Leopoldina
Nationale Akademie
der Wissenschaften

2024 | Discussion No. 34

Generative AI – Beyond Euphoria and Simple Solutions

Judith Simon | Indra Spiecker gen. Döhmann | Ulrike von Luxburg

Imprint

Publisher
Deutsche Akademie der Naturforscher Leopoldina e. V.
– German National Academy of Sciences –
President: Prof. (ETHZ) Dr. Gerald H. Haug
Jägerberg 1, 06108 Halle (Saale), Germany

Editors
Christina Hohlbein, Dr. Sebastian Wetterich, Dr. Charlotte Wiederkehr,
Dr. Matthias Winkler
German National Academy of Sciences Leopoldina
Contact: politikberatung@leopoldina.org

Translation
GlobalSprachTeam – Sassenberg e. K., Berlin

Design and typesetting
Klötzner Company Werbeagentur GmbH, Reinbek

DOI
https://doi.org/10.26164/leopoldina_03_01245

Bibliographic information published by the German National Library
The German National Library lists this publication in the German National Bibliography. Detailed bibliographic data is available online at https://portal.dnb.de.

Recommended citation
Simon, J., Spiecker gen. Döhmann, I. & von Luxburg, U. (2024): Generative AI – Beyond Euphoria and Simple Solutions. Discussion No. 34, Halle (Saale): German National Academy of Sciences Leopoldina.

Editorial deadline
November 2024

First published in German by Nationale Akademie der Wissenschaften Leopoldina (2024): Generative KI – Jenseits von Euphorie und einfachen Lösungen. Diskussion Nr. 34, Halle (Saale).

Generative AI – Beyond Euphoria and Simple Solutions

Judith Simon | Indra Spiecker gen. Döhmann | Ulrike von Luxburg

Contents

1 Introduction

Since the release of text generator and chatbot ChatGPT by the American software company OpenAI in 2022, generative Artificial Intelligence (AI) has taken the digital world by storm; AI applications are now generally accessible and have a versatile range of uses. ChatGPT alone reached around 100 million users within just two months. In addition, tools for the automatic creation of photorealistic images and videos such as Midjourney, Dall-E or Gemini are also already being widely used and many tools now provide multi-modal outputs. At the click of a button, these applications can create high-quality texts, images or videos by means of generative AI. Immediate use of such tools via free access over the Internet and simple interfaces is responsible for this triumph; users need almost no prior knowledge and only a few technical requirements in order to receive answers to all kinds of questions or to generate text, images and videos in a matter of seconds.

The foundation and core of generative AI is the capacity to create new linguistic or visual products based on learned patterns from multifarious data of all sorts of origins and quality. Importantly, this creation of new content is purely based on correlations or probabilities, but not on a real understanding.

ChatGPT is a so-called large language model (LLM) that is trained with huge amounts of text: Websites, books, articles, song lyrics, posts, tweets, comments or other statements – in short, with all text types that can be found on the Internet. The training specifically consists of predicting the next word for provided sentence segments based on linguistic patterns that are learned from these data. For this purpose, ChatGPT first analyses the context of the sentence by using statistical procedures and then produces the next word based upon probability calculations. This way ChatGPT can answer questions word for word in a statistically plausible manner and produce new texts. While a statement

generated in this way can be correct, this is not necessarily the case – in fact, the chatbot's answers are often false. ChatGPT creates sentences that sound reasonable but are completely made up ("hallucinated"). This means that the large language model understands neither the content that it analyses nor anything that it outputs – it has simply learned which words and characters usually follow each other in certain contexts and applies this pattern to its own text production.

Generative models for the creation of images and videos work in a similar way. They are trained with huge amounts of image and text material primarily from the Internet in order to learn the connection between linguistic description and an image. Such tools then create new images or videos by "refining" a starting image containing only random pixels into an image or video that matches the linguistic instruction.

All these forms of generative AI have one thing in common: they create (generate) media products that have never existed before. This means that AI-generated content is more than a simple reproduction of existing texts, images or videos. In addition to its wide range of usage, it is above all the ability of generative AI that continues to fascinate; with just a few clicks it is possible to create realistic-looking pictures and original texts that sound perfect. Users often only – if at all – realise that these are AI-generated images and texts at second glance. This causes different forms of deception – deception about the fact that we are interacting with an AI, deception about the produced materials but also about the capabilities attributed to AI. Moreover, it is important to note that newly generated text and images are completely based on training data, i.e. on uncritical absorption of randomly chosen text and image material available online. This means, however, that all AI applications also reproduce the values expressed in the training data, including prejudices, distortions and limitations. These aspects will be examined in more detail in the following.

The various usage possibilities and potential economic gains promised by the use of AI in general, but also the ethical and regulatory challenges connected to this use, have already been addressed multiple times.[1] Generative AI is a catalyst which heightens the opportunities and risks of AI technologies even further – particularly in light of the crucial role of language and images for human interaction. While language is the central medium of human communication and information transfer, images and videos are significant for questions of evidence, but also for the conveyance and creation of emotions.

