insilico-gains-fdas-first-orphan-drug-designation-for-ai-candidate/

Qi, X., Zhu, Z., & Wu, B. (2023). The promise and peril of ChatGPT in geriatric nursing education. What we know and do not know. *Journal of Translational Medicine*, *3*(2), 100136. https://doi.org/10.1016/j.ahr.2023.100136

Quekel, S., & Hoijtink, D. (2023, 13 July). 'Briljante' Google-chatbot Bard blundert volop. 'Hugo de Jonge in race voor VVD.' *Algemeen Dagblad*. https://www.ad.nl/binnenland/briljante-google-chatbot-bard-blundert-volop-hugo-de-jonge-in-race-voor-vvd~aebf0eb3/

Raad voor de leefomgeving en infrastructuur. (2021). *Digitaal duurzaam*. https://www.rli.nl/publicaties/2021/advies/digitaal-duurzaam

Raad voor het Openbaar Bestuur. (2021). *Sturen of gestuurd worden? Over de legitimiteit van sturen met data* [Publicatie]. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. https://www.raadopenbaarbestuur.nl/documenten/publicaties/2021/05/25/advies-sturen-of-gestuurd-worden

Rahman, M., & Watanobe, Y. (2023). ChatGPT for Education and Research: Opportunities, Threats, and Strategies. *Applied Sciences*, *13*(9), 5783. https://doi.org/10.3390/app13095783

Raso, F., Hilligloss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). *Artificial Intelligence & human rights. Opportunities & risks*. Berkman Klein Center for Internet & Society at Harvard University. https://dash.harvard.edu/bitstream/handle/1/38021439/2018-

09_AIHumanRights.pdf
Rathenau Instituut. (2015). *Werken aan de robotsamenleving. Visies en inzichten uit de wetenschap over de relatie technologie en werkgelegenheid*. (auteurs: Kool, L. & Van Est, R.). https://www.rathenau.nl/nl/digitalisering/werken-aan-de-robotsamenleving
Rathenau Instituut. (2017). *Opwaarderen. Borgen van publieke waarden in de digitale samenleving*. (auteurs: Kool, L. & Van Est, R.). https://www.rathenau.nl/nl/digitalisering/opwaarderen
Rathenau Instituut. (2019). *Verantwoord virtueel. Bescherm consumenten in virtual reality*. (auteurs: Kool, L. & Van Est, R.). https://www.rathenau.nl/nl/digitalisering/verantwoord-virtueel
Rathenau Instituut. (2020a). *Hoor wie het zegt. Hanvatten voor het verantwoorde gebruik van spraaktechnologie*. (auteurs: Hamer, J. & Kool, L.). https://www.rathenau.nl/nl/digitalisering/hoor-wie-het-zegt
Rathenau Instituut. (2020b). *Nep echt. Verrijk de wereld met augmented reality*. (auteur: Van Est, R.). https://www.rathenau.nl/nl/digitalisering/nep-echt
Rathenau Instituut. (2020c). *Rathenau Manifest. Stel nu 10 ontwerpeisen aan de digitale samenleving van morgen*. https://www.rathenau.nl/nl/digitalisering/rathenau-manifest-stel-nu-10-ontwerpeisen-aan-de-digitale-samenleving-van-morgen
Rathenau Instituut. (2020d). *Werken op waarde geschat. Grenzen aan digitale monitoring op de werkvloer door middel van data, algoritmen en AI*. (auteurs: Das, D. & Kool, L.). https://www.rathenau.nl/nl/digitalisering/werken-op-waarde-geschat
Rathenau Instituut. (2021). *De stand van digitaal Nederland. Naar zeggenschap en vertrouwen in de digitale samenleving*. (auteurs: Hamer, J. & Kool, L.). https://www.rathenau.nl/nl/digitalisering/de-stand-van-digitaal-nederland
Rathenau Instituut. (2022a). *Beter beslissen over datacentra. De noodzaak van een breed publiek perspectief op de digitale infrastructuur*. (auteur: Van Est, R.). https://www.rathenau.nl/nl/digitalisering/beter-beslissen-over-datacentra
Rathenau Instituut. (2022b). *Naar hoogwaardig digitaal onderwijs*. (auteurs: Karstens, B. & Kool, L.). https://www.rathenau.nl/nl/digitalisering/naar-hoogwaardig-digitaal-onderwijs
Rathenau Instituut. (2022c). *Non-discriminatie bij algoritmes*. https://www.rathenau.nl/nl/berichten-aan-het-parlement/non-discriminatie-bij-algoritmes
Rathenau Instituut. (2023a). *Gezondheidstechnologie speciaal voor vrouwen. FemTech en de gezondheidskloof*. (auteurs: Pison, I. & Edelenbosch, R.). https://www.rathenau.nl/nl/gezondheid/gezondheidstechnologie-speciaal-voor-vrouwen
Rathenau Instituut. (2023b). *Immersieve technologieën*. (auteurs: Ex, L., Nieuwenhuizen, W., Hijstek, B., Roolvink, S. & Huijstee, M.). https://www.rathenau.nl/nl/digitalisering/immersieve-technologieen
Ravindran, A. (2023, 8 May). *Brain-computer interfaces and AI language models. A new frontier*. LinkedIn. https://www.linkedin.com/pulse/brain-computer-interfaces-ai-

