HELMHOLTZ

Spitzenforschung für
große Herausforderungen.

Recommendations for the use of artificial intelligence

Version 1.0

(version approved by the Helmholtz General Assembly on September 18, 2024)

Content

1. Preamble

Artificial Intelligence (AI), its many applications, and the potential risks associated with it are currently of concern to large segments of society. This is particularly due to the development and widespread use of generative AI systems for text, video, image or music creation (e.g. ChatGPT, Dall-E).

Even before the release of well-known generative AI systems, AI was integrated into the professional and private lives of many people. While navigation software (e.g. Google Maps) uses AI for route planning, streaming services (e.g. YouTube, Netflix) and sales platforms (e.g. Amazon) use AI for personalized recommendations based on user behavior and preferences. In the professional context, AI is widely used for spelling and grammar checking (e.g. Microsoft Word, Grammarly), analysis and data visualization (e.g. Microsoft Excel), or automatic background noise suppression (e.g. Zoom). But it was not until the widespread use of generative AI systems—large AI models that are trained on large amounts of data and then use that data to generate new content that is indistinguishable from human-generated content—that the enormous potential of AI for society as a whole became clear[1].

AI is a relevant research topic in the Helmholtz Association. Developments in this field are therefore of great importance. On the one hand, Helmholtz has large, self-hosted computing capacities (e.g., JUWELS Booster, HoReKa and soon JUPITER), and makes them available to researchers and (external) users for the development and application of (generative) AI systems. On the other hand, Helmholtz actively promotes the development of its own AI models for various research areas (e.g. Helmholtz Foundation Model Initiative (HFMI)). At the same time, Helmholtz researchers are developing new AI methods or applying AI technologies and systems to help develop solutions for the grand challenges of our time.

As Europe's largest research organization, Helmholtz is not only committed to helping shape the world of AI by developing its own AI systems, but also wants to encourage and enable employees in all areas of the association to use AI responsibly.

The aim of these recommendations is therefore to highlight the opportunities that AI systems offer, as well as potential risks, and recommendations for mitigating them. This should empower all employees of the Helmholtz Association to use and develop AI systems in an informed, reflective and responsible manner.

2. Explanation of key concepts

There is currently no universally accepted definition of AI, and the pace of progress makes it unlikely that a consensus will be reached soon. However, some characteristics that are commonly used to describe AI can be identified.

Artificial intelligence (AI) refers to the field of computer science that deals with the development of systems and algorithms that aim to solve tasks in the digital and physical world.

Traditional AI systems, such as rule-based or symbolic AI systems, rely on predetermined rules and algorithms to make decisions and solve problems. This type of AI uses logical inference and explicitly coded knowledge bases to accomplish tasks.

[1] See, for example, "Science in the age of AI" by the Royal Society.

Generative AI systems are a special form of artificial intelligence that transcends mere automation and classification. They produce new content (text, images, videos, music, etc.) by learning patterns and structures in large amounts of data in a training process and subsequently replicating them.

3. Opportunities created by AI

AI systems have the potential to facilitate many tasks in research, management and administration.

In research, artificial intelligence has been an integral part of the repertoire of methods for a long time. The generative AI systems developed recently promise dramatic simplifications of data science processes. Particularly in the exploratory phase of research, in which hypotheses are generated and tested, AI systems can help accelerate the knowledge-building process, for example when programming new tools. In addition, generative AI systems can simplify the application of complicated analysis algorithms, for example through natural language interfaces, and thus make them accessible to a broader base of researchers. But the automation of repetitive tasks can also create efficiencies, as with data preparation or report generation. This allows researchers to invest more time in creative and strategic activities, significantly increasing innovation and scientific productivity.

(Generative) AI systems also provide many administrative and management applications. From chatbots that interactively enable access to knowledge in databases or cloud systems, to writing meeting notes, job advertisements or texts for a variety of other contexts—large language models can make work easier and enable more effective processes. Image-generating AI tools can help to create graphics and visualizations for presentations, websites or social media, freeing up capacity for users to work on other tasks.

The latest developments also enable the use of so-called multimodal AI systems, which can process and generate not only written text but also audio and video. Such systems can, for example, help with the creation of press releases, where not only text but also images and short videos are to be used.