An especially problematic aspect of generative AI tools such as ChatGPT, which has up until now been insufficiently addressed, concerns its integration in numerous applications that have not been explicitly curated for this. In many countries including Brazil, China, Estonia or the USA for example, applications are already being developed for legal uses intended to support or even take over court decisions on the basis of AI.[2] This means that, unless the AI has been very specifically developed, jurisdiction influenced by AI is based on various types of data, values, legal experiences and prejudices that are skimmed for training purposes from the most diverse sources, including the Internet, instead of being based on a pillar of legal values and interests according to the corresponding legal framework.

It is therefore necessary to look at the real and urgent risks related to generative AI systems and to ensure their responsible development and use. Firstly, it is important not to be distracted by pseudo debates about singularity, the end of humanity due to AI, the dissolution of the labour market, a required legal personality, or claims about chatbots seemingly having consciousness. However intuitive AI systems may appear to be, they currently do not have any understanding or consciousness – instead they recognise patterns in enormous amounts of data and learn to recombine characters and patterns and to reproduce

1 For example, High-Level Expert Group on Artificial Intelligence (2019a, 2019b, 2020a, 2020b); Data Ethics Commission (2019); German Ethics Council (2023); Leopoldina, acatech, Union of the German Academies of Sciences and Humanities (2021); Orwat (2019), Spiecker gen. Döhmann & Towfigh (2023).

2 Research and Documentation Services of the German Parliament (2021).

these in new contexts. As explained above, these are purely statistical applications. The production of text, images or videos thus takes place without any understanding of the content or context – even if it appears otherwise to users.

Seven problem areas are sketched out below. While they are not exclusive for generative AI, they become particularly evident in this context:

1. The necessity and suitability of de-biasing;
2. The possibilities and limits of explainable AI;
3. The risk of fourfold deception;
4. Power and power asymmetries;
5. AI as a critical infrastructure;
6. Responsibility diffusion and control deficits, and
7. The availability and openness of AI systems.

2 Ethical and social challenges posed by generative AI

2.1 AI as a conservative tool and the role of de-biasing

A fundamental challenge in the development and use of data-based AI systems exists in regard to its accuracy, neutrality and objectivity. Even if such systems indeed open up the possibility of improving decision processes, the values and assumptions contained in the data material constantly lead to systematic distortions. Numerous examples provide evidence of such distortion – biases – and the resulting discrimination against people or social groups.[3] This is concerning yet not surprising as data-based AI systems are inherently conservative: If not actively counteracted, such systems mirror societal and cultural norms, values and relationships, including existing inequalities and injustices stemming from their data bases or methodological choices. If these flow into AI-generated predictions and decisions, conditions of the past are perpetuated and then cemented – hidden in seemingly neutral algorithms. This means that, even in 2024, an AI based on data from 1950 would thus make or suggest decisions according to the parameters of 1950.

The most recent debates about the AI system Gemini, Google's AI assistant, show just how complex the problem of generative AI technology actually is.[4] Google's AI assistant has, among other things, depicted the founding fathers of the USA, who in reality were all white and male, as more diversified, i.e. with different skin colours and genders. The consequence was a heated debate that exposed a fundamental question: Should text, images and videos of societal realities with all their inequalities and injustices be reproduced or should this be actively counteracted – for example using "de-biasing metrics"? The question can clearly not be categorically solved as its answer depends on both specific values and on context. For example, are we talking about the

3 For example Angwin et al. (2016).

4 Frankfurter Allgemeine Zeitung, 28/02/2024.

representation of historical events or about illustrating an advertisement campaign? Normative (pre)decisions in the development and training of generative AI are unavoidable, yet they are hardly being addressed. As a consequence, these decisions remain hidden and there is a lack of awareness among many users.

Methodological decisions in the development of AI models thus often have ethical and political implications. Here it is important to consider that both foregoing de-biasing and the decision to carry out de-biasing, as well as the choice of specific de-biasing methods or fairness metrics, are of enormous ethical and political significance. Such decisions require both technical-mathematical and political-ethical expertise and should therefore not be left to developers alone. They require, especially in the case of influential and far-reaching AI systems, a wider social and interdisciplinary discourse and the participation of societal groups potentially affected by discrimination.

2.2 Explainable AI caught between overblown hope and deception

Another problem area is the technically inherent lack of transparency, accountability and control in AI systems.[5] This problem is particularly visible in generative AI models. For users it is not understandable how generative AI tools produce their texts, which training data they draw from, and to what extent the information they provide is actually correct and thus reliable. However, transparency and accountability are urgently needed for myriad reasons. On the one hand, in order to assess the quality of the output and to reveal assumptions inherent in the model and the systematic distortion of facts (bias), or to demonstrate the discrimination of persons and groups (see also 2.1.). And on the other, in order to interpret predictions and to justify decisions as well as to counteract the different forms of deception (see also 2.3.)