language-models-new-aniruddh-ravindran/
Riemens, R., Nast, C., Pelzer, P., & van den Hurk, M. (2021). An assessment framework for safeguarding public values on mobility platforms. *Urban Transformations*, *3*(1), 7. https://doi.org/10.1186/s42854-021-00023-3
Rijksoverheid. (2022a, 29 August). *Werken 'in de cloud' wordt mogelijk voor Rijksoverheid* [Nieuwsbericht]. https://www.rijksoverheid.nl/actueel/nieuws/2022/08/29/werken-in-de-cloud-wordt-mogelijk-voor-rijksoverheid
Rijksoverheid. (2022b). *Werkagenda Waardengedrageven Digitaliseren*. http://www.digitaleoverheid.nl/werkagenda
Rooij, I. V. (2023, 14 January). Stop feeding the hype and start resisting. *Iris van Rooij*. https://irisvanrooijcogsci.com/2023/01/14/stop-feeding-the-hype-and-start-resisting/
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots. Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, *6*(1), Article 1. https://doi.org/10.37074/jalt.2023.6.1.23
Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education. Quick start guide*. UNESCO International Institute for Higher Education in Latin America and the Caribbean. https://unesdoc.unesco.org/ark:/48223/pf0000385146
Sadeghi, M., Arvanitis, L., Padovese, V., Pozzi, G., Badilini, S., Vercellone, C., Roache, M., Wang, M., Brewster, J., Huet, N., Schimmel, B., Fishman Z., Pfaller, L. & Adams, L. (2024) Tracking AI-enabled Misinformation: 739 'Unreliable AI-Generated News' Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools. *Newsguard*, https://www.newsguardtech.com/special-reports/ai-tracking-center/
Schaeffer, R., Miranda, B., & Koyejo, S. (2023). *Are emergent abilities of large language models a mirage?* (arXiv:2304.15004). arXiv. http://arxiv.org/abs/2304.15004
Sharon, T. (2016). The Googlization of health research: from disruptive innovation to disruptive ethics. *Personalized Medicine*, *13*(6), 563–574. https://doi.org/10.2217/pme-2016-0057
Sharon, T., & Gellert, R. (2023). Regulating Big Tech expansionism? Sphere transgressions and the limits of Europe's digital regulatory strategy. *Information, Communication & Society*, *0*(0), 1–18. https://doi.org/10.1080/1369118X.2023.2246526
Simon, Julien. (n.d.). *Large Language Models: A New Moore's Law?* Retrieved September 20, 2023, from https://huggingface.co/blog/large-language-models
Sohail, S. S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., Atalla, S., & Mansoor, W. (2023). *The future of GPT. A taxonomy of existing ChatGPT*

research, current challenges, and possible future directions (SSRN Scholarly Paper No. 4413921). https://doi.org/10.2139/ssrn.4413921

Solaiman, I. (2023, 24 May). Generative AI systems aren't just open or closed source. *Wired*. https://www.wired.com/story/generative-ai-systems-arent-just-open-or-closed-source/

