NOT IF, BUT HOW

HSB | Munich RE

Whitepaper

Insuring Generative AI: Risks and Mitigation Strategies

Balancing creativity and responsibility to enable adoption

Content

1 Introduction

With the introduction of ChatGPT and GPT4 in late 2022 and early 2023 respectively, consumer-facing generative AI (GenAI) tools have captured the public's attention, despite having fascinated experts for years.

GenAI is a transformational technology which impacts how organisations and people operate[1]. It is estimated that GenAI will significantly contribute to the global economy[2] and is predicted to affect two-thirds of US occupations[3]. With its capacity to produce assets like images or text, make unstructured data accessible, enable AI access for a layman, unlock new business opportunities and drive advancements across the organisation, both GenAI's attractiveness, but also the risks that come with the use of this technology, need to be examined.

With new risks, the question of available risk transfer solutions arises. Having insured our first AI risk in 2018 and our first Large Language Model (LLM) in 2019, Munich Re has been following the recent advancements in GenAI with particular interest. We believe that insurance will become vital for a smooth and widespread adoption of GenAI and management of emerging AI risks over the years to come. Collaboration of the insurance and tech industries can aid in unlocking the tremendous potential of GenAI for all.

This document will provide risk management considerations for GenAI use cases for decision makers. After a short introduction into GenAI, we share our thoughts on the novel risks that GenAI introduces – compared to other types of AI models – and will provide a risk management recipe of how one would assess, price and insure some of these risks.

2 What is generative AI?

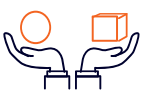
GenAI refers to AI models that generate new data, such as text, images, or audio in general or for specific tasks, like new efficient bridge designs or software code. The models do so by learning patterns from their training or input data. Depending on the input data and the desired outcome, the techniques used to generate new data differ:

Model Types	Examples	Use Cases
Image-to-Image	– Variational Autoencoders (VAE) – Generative Adversarial Networks (GAN)	Image denoising and inpainting, image editing, style transfer
Text-to-Text	– Generative Pre-trained Transformer (GPT) – Sequence-to-Sequence (Seq2Seq) – Text-based Variational Autoencoders (VAEs)	Text generation, machine translation, text summarisation, chatbots
Multimodal Model	– Contrastive Language-Image Pre-training (CLIP) – Stable Diffusion – DALL-E	Image captioning, visual question answering, text-to-image generation, video analysis

Table 1: Generative AI models and use cases

– For **image generation**, different models have been developed to create an image that is similar to the input image. Such models are Variational Auto-encoders (VAE) and Generative Adversarial Networks (GAN).

– For **text generation**, LLMs are trained in a similar fashion to models which generate images. By trying to recover the masked word(s) in a sentence, the model learns the structure in text documents of a certain language. The more recent LLMs are based on transformers, which utilise an attention mechanism to only learn from the important words to predict the next word efficiently. This approach makes it possible to develop broad **"foundation models"**[33] that are trained on a broad set of unlabelled text data, which can be fine-tuned for different specific tasks. The recent GPT-3 and GPT-4 models have shown potential to be used as the foundation model for many tasks, such as building chatbots, solving maths problems, generating programming code, summarising legal documents, translation, etc.



– Combining the benefits of language and image-generation models, **multi-modal models** are opening the door to many new applications. For example, the text-to-image models that are trained on image and text pairs can generate images by following human descriptions (i.e. prompts), which makes them easy to be used by the general public. Stable Diffusion and DALL-E are examples of this type.