Explainable AI (XAI) promises to deliver a solution for these problems. In this still young field of research, methods are being developed to make the AI-generated suggestions or decisions understandable in

5 Bordt et al. (2022); Crawford (2024).

retrospect. If we take the example of an AI decision-making process in a bank's internal assessment of loan applications, the case could look like the following. If the AI model rejects a loan application, the explanation system would then justify this decision: "Mr Schmidt will not receive the loan because he is too old and his income is too low." However, it is important to understand how such explanations do or rather how they do not come about. In the case of complex AI tools, especially in the context of neural networks, it is technically impossible, due to this complexity, to provide the real reason for the decision – because one as such does not exist. Instead, explanation processes try to find plausible reasons in hindsight. However, there are different ways to produce such explanations at a later stage, and although all these possibilities are in themselves technically plausible, in practice they often lead to completely different explanations – which is why the real underlying mechanisms which had remain hidden.[6]

Moreover, it is clear that algorithms producing explanations can be manipulated. For example, in the case of a loan decision, a discrimination-based decision system can be combined with an explanation system in a way that generally understandable, indisputable justifications are always generated – yet do not provide information on an applicant's characteristics that are actually relevant to the decision.[7] In such a case, the absurd consequence would be that the "explanation" would not serve transparency but would instead be an effective means of deception. In this context, the fact that it is not possible to find out, even by means of external comprehensive technical checks, whether a decision was in fact reasonable or whether it was purposefully manipulated is especially critical.[8] In view of the legal standards that require transparency and explainability from AI processes, this is a serious problem[9] because, in the legal context, explanations only make sense when they are checkable – which is not yet technologically possible.

6 Bordt et al. (2022).

7 Sharma et al. (2024).

8 Bordt et al. (2022).

9 European General Data Protection Regulation (abbreviated: EU GDPR); European AI Act (abbreviated: EU AI Act).

Just how difficult it is to make the decision processes of AI systems transparent becomes even more obvious when attempting to explain the results of generative AI models. Large language models make decisions based on the context of how a specific text should be designed. To explain why a language model generates a certain text, you would first have to show which context the model is referring to in the specific case; this depends on the current prompt and the already generated text, among other things. Currently, however, there is no technological approach that could track and explain the specific context connection of an AI-generated text. And it is difficult to imagine that such an approach could be found. To circumvent this problem, scientists are currently developing methods to enable large language models to explain their approaches and results themselves. The charming idea behind this is that the language model itself knows best why it generated a certain text. Unfortunately, the corresponding explanations sound logical, but are unreliable.[10] This is because, as illustrated above, a language model has neither a semantic understanding of text nor one of its own functions. It simply produces statistically plausible texts. The statements contained there can be true or made-up; and it is not technologically possible to reliably differentiate between these two categories. To quote American philosopher Harry Frankfurt[11], such systems produce "bullshit" – they make the boundary between truth and lies disappear behind reasonable sounding texts.[12] This means that, in the worst-case scenario, explainable AI systems contribute even more to deception instead of avoiding it. Legally considered, such approaches thus do not provide a reliable evaluation tool.

What do these insights imply? Methods in XAI which seek to establish ex-post explanations for deep learning in general and generative AI in particular do not produce reliable transparency. It is thus important to decide for which applications the inextricable non-transparency of generative AI is acceptable with regard to the benefits and risks. In the case of language models, this applies to simple applications like the execution of routine tasks, where mistakes are either not critical or can

10 Tanneru et al. (2024); Turpin et al. (2023).

11 Frankfurt (2005).

12 Hicks et al. (2024).

be quickly found and easily corrected, for example, the phrasing of an email based on keywords, or the creation of an application based on a CV and a job vacancy. In which contexts is transparency non-negotiable? Applications in the legal context should be named here – for example an AI system that summarises criminal or civil trial records or provides the judiciary with decision templates based on earlier court decisions. In this case, generative AI systems should be used with extreme caution, since neither the AI-generated results themselves nor the AI-generated explanation of how they were produced are of reliable assistance as shown above. While the idea of human supervision in this context generally appears to be a possible corrective measure, who should check the AI summary of an extensive criminal trial record in legal day-to-day work? And if such a check is left to humans, would not the gains in efficiency from using AI be forfeited?

XAI is thus an interesting and important field of research that should at the same time not be overloaded with expectations. When it comes to complex decision or prediction models, ex-post explanations do not help because the models do not understand the actual processes and simply deliver plausible approaches. In specific situations and contexts in which the explainability of results and how they were generated are non-negotiable, methods must therefore be used whose prediction models are less mathematically complex yet can be directly interpreted by humans – even if this may lead to less suitable results. An example for such simple prediction models are decision trees.