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S., Daumé, I., Dodge, J., Evans, E., Hooker, S., Jernite, Y., Luccioni, A., Lusoli, A., Mitchell, M., Newman, J., Png, M.-T., Strait, A., & Vassilev, A. (2023a). *Evaluating the Social Impact of Generative AI Systems in Systems and Society*.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Daumé, H., Dodge, J., Evans, E., Hooker, S., Jernite, Y., Luccioni, A. S., Lusoli, A., Mitchell, M., Newman, J., Png, M.-T., Strait, A., & Vassilev, A. (2023b). *Evaluating the Social Impact of Generative AI Systems in Systems and Society* (arXiv:2306.05949). arXiv. https://doi.org/10.48550/arXiv.2306.05949

Sparnaaij, K., Van Eekeren, P., & Vasseur, J. (2023). *AI Monitor Ziekenhuizen. Editie 2023*. M&I Partners. https://mxi.nl/uploads/files/publication/ai-monitor-ziekenhuizen-2023.pdf

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: many scientists disapprove. *Nature*, *613*(7945), 620–621. https://doi.org/10.1038/d41586-023-00107-z

Sweeney, C. (2021, 20 November). Artificial intelligence. Good or bad for public health? *World Economic Forum*. https://www.weforum.org/agenda/2021/11/artificial-intelligence-technology-public-health/

Takagi, Y., & Nishimoto, S. (2022). *High-resolution image reconstruction with latent diffusion models from human brain activity* (p. 2022.11.18.517004). bioRxiv. https://doi.org/10.1101/2022.11.18.517004

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*(5), 858–866. https://doi.org/10.1038/s41593-023-01304-9

Turkle, S. (2015). *Reclaiming conversation. The power of talk in a digital age.* Penguin Press.

UNESCO. (2022). *Recommendation on the ethics of Artificial Intelligence*. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

UNESCO. (2023). *Guidance for generative AI in education and research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

UNESCO International Bioethics Committee. (2022). *Ethical issues of neurotechnology*. https://unesdoc.unesco.org/ark:/48223/pf0000383559

United Nations. (2023). *Our common agenda policy brief 9. A new agenda for peace.* [Policy brief]. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf

Unlearn.AI. (n.d.). *AI-powered digital twins of individuals patients*. Retrieved August 2, 2023, from https://www.unlearn.ai/technology

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2023). AI in drug discovery. A wake-up call. *Drug Discovery Today*, *28*(1), 103410. https://doi.org/10.1016/j.drudis.2022.103410

U.S. Department of Defense. (2023, 10 August). *DOD announces establishment of generative AI task force*. https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/

Van Bekkum, M., & Zuiderveen Borgesius, F. (2023). De spanning tussen het non-discriminatierecht en het gegevensbeschermingsrecht. Heeft de AVG een nieuwe uitzondering nodig om discriminatie door kunstmatige intelligentie tegen te gaan? *Nederlands Juristenblad*, *27*, 2271–2280.

Van Buchem, M., Boosman, H., Bauer, M., Kant, I., Cammel, S., & Steyerberg, E. (2021). The digital scribe in clinical practice. A scoping review and research agenda. *Npj Digital Medicine*, *4*. https://doi.org/10.1038/s41746-021-00432-5

Van de Ven, L. (2023, 28 July). OM gaat maker deepfakes van Welmoed Sijtsma vervolgen. *NRC*. https://www.nrc.nl/nieuws/2023/07/28/om-gaat-maker-deepfakes-van-welmoed-sijtsma-vervolgen-a4170774

van den Berge, W., & ter Weel, B. (2015). *Baanpolarisatie in Nederland*. Centraal Planbureau. https://www.cpb.nl/publicatie/baanpolarisatie-in-nederland

van Dijck, J., Poell, T., & de Waal, M. (2018). *The Platform Society*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems*. Conference on Neural Information Processing Systems, Long Beach. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Véliz, C. (2023). Chatbots shouldn't use emojis. *Nature*, *615*(7952), 375. https://doi.org/10.1038/d41586-023-00758-y

Verdecchia, R., Sallou, J. & Cruz, L. (2023) A systematic review of green AI. *Wire 13(4),* https://doi.org/10.1002/widm.1507