Due to the rapid development in the field of (generative) AI, this list can only provide a snapshot of the possibilities, which will need to be regularly updated. Therefore, the related new challenges, uncertainties and risks (see 4) can currently only be outlined in broad lines. The societal embedding of AI systems is, to a large extent, still developing. The experience of scientists and science managers with the practical possibilities and limitations of AI systems are an important resource for social reflection and a significant instance of societal learning with regard to the future handling and design of such systems. It is the express goal of the Helmholtz Association to create the conditions for the association and its employees to be able to contribute to this comprehensively.

4. Risks and recommendations for their mitigation

AI systems, in particular generative AI systems, offer many possible applications. However, they also entail risks and challenges. Some of these are based on technical limitations of individual AI systems, while others are due to the (intentional or unintentional) misuse of AI systems. The following section highlights individual risks and recommendations for minimizing them.

In principle, users of AI systems are always responsible for

1. the content they provide to the AI system (input), and

2. the use of the content that the AI system produces (output).

AI systems are not moral or legal agents and cannot be held legally or morally responsible for products. In particular, AI systems cannot claim authorship and cannot be held accountable for misinformation, data protection violations, copyright infringement, or other legally or morally problematic information. Responsibility requires conscious decision-making ability and moral judgment, which AI systems do not possess. Therefore, it is the responsibility of the users to use AI systems and their results in such a way that moral standards and legal regulations are observed, since users may be liable for any violation.

4.1 Risk: Privacy violation and unintentional disclosure of information

(Generative) AI systems often store and/or use the information entered by users, for example to further train the AI. In this way, the information entered by users can reach both the service provider and other future users. Users of certain AI systems thus run the risk of passing on sensitive, confidential or personal data to others.

This can have various consequences:

- Potential violation of the General Data Protection Regulation[2] (GDPR), which protects personal data.
- Violation of the Trade Secrets Act[3] through the disclosure of confidential business information/secrets.
- Unpublished research results/data/projects and manuscripts are made accessible to third parties, so that they can be used and possibly published by competitors.

Recommended action for risk mitigation

- When using (external) AI systems, care should be taken to ensure that no sensitive data is passed on to the service providers. This includes
  - Personal data (see GDPR for details)
  - Trade secrets, confidential/strategic business information
  - Unpublished research results/data, scientific manuscripts, research and project proposals
- If AI systems are used nonetheless, care must be taken to ensure that the sensitive data is obfuscated so that it cannot be reconstructed or used by either the external service provider or other users.
- When sensitive data cannot be obfuscated because all information in the document is necessary for a meaningful evaluation, e.g., of scientific manuscripts or research and project

[2] Further information on the GDPR can be found here.
[3] Further information on the Act on the Protection of Trade Secrets can be found here.

proposals, evaluations and assessments are on principle to be carried out without the support of AI systems[4].

4.2 Risk: Violation of copyright and other intellectual property rights

AI systems may have been trained with copyrighted data or by using other intellectual property, some or all of which may be reflected in the generated results. If users use or distribute these results (images, texts, videos, films, etc.) without checking them, this leads to copyright infringement[5] or violation of other intellectual property rights. In this case, the users are liable and may be legally prosecuted, even in cases of unintentional infringement of copyrights and other intellectual property.

Recommended action for risk mitigation

- AI-generated content should always be checked for trademarks and other elements that are obviously protected by copyright, trademark law, or other laws (e.g., logos on generated images), before further processing. If protected elements are found in the results, the service in question and the content generated with it should not be used if at all possible.
- Preferably, established services should be used, that are based on open-source models, provide information about their training data, and support verification of the copyright status of their sources.
- When creating content with high visibility and reach (e.g., images for social media or scientific publications), users should always carefully consider the use of AI tools against the risk of unintentional copyright infringement.

4.3 Risk: Dissemination of false information and violation of (scientific) information integrity

When using AI-generated content, there is a risk of spreading misinformation and plagiarism. This is not compatible with the principle of (scientific) information integrity.

- **False information:** As described above, AI systems, and generative AI systems in particular, generate their results based on training data, i.e., on existing texts, data, images, videos, etc. In doing so, the AI system does not check whether the data used for training is factually correct, or whether the results are true or meaningful. Consequently, there is a risk that AI-generated content may contain false information, which could be (unintentionally) disseminated when it is used. Users are responsible for checking the output for accuracy, as they bear the responsibility for its continued use. Disclosure and documentation of the use of AI systems is crucial here.
- **"Hallucinations":** AI Systems sometimes assemble elements of their training data to new content in such a way, that the output is incorrect, or even nonsensical. Such

[4] See also the statement by the German Research Foundation (DFG)
[5] You can find more information on copyright here.