2.3 The problem of fourfold deception

As already described, generative AI models can now produce texts, images and videos in very high quality, which involves the risk of deceiving users. In this context at least four different dimensions of deception can be identified, which presents a core problem when using generative AI.[13]

13 Messeri & Crockett (2024); German Bundestag, Committee on Education, Research and Technology Assessment, committee document 20(18)108b, 21 April 2023. Expert interview on ChatGPT, Professor Dr Judith Simon, Universität Hamburg; https://www.bundestag.de/resource/blob/944448/004ca2f7a9fcf586a07113c6ba72b689/20-18-108b-Simon-data.pdf [last accessed: 20/08/2024].

First, this affects the interaction between user and chatbot as long as it remains unclear that the latter is not a person, but an AI system. The question as to whether you are interacting with a chatbot or with another person is already relevant for customer service and the moderation of social media. However, it is much more relevant for especially sensitive contexts, for example in psychotherapy.

Further deception potential exists regarding the capabilities of generative AI models. Even though currently available AI systems neither have a semantic understanding nor consciousness, it can still appear to be the case for users – even when they know they are interacting with a technical system. Early experiences with the software ELIZA[14], which was developed in the 1960s by German-American computer scientist Joseph Weizenbaum, have already shown this, as do current reports on the interaction of users with ChatGPT: People clearly tend to attribute understanding, empathy or consciousness to a chatbot if it communicates in a plausible way. While such attribution of human capabilities does not say anything about the actual functions and capabilities of the machine, it does say something about the human tendency to anthropomorphise technology.

The third dimension of the deception problem relates to the results of generative AI systems that can involve manifold risks: from fake news and deep fakes for propaganda, defamation and bullying, to the criminal use of faked voices in order to defraud relatives or to provide misleading evidence in court trials. Deception and manipulation are certainly not new phenomena, but the quality, simplicity and the generally low technological and technical requirements as well as the speed at which texts, images or videos are produced and spread now open up a completely new scope for possible misuse. Despite all the opportunities that generative AI models create, ChatGPT and other systems harbour a real threat for our democratic constitution as fundamental information and communication processes can be quickly and easily disrupted, and evidence and credibility no longer constitute reliable categories. This situation is emphasised by current examples from election campaigns in Slovakia, the USA, and the European Union.[15]

14 Weizenbaum (1966, 2023).

15 For example Wired, 03/10/2023.

Ultimately, users can also be deceived about the functions of gener-ative AI when it is integrated into other systems. The problem especially comes to light where the provision of existing information is mixed with the generation of new information. After all, if information available online has different truth values, i.e. it could well be true or false, there is a massive difference as to whether clicking a link opens already existing information or whether new information is generated.

The previously separate processes of information retrieval and information creation are increasingly blurred as generative AI is being built into more and more applications, software systems and tools – from PDF readers, email programmes and Internet search engines to entire software packages such as the AI assistant Microsoft Copilot. This means that it is becoming increasingly difficult for users to differentiate between existing and newly generated content, which makes it even more difficult to classify and assess the information's quality and origin.

2.4 Questions of power and power asymmetries

Another challenge for the development and use of generative AI exists with regard to potential power asymmetries, particularly when it comes to the use and evaluation of personal data. Multiple risks for users' privacy, autonomy and independence arise from the processing of both original, freely available training data and user data that has been deliberately provided. Especially in the context of scoring or personalised offers, companies and the state are already using multiple data to make users specifically customised offers. Even if this may be helpful, they can also use this advantageous knowledge about individuals or groups against them. This happens, in the case of so-called "dark patterns" when decisions are influenced subliminally or for the one-sided design of contract conditions or access to state services because the preferences or constraints of users are exploited. Such possibilities of misuse are hugely increasing under AI. The results of corresponding data-analyses are in no way always appropriate, let alone normatively desirable. Power asymmetries increase between those who make assessments using AI systems and those who are assessed by AI systems.

Users of AI models do not know what is being concluded from their behaviour, which means that self-defence strategies mostly prove useless. This affects all users of AI systems, but above all socially marginalised individuals and groups.

Furthermore, questions of intellectual property and copyright are also more pressing in the context of generative AI, as the models are (often) trained on the basis of data whose copyright holders have been neither informed about the use of this data nor have they received appropriate compensation – let alone a share in the added value generated by AI. As the first court proceedings on copyright violations in the context of AI show, the issue generally also concerns access rights to data which is especially easily available online and can be used one-sidedly for the development of AI.

To date, however, there is no convincing concept to involve the people individually and collectively whose data an AI application's success depends on to establish a fair distribution of opportunities and risks. The same applies to the question of how to guarantee that copyright holders retain the power of ultimate decision when it comes to the use and transfer of their data by generative AI systems.