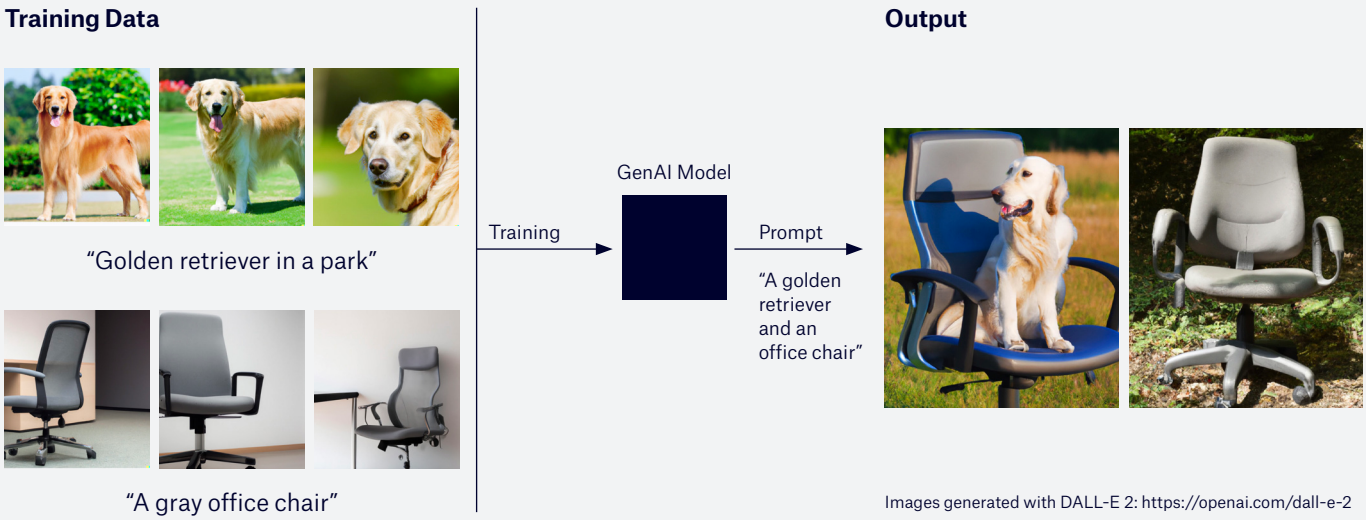


Figure 1: Example of multimodal model: the text inputs interact with the learned patterns from images to generate new images.

For an example of the workings of a multimodal model: The model is trained on images with a descriptive label. When the model is prompted, it creates a new image using the learned patterns. The results can fit to the expectation of the user (the left output) or not (the right output). For the right output, the model did not understand what the training data images contain and created a park based on the training data. This risk can be reduced by a more diverse and bigger training dataset.

The above GenAI techniques are beginning to inspire many different business cases in text and image generation for advertisement, automating legal tasks, automating software development, and many more. As companies rely more and more on GenAI tools for process automation, risk managers need to evaluate the risks that emerge for their organisation by relying on AI.

3 Risks of generative AI

GenAI applications share many of the risks also faced by AI. Their output is dependent on the quality of the data inputs – both during training and by users. Furthermore, their lack of control over the desired output as well as the general opaqueness of the "black box AI system" can expose users and businesses to a number of well-documented risks.

Due to network effects, as well as positive feedback loops, GenAI models get better the more they are being used. This can result in a concentrated market, with first-mover models being vital building blocks for fine-tuned models[4]. Built upon pre-trained foundation models, GenAI's output can be significantly different from the input data that the business users feed it for fine-tuning the models to fit to their respective applications. This opaqueness in the models makes it harder for the users, as well as the downstream application developers, to understand and control the risks well.

Due to the additional layer of opaqueness specific to GenAI models, it is vital to be aware of the potential risks as well as knowing how to mitigate and transfer them. In this chapter, we will list some of the more prominent risks that GenAI exhibits, as well as possible ways to mitigate those risks. Afterwards, we will go one step further to discuss some of the risk transfer options currently available.

Risk Types	Description
Hallucination and false information	Generation of false information or misleading content.
Bias and fairness	Generating unfair or biased output resulting in the discrimination of a protected class.
Privacy infringement	Leak of private or sensitive information.
Intellectual property violations	Generation of content that is trained on IP-protected materials without permission; or generation of content that mimics licensed material (derivative work).
Harmful content	Generation of offensive or malicious content, illegal materials.
Other risks and environmental risks	Increasing number of parameters increases need for training/retraining the models, which increases the energy consumption.