Villasenor, J., & West, D. M. (2023, 10 April). Will generative AI kill jobs? *Brookings*. https://www.brookings.edu/blog/techtank/2023/04/10/will-generative-ai-kill-jobs/

Vipra, J., & Myers West, S. (2023). *Computational power and AI*. AI Now Institute. https://ainowinstitute.org/publication/policy/compute-and-ai

Visser, D. (2023). Robotkunst en auteursrecht. *Nederlands Juristenblad*, *7*, 504–515. https://www.ipmc.nl/wp-content/uploads/2023/06/Robotkunst-en-auteursrecht-1.pdf

Vogt, M. (2023). Exploring chemical space. Generative models and their evaluation. *Artificial Intelligence in the Life Sciences*, *3*, 100064. https://doi.org/10.1016/j.ailsci.2023.100064

Vries, A. (2023) The growing energy footprint of artificial intelligence. *Joule, 7(10)*, 2191 0 2194. https://doi.org/10.1016/j.joule.2023.09.004

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated. Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567. https://doi.org/10.1016/j.clsr.2021.105567

Wang, J., Mao, J., Wang, M., Le, X., & Wang, Y. (2023). Explore drug-like space with deep generative models. *Methods*, *210*, 52–59. https://doi.org/10.1016/j.ymeth.2023.01.004

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by Language Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Went, R., Kremer, M., & Knottnerus, A. (red.). (2015). *De robot de baas. De toekomst van werk in het tweede machinetijdperk (WRR-Verkenning nr. 31)*. Amsterdam University Press. https://www.wrr.nl/publicaties/verkenningen/2015/12/08/de-robot-de-baas

Wetenschappelijke Raad voor het Regeringsbeleid. (2021). *Opgave AI. De nieuwe systeemtechnologie* (Rapport No. 105). Ministerie van Algemene Zaken. https://www.wrr.nl/publicaties/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie

Whang, O. (2023a, 1 May). A.I. is getting better at mind-reading. *The New York Times*. https://www.nytimes.com/2023/05/01/science/ai-speech-language.html

Whang, O. (2023b, 24 May). Brain implants allow paralyzed man to walk using his thoughts. *The New York Times*. https://www.nytimes.com/2023/05/24/science/paralysis-brain-implants-ai.html

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., & Henderson, J. M. (2023). A high-performance speech neuroprosthesis. *Nature*, 1–6. https://doi.org/10.1038/s41586-023-06377-x

Wong, M. (2023, 2 October). Artists are losing the war against AI. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/10/openai-dall-e-3-artists-work/675519/

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). Weaponized AI for cyber attacks. *Psychology & Marketing*, *57*, 102722. https://doi.org/10.1016/j.jisa.2020.102722

Yasar, A. G., Chong, A., Dong, E., Gilbert, T. K., Hladikova, S., Maio, R., Mougan, C., Shen, X., Singh, S., Stoica, A.-A., Thais, S., & Zilka, M. (2023). *AI and the EU Digital Markets Act. Addressing the risks of bigness in generative AI* (arXiv:2308.02033). arXiv. https://doi.org/10.48550/arXiv.2308.02033

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A survey of Large Language Models* (arXiv:2303.18223). arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zubascu, F. (n.d.). EU and US hatch transatlantic plan to rein in ChatGPT. *Science|Business*. Retrieved June 19, 2023, from https://sciencebusiness.net/news/AI/eu-and-us-hatch-transatlantic-plan-rein-chatgpt

Authors
Jurriën Hamer, Linda Kool, Bo Hijstek, Quirine van Eeden and Djurre Das

Illustrations
Rathenau Instituut

Cover photo
Stock-Asso / Shutterstock

Please cite as:
Rathenau Instituut (2023). *Generatieve AI.* The Hague. Authors: Hamer, J., L. Kool, B. Hijstek, Q. van Eeden and D. Das

Contact details
Anna van Saksenlaan 51
Postbus 95366
2509 CJ Den Haag
070-342 15 42
info@rathenau.nl
www.rathenau.nl

The Rathenau Instituut stimulates public and political opinion forming on social aspects of science and technology. We perform research and organise debate relating to science, innovation and new technologies.

Rathenau Instituut