2.5 AI as critical infrastructure

A particularity of generative AI models is, as already described, their wide applicability. Large language models in particular are not only utilised directly by users via web interfaces but are also increasingly being integrated into a number of products, processes, and services. This takes place by incorporating so-called foundation models.[16] AI systems in general as well as generative AI and language models in particular thus form a type of critical infrastructure in two regards: firstly, in relation to their significance for the variety of different uses that make AI systems unavoidable and create a great dependency, and secondly, in terms of their invisibility.

16 These are base models trained by means of machine learning with huge amounts of data. They can be used for different downstream applications. This also includes large language models.

Several risks arise here. The incorporation of generative AI in different software systems first leads to the problem of fourfold deception mentioned above (see 2.3.) – i.e. to the blurring of the boundary between existing and newly generated content. Furthermore, it creates or intensifies problems that have already been discussed for algorithmic systems in general under the term of "algorithmic monoculture"[17]; there is ample evidence that complete foundation models or individual components, in particular training data, but also software packages are shared for both the development and use of AI applications. Homogenisation of these models' output can be a consequence of this shared use. This means that different application systems, for example text or image generators, possibly create very similar results because they were trained with the same data and AI models. Subsequently, errors as well as biases and possibly resulting discrimination can systematically take effect in the results[18], which, however, goes unnoticed because the foundation models in the different applications remain invisible. In light of this, it becomes clear why high transparency and quality standards must apply for foundation models in particular.

2.6 Responsibility diffusion and control deficit

The many different actors and influences in the life cycle of generative AI systems, ranging from development to integration in other application systems and up to the use of such models, lead to an enormous diffusion of responsibility. From the outside it is mostly not identifiable what these influences are specifically, and which actors have which effects on the functions and results of generative AI models. Sometimes, this intransparency is even intended, for instance in the case of deep fakes. In this way, however, it remains unclear who can be effectively assigned the responsibility for both the positive and negative effects of this technology and who should be responsible for the political, legal, social and economic consequences – this poses a risk for democracy and the constitutional state.

17 Kleinberg & Raghavan (2021).

18 Bommasani et al. (2022a, 2022b).

This is even more true in the case of generative AI, not only because the personal and institutional actors and their roles remain unknown, but also because the technological background and functioning are inaccessible. It is usually not possible to find out whether an image or text has been created or altered and if so, by whom, and to what end. Controlling how decisions are made in the context of generative AI, however, requires access to relevant parameters such as content, purpose, base values, selection criteria, authorship and decision requirements, whether it is carried out by private persons, the press, state supervisory authorities, courts or others. As a consequence, users must blindly trust each AI model and the corresponding application system if they want to use them – without such trust being actually justified, which means that legal and ethical rules remain unaccounted for.

This responsibility and control deficit is reinforced by three other circumstances. First, AI can support people and institutions with decisions in various ways, which makes assigning responsibility more difficult. It can be used as a supporting tool wherein the actual decision is made by the person, e.g. when writing a job application. AI is conferred much more weight if the decision is AI-based and the person can merely intervene in exceptional cases to change the AI decision. However, the loss of control is most severe when there is no possibility for the person to change the AI-based decision, for instance when generative AI is incorporated in other applications, software systems and tools. Especially when it comes to the use of AI in complex autonomous systems, the latter is considered essential to its functionality.

The second circumstance is that it cannot be concluded which information and how the information has flowed into the produced decisions. Thus, it cannot be traced how an AI application – provided it is recognisable that such an application is being used – detects, asseses, sorts and processes information and according to which standard criteria this takes place. In this respect, however, it is not possible to assess whether the AI-generated result has been generated correctly according to certain criteria. The continual, large-scale adjustment of AI is without a doubt its particular technological strength, yet it makes technical solutions for parallel or subsequent checks hard to imagine.

Thirdly, the control deficit is exacerbated by the technical impossibility of reproducing the decisions of an AI application for control or monitoring purposes. The problem of creating transparency already mentioned above is thus closely related.

Due to the number of actors involved, the far-reaching integration of generative AI in numerous products, applications, services and tools works against a clear assignment of responsibility, the maintenance of legal rules and ethical standards, and effective control by users or state organs.

2.7 Availability and openness of AI

The necessity of and the boundaries between transparency and accountability pose questions in relation to the specific modalities of availability and openness of generative AI systems.

Open systems, in particular open-source solutions, present various advantages, the first of which is improved checkability. While the functions of proprietary systems usually are in the economic interest of the owners or are conceived as shareholder-oriented, wherein non-transparency as well as political and social risks are connected, open systems are more transparent; they facilitate external control and checks regarding the data, methods and models used. At the same time, full transparency in the open-source segment is not always achievable because some training data cannot be published due to reasons of data protection.