"hallucinations" are to be considered as a special kind of false information, generated by the AI system itself, without any basis in the underlying data or in reality. Therefore, users should never use results unchecked, and on principle critically review AI-generated products.

- **Plagiarism:** AI-generated content may contain plagiarized content, originating directly from the training data. If users reuse this content unchecked, they are liable for plagiarism. If users use AI-generated content and pass it off as their own intellectual property, and the generated content contains the intellectual property of others, they are plagiarizing. Users are responsible for the content they use. When further processing the generated content, authorship lies with the users, not with the AI tool.

Recommended action for risk mitigation

- Authorship can only lie with natural persons. It is these authors who, in their role as authors, take responsibility for content and ensure its quality and accuracy. Content must therefore always be thoroughly checked for consistency, errors and plagiarism.
- The use of AI systems must be disclosed and documented in a comprehensible and transparent manner. Users should therefore always provide information when AI has been used (e.g. via a disclaimer, footnote, note, citation, etc.). This task of maintaining scientific integrity is a guiding principle of good scientific practice. In scientific work, AI systems must be cited in detail, for example, to make data analyses comprehensible and, as far as possible, replicable.
- Open science as a guiding paradigm of scientific work in Helmholtz must also be followed when applying AI systems. Beyond documentation, data, models and further materials are to be made available in trustworthy repositories in accordance with the Helmholtz Open Science Policy[6], following the motto "as open as possible, as closed as necessary".

4.4 Risk: Bias / prejudice due to training data

Bias or prejudice present in the data used to train an AI system can persist and even be amplified in generated content. These biases may be towards ethnic groups, religious communities etc., but they may also influence which data, metadata or sources are used by AI systems in their output, how often and to what extent. As a result, the unchecked use of AI-generated content can itself lead to discriminatory behavior on the part of users. This in turn can lead to further dissemination of such biases and prejudices, thus reinforcing them in the long term.

Recommended action for risk mitigation

- When using AI-generated content, users are advised to carefully check for bias/prejudice. This applies in particular if AI-generated results are used for decision making that have a significant impact on third parties.

[6] See Helmholtz Open Science Policy

4.5 Risk: Violation of AI-related regulation

AI is increasingly becoming a commodity. Even moderate programming skills are now sufficient to set up and provide AI services, and it is to be expected that this will become even easier in the future. Therefore, there is a growing risk of unknowingly using and/or developing services that implement legal regulations insufficiently, or not at all.

Recommended action for risk mitigation

- The "EU AI Act" is the first comprehensive law to create a legal framework for the use of AI services. The EU AI Act[7] takes a risk-based approach to regulating AI systems and, depending on how an AI system is classified, imposes strict obligations on providers, operators and developers of AI systems. Violations of these obligations can be punished with significant fines. Therefore, developers, operators and providers should seek legal advice before offering or operating AI services.
- Users should make sure that the AI services they choose fully comply with legal requirements.

5. Good practice when using and developing AI systems

Reviewing content: All AI-generated content should be carefully reviewed by users to mitigate the risks listed above as much as possible. Users are responsible for the content they use/distribute and may be held liable if the content violates the law. This applies in particular if the AI-generated results are used for decision-making or evaluation.

Transparency and reproducibility: Users should always be transparent and openly document when and to what extent AI was used (e.g. via a disclaimer, footnote, note, citation, etc.). Developers of AI systems should make sufficiently transparent to users and experts their basic operation, and the data used for training. In the spirit of Open Science, models and data should be made accessible on trusted infrastructures, following the motto "as open as possible as closed as necessary" in line with the Helmholtz Open Science Policy. In addition, users should always be informed when they are interacting with an AI system rather than a human.

Responsible (non-)use: AI services should not be used if there is a risk of sensitive data being disclosed, e.g., personal data, confidential business data, trade secrets, unpublished research results/data, research proposals, etc.

Use of certain services: Where possible, users should use AI services that provide transparent information about their data sources. For data protection reasons, users should also favor AI systems that are hosted in Europe.

[7] More information on the EU AI Act can be found here.

HELMHOLTZ

Spitzenforschung für
große Herausforderungen.

Imprint

Hermann von Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V.
Geschäftsstelle Berlin
Anna-Louisa-Karsch-Straße 2, 10178 Berlin
www.helmholtz.de