Table 2: Some risks associated with GenAI and their meaning

3.1 Hallucinations and false information

In text-based GenAI, one may find that some generated answers or articles look plausible and follow an inherent logic, while the statements are factually false. Image-generating AI can create fictional images that can be misleading – with potentially significant financial and reputational consequences. The phenomenon of GenAI generating misinformation or inaccurate output is called "hallucinations".

Being able to trust the output of GenAI is essential for businesses that use such models to automate tasks or generate information. Hallucinations erode that trust. As an example, a US lawyer encountered the consequences that the risk of hallucinations can entail in 2023 first hand. The lawyer used an AI chatbot to aid him in researching relevant case law. The cases cited by the AI chatbot – as well as the subsequent reassurance by the AI that the cases were, in fact, real – were a hallucination of the AI[5]. This "bogus judicial decision with bogus quotes and bogus internal citations"[6] received a "stern admonishment"[7] by the federal judge ruling the case, tarnishing his law firm's reputation.

There are a few technical ways to mitigate the risk of hallucinations and false information, although the risk can never be fully avoided and a residual risk will remain. One way is fine-tuning using labelled data, which can improve the accuracy of foundational GenAI models on specific tasks. Furthermore, recent developments on adopting reinforcement learning from human feedback (RLHF) in LLMs is a direct step towards aligning LLMs with human intent[8], and has been shown to be powerful for reducing the likelihood of hallucinations.

Despite the available tools and techniques available to mitigate the risk of GenAI models hallucinating, even the best GenAI models will still occasionally generate factually inaccurate content – leading to the undesired consequences described above. In order to effectively manage the risk of models hallucinating, technical as well as financial risk management instruments are required. In addition to the technical mitigation tools available, the risk mitigation toolbox for companies working with GenAI should also include insurance solutions.

3.2 Bias and fairness

The risks of bias in AI models has been discussed at length, especially when machine learning models are used in high-stake applications. They have the potential to result in unfair and discriminatory outcomes, especially worrying in areas of e.g. credit scoring, loan approval, employment or advertising. This discussion has been taken to the regulators, leading to two recent settlements on discriminatory practices by AI with respective regulatory fines for discriminatory marketing[9], and discriminatory hiring practices[10].

Biased GenAI models could have an even broader and more impactful effect on society by magnifying unwanted bias already found in data sets. For example, if a model is trained on a homogenous dataset, the trained model may reflect the characteristics of this dataset accurately (including its undesirable properties), reinforcing harmful stereotypes that could lead to far-reaching discriminatory events and perpetuate the oppression of historically underrepresented groups or opinions in the public space.

Researchers have always been aware of bias in AI. Mitigating bias in AI models is, however, a balancing act. Firstly, usually a compromise needs to be made between fulfilling fairness metrics and achieving a high model accuracy – recently leaving researchers to thread the needle through the optimisation approach[11]. Furthermore, fairness metrics can be split into two different types: group fairness (parity between different protected groups, such as those defined by gender or race) and individual fairness (similar individuals being treated similarly). Increasing the difficulty level further, different metrics as well as simultaneously achieving group and individual fairness is not possible. The fairness metrics are mutually exclusive – the so-called impossibility theorem of fairness[12]. Finally, there is a mismatch between the many mathematical definitions of bias and the legal interpretation used by the courts to determine unlawful discriminatory behaviour.

Fully eliminating the risk of biased models through technical mitigation is impossible. While the risks can be mitigated, a residual discrimination risk will always remain. Furthermore, regulators have issued a clear statement that relying on the assurance of an AI vendor that the AI tool will not be discriminatory will not exculpate its users[13]. With potentially expensive lawsuits looming, insuring against liability due to model bias could be an effective way to manage the residual risk caused by relying on GenAI models in sensitive areas of society and business. The ability to transfer the risk can ultimately make the difference for corporate decision makers between avoiding AI and embracing AI. Embracing AI in a risk-conscious way can create a first-mover advantage for companies and increase an organisation's competitiveness.