Beyond this, open-source solutions generally promise reliable accessibility, which means that the developed systems will also be available without restrictions in future. We can then, for example, rule out a situation in which public administration would develop a language model-based application system and that the use of the language model would then become limited or expensive. The example of the AI tool AlphaFold 3 shows that precisely this can happen in the proprietary

systems segment.[19] In addition, open-source systems also facilitate the development of applications with a lower demand, i.e. where there is no (or little) economic interest on the part of the producers. This applies, for example, of language models for languages spoken by a minority.

Nevertheless, open-source systems also present many disadvantages in the field of generative AI because an abusive use of the technology is hugely facilitated by the public nature of the source code. Every person, group or institution can use and alter this source code for their own purposes, such as generating fake news or large numbers of hate comments. Long-term availability and security also naturally depend on how each open-source solution is continually operated.

It is therefore paramount to have a public and open discussion about the benefits and disadvantages of open AI models in order to provide balanced, practical and democratic technological solutions in the future. The present constellation of a simple yet only seemingly free access to opaque, proprietary systems likely represents the worst possible combination for generative AI technology; in this way freely available knowledge and user data are drawn upon without consent and subsequently made available according to non-transparent, commercially motivated rules and requirements, while these processes lack an assessment including societal perspectives and the guarantee of legal controls or monitoring.

19 AlphaFold is an AI tool for researching proteins. It was developed by the company DeepMind, a Google subsidiary, and is implemented worldwide in research. In the newest version from May 2024, DeepMind suddenly largely restricted the free-of-charge access of public research and no longer provides free access to the tool's source code, so that it is not possible to check what is actually happening in the background.

3 Conclusion

The objective of this discussion paper is to take a realistic look at the opportunities and risks in the development and use of generative AI – and with this to counteract both the utopian promises of salvation and the dystopian warnings characterising much of the contemporary debate. The considerations formulated here focus on the risks for individuals, democracy, economy and society that have not yet been reflected upon enough in the public discourse. These risks are partially inherent in the functional logic of the technology itself – such as the inexplicability, the non-controllability, non-neutrality and non-objectivity of generative AI technology. Others only arise during specific use cases or due to the interaction between humans and technology in different contexts and organisational framework conditions – such as diffusion of responsibility, ease of manipulation or illusions about capabilities of AI. Many of these risks are either not at all or only insufficiently addressed by the European Artificial Intelligence Act and other laws[20], and are thus presently not part of normative guardrails for the development of generative AI in Germany and Europe.

Yet such a regulation is urgently needed. Due to the inherent normativity of processes of data collection, preparation and classification and the shaping of decision rules in generative AI systems, such systems cannot produce objective and neutral decisions which are necessarily superior to those taken by humans. Every generative AI reflects both the training data it is based on and the goals and purposes of its development that are inextricably linked to its logic of functionality. However, these are not discernible, neither in the AI itself nor in the corresponding applications, and therefore evade control and regulation via established procedures, institutions and normative orders.

20 For example the General Data Protection Regulation, European Data Act, copyright laws or procedural law for public administration.

Users of generative AI receive information based on the specific values of others without usually reflecting on these consciously. This fact enables deception and manipulation disguised by a seemingly superior and apparently neutral, objective technology. However, this fundamentally undermines our ways of perceiving the world and our power to judge based on sensory experiences in the long term. Text and images ultimately lose their evidential value.

In light of ongoing discussions about the development of trustworthy AI in Europe and a competitive advantage connected to this, measures to prevent damage need to be accelerated. This affects, for example, methods that uncover or minimise biases (so-called de-biasing) and increase the transparency of the technology (explainable AI)[21]. At the same time, it is important to warn against exaggerated expectations directed at these developments as they also have limitations or even create risks of their own, such as the exploitation of open-source codes for the creation and widespread distribution of deep fakes.

The aspects addressed in this discussion paper are of course not to be understood as exhaustive. Other critical aspects include the often precarious working conditions – particularly in countries of the Global South – in the development and deployment of numerous AI models as well as the high energy and resource consumption required for the training and use of generative AI-based tools.[22] There is no simple, general answer or ready-to-use solution to any of this. Nevertheless, and precisely for this reason, it is necessary to present all of these aspects in their entire ambivalence in order to make them accessible to a critical public debate.

21 Asghari et al. (2022).

22 Crawford (2024).

References

Asghari, H., Birner, N., Burchardt, A., Dicks, D., Faßbender, J., Feldhus, N., Hewett, F., Hofmann, V., Kettemann, M. C., Schulz, W., Simon, J., Stolberg-Larsen, J., Züger, T. (2022). *What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making.* Berlin: Alexander von Humboldt Institute for Internet and Society. https://doi.org/10.5281/zenodo.6375784

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica,* 23/05/2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Last accessed: 06/12/2024].

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., Liang, P. (2022a). Picking on the same person. Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems,* 35, 3663–3678. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/17a234c91f746d9625a75cf8a8731ee2-Paper-Conference.pdf [Last accessed: 31/07/2024].