3.3 Privacy infringement

GenAI provides output based on given inputs and learned patterns from training data. Sensitive information in the training data and input data can be captured by the model, and may be leaked in the model outputs. A recent study on AI app usage on 10,000 employees has shown that 15% of employees paste data into GenAI and that 6% of employees paste sensitive data[14]. Even anonymised data can cause damage due to the potential for reconstruction, as alleged in Doe v. Netflix[15]. In this class action suit, a lesbian mother sued for the invasion of her privacy. She alleged that the streaming platform had made it possible for her to be outed when disclosing insufficiently anonymous data as part of a contest to improve its recommendation system, claiming that Netflix had "perpetrated the largest voluntary privacy breach to date"[16].

In order to mitigate the risk of invading privacy or breaching data protection laws, researchers in the field of AI have adopted a mathematical framework called Differential Privacy (DP) for ensuring the privacy of individuals in datasets. This is usually done by adding noise to individual data records before the input is fed into the model. The added noise makes it difficult to recover the raw data from model output. Fine-tuning self-supervised generative models with DP techniques bounds the probability that the generated output will not reveal the sensitive data with which the model was fine-tuned, providing a probabilistic data protection layer.

While DP techniques are mostly robust against unknown privacy attacks, a trade-off is to be made between privacy and accuracy. The trade-off depends on the parameters. In current GenAI models, it is still uncertain how sensitive enterprise data in training or in input could be used by the GenAI model such that it cannot be reconstructed by the outside world[17]. We expect, based on the formalised privacy definition described above, that the risk of data breach of GenAI will at some point be quantifiable and consequently insurable. How a clear mathematical definition will correspond with the legal vagueness of the concept "personally identifiable information" remains to be seen.

3.4 Intellectual property violations

Whereas the previously mentioned risks are relevant for AI models and GenAI alike, one of the risks novel to GenAI is the risk of intellectual property (IP) infringements by the generated output. The violation of IP rights can arise from infringements of copyrights, patents, trademarks, industrial design and trade secrets.

One of the ways in which GenAI can commit IP violations is if it is trained on original, licensed works, or the resulting works are insufficiently transformative from existing, protected works[18], and, as a result, are unauthorised derivative work[19]. A different type of risk that businesses using GenAI platforms face is the risk of accidentally sharing confidential trade secrets or business information by inputting that data into GenAI tools. An example is an incident by Samsung in April 2023, in which Samsung engineers accidentally leaked trade source codes while utilising ChatGPT[20].

One approach of mitigating IP infringement risks is to compare the likelihood of generating an output from the full training data with the likelihood of generating that output from just a part of the training data (excluding licensed works). If the likelihoods are similar, one could argue that there is little risk of copyright infringement. Following this idea, GenAI models can be trained with limited access to the copyrighted data in the training set[21], a so-called "near-access freeness" technique. Another approach is setting up a similarity-based algorithm to help retrieve the licensed materials that are most similar to the generated data. Using such algorithms for screening or alerting users whether generated output is too similar to licensed materials may help reduce the risks of IP infringement for GenAI.

The current legal uncertainty around IP Infringement of GenAI plays a big part in the adoption aversion of GenAI. Unanswered questions to date circle around how the courts will interpret the "fair use doctrine" of Section 107 of the Copyright Act, as well as uncertainty around which similarity metrics to pick and how these similarity metrics will uphold in front of a judge. These unanswered questions could prove to be expensive, especially if the use of the GenAI in and of itself is classified as wilful infringement. If a business user is aware that the training data could potentially include unlicensed works or that the works generated by the GenAI are probably not covered by the "fair use doctrine", a business could face charges of wilful infringement with statutory damages alone of up to US$ 150,000 for each instance (17 US Code § 504 (c)).

3.5 Producing harmful content

Individuals with