Bommasani, R., Hudson, D. A. ..., Liang, P. (2022b). On the opportunities and risks of foundation models. *arXiv preprint,* 12/07/2022, arXiv:2108.07258v3. https://doi.org/10.48550/arXiv.2108.07258

Bordt, S., Finck, M., Raidl, E., von Luxburg, U. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. *FAccT 22. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* 891–905. https://doi.org/10.1145/3531146.3533153

Crawford, K. (2024). Generative AI's environmental costs are soaring. And mostly secret. *Nature,* 626, 693. https://doi.org/10.1038/d41586-024-00478-x

Data Ethics Commission of the Federal Government (2019). *Gutachten der Datenethikkommission.* Berlin: Federal Ministry of the Interior and Community, Federal Ministry of Justice and Consumer Protection. URL: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf [Last accessed: 06/12/2024].

Frankfurt, H. G. (2005). On bullshit. Princeton: Princeton University Press.

Frankfurter Allgemeine Zeitung, 28/02/2024. *Wir fragen eine Ethikerin. Welche Gesellschaft soll Gemini abbilden?* By Hendrik Wieduwilt. URL: https://www.faz.net/pro/d-economy/kuenstliche-intelligenz/wir-frageneine-ethikerin-welche-gesellschaft-soll-gemini-abbilden-19550166.html [Last accessed: 01/08/2024]

German Ethics Council (2023). *Mensch und Maschine. Herausforderungen durch Künstliche Intelligenz* (Statement). Berlin: Deutscher Ethikrat. URL: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf [Last accessed: 12/05/2024].

German National Academy of Sciences Leopoldina, acatech – Akademie der Wissenschaften Leopoldina, acatech – German Academy of Science and Engineering, Union of German Academies of Sciences (2021). *Digitalisierung und Demokratie.* Halle (Saale). https://doi.org/10.26164/leopoldina_03_00348

Hicks, M. T., Humphries, J., Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology,* 26, 38. https://doi.org/10.1007/s10676-024-09775-5

High-Level Expert Group on Artificial Intelligence (2019a). *Ethics guidelines for trustworthy AI.* Brussels: European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Last accessed: 20/08/2024].

High-Level Expert Group on Artificial Intelligence (2019b). *Policy and investment recommendations for trustworthy AI.* Brussels: European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence [Last accessed: 20/08/2024].

High-Level Expert Group on Artificial Intelligence (2020a). *Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment.* Brussels: European Commission. URL: https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1 [Last accessed: 20/08/2024].

High-Level Expert Group on Artificial Intelligence (2020b). *Sectoral considerations on the policy and investment recommendations for trustworthy AI.* Brussels: European Commission. URL: https://op.europa.eu/en/publication-detail/-/publication/6d266f0f-f7c3-11ea-991b-01aa75ed71a1/language-en [Last accessed: 20/08/2024].

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences,* 118(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature,* 627, 49–58. https://doi.org/10.1038/s41586-024-07146-0

Orwat, C. (2019). *Diskriminierungsrisiken durch Verwendung von Algorithmen.* Baden-Baden: Nomos. URL: https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/studie_diskriminierungsrisiken_durch_verwendung_von_algorithmen.html [Last accessed: 01/03/2024].

Research and Documentation Services of the German Parliament (2021). *Künstliche Intelligenz in der Justiz: Internationaler Überblick.* WD 7 - 3000 - 017/21. URL: https://www.bundestag.de/resource/blob/832204/6813d064fab52e9b6d54cbbf5319cea3/WD-7-017-21-pdf-data.pdf. [Last accessed: 06/09/2024].

Sharma, R., Redyuk, S., Mukherjee, S., Sipka, A., Vollmer, S., Selby, D. (2024). X Hacking. The threat of misguided AutoML. *arXiv preprint,* 12/02/2024, arXiv:2401.08513v2. https://doi.org/10.48550/arXiv.2401.08513

Spiecker gen. Döhmann, I., & Towfigh, E. V. (2023). *Coded Bias: The General Equal Treatment Act and protection against discrimination by algorithmic decision-making systems.* Berlin: Federal Anti-Discrimination Agency. URL: https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/EN/publikationen/ki_study.pdf?__blob=publicationFile&v=2 [Last accessed: 21/10/2024].

Tanneru, S. H., Agarwal, C., Lakkaraju, H. (2024). Quantifying uncertainty in natural language explanations of large language models. Proceedings of the 27th International Conference on Artificial Intelligence and Statistics. *Proceedings of Machine Learning Research,* 238, 1072–1080. URL: https://proceedings.mlr.press/v238/harsha-tanneru24a/harsha-tanneru24a.pdf [Last accessed: 20/08/2024].

Turpin, M., Michael, J., Perez, E., Bowman, S. R. (2023). Language models don't always say what they think. Unfaithful explanations in chain-of-thought prompting. *NIPS '23. Proceedings of the 37th International Conference on Neural Information Processing Systems,* 3275, 74952–74965. URL: https://dl.acm.org/doi/10.5555/3666122.3669397. [Last accessed: 31/07/2024].

Weizenbaum, J. (1966). ELIZA. A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery,* 9, 36–45. https://doi.org/10.1145/365153.365168.

Weizenbaum, J. (2023). *Die Macht der Computer und die Ohnmacht der Vernunft* (16th Edition). Frankfurt am Main: Suhrkamp.

Wired, 03/10/2023. *Slovakia's election deepfakes show AI is a danger to democracy. Fact-checkers scrambled to deal with faked audio recordings released days before a tight election, in a warning for other countries with looming votes.* By Morgan Meaker. URL: https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/. [Last accessed: 21/08/2024].

Contributors

Authors

Prof. Dr. Judith Simon	Chair of Ethics in Information Technology, University of Hamburg
Prof. Dr. Indra Spiecker gen. Döhmann	Chair of Public Law and Law of Digitalization, University of Cologne
Prof. Dr. Ulrike von Luxburg ML	Professor for Computer Science, University of Tübingen

In accordance with the published "Rules for Dealing with Conflicts of Interest in Science-Based Advisory Activities of the National Academy of Sciences Leopoldina" (Regeln für den Umgang mit Interessenkonflikten in der wissenschaftsbasierten Beratungstätigkeit der Nationalen Akademie der Wissenschaften Leopoldina), the contributing scientists have been obliged to disclose facts that may be likely to lead to conflicts of interest. In addition, reference is made to the present rules.

Scientific Officers and Coordination

Dr. Sebastian Wetterich	German National Academy of Sciences Leopoldina
Dr. Charlotte Wiederkehr	German National Academy of Sciences Leopoldina
Dr. Matthias Winkler	German National Academy of Sciences Leopoldina

ML – Member of the Leopoldina

Selected publications from the “Leopoldina Discussion“ series

No. 35: Die gemeinsame Verantwortung für das archäologische Erbe. Warum der archäologische Kulturgutschutz besser in die akademische Ausbildung integriert werden muss – 2024 *

No. 33: Vernetzte Notfallvorsorge für Kulturgüter. Eine Umfrage unter den Notfallverbünden Deutschland – 2023*

No. 32: Ein öffentlicher Dialog zur Fortpflanzungsmedizin – 2023*

No. 31: Den kritischen Zeitpunkt nicht verpassen. Leitideen für die Transformation des Energiesystems – 2023*

No. 30: Organisatorische Voraussetzungen der Notfallvorsorge für Kulturgüter – 2022*

No. 29: Die rechtlichen Grundlagen der Notfallvorsorge für Kulturgüter – 2022*

No. 28: Ärztliche Aus-, Weiter- und Fortbildung – für eine lebenslange Wissenschaftskompetenz in der Medizin – 2022*

No. 27: Nutzen von wissenschaftlicher Evidenz – Erwartungen an wissenschaftliche Expertise – 2021*

No. 26: Neuregelung des assistierten Suizids – Ein Beitrag zur an Debatte – 2021*

No. 24: Global Biodiversity in Crisis – What can Germany and the EU do about it? – 2020

No. 23: Traces under water – Exploring and protecting the cultural heritage in the North Sea and Baltic Sea – 2019

No. 22: Übergewicht und Adipositas: Thesen und Empfehlungen zur Eindämmung der Epidemie – 2019*

No. 21: Wie sich die Qualität von personenbezogenen Auswahlverfahren in der Wissenschaft verbessern lässt: Zehn Prinzipien – 2019*

No. 20: Gemeinsam Schutz aufbauen – Verhaltenswissenschaftliche Optionen zur stärkeren Inanspruchnahme von Schutzimpfungen – 2019*

* available only in German

All publications from the series are available for free in PDF format at:
https://www.leopoldina.org/en/publications/statements/discussion-papers

Deutsche Akademie der Naturforscher Leopoldina e. V.
– German National Academy of Sciences –

Jägerberg 1
06108 Halle (Saale)
Phone: 0049 345 472 39-867
Email: politikberatung@leopoldina.org

Berlin Offices:
Reinhardtstraße 16
10117 Berlin

Unter den Linden 42
10117 Berlin

The Leopoldina originated in 1652 as a classical scholarly society and now has 1,700 members from almost all branches of science. In 2008, the Leopoldina was appointed as the German National Academy of Sciences and, in this capacity, was invested with two major objectives: representing the German scientific community internationally, and providing policymakers and the public with science-based advice.

The Leopoldina champions the freedom and appreciation of science. It is the role of the Leopoldina to identify and analyse scientific issues of social importance. The Leopoldina presents its policy recommendations in a scientifically qualified, independent, transparent and prospective manner, ever mindful of the standards and consequences of science.

www.leopoldina.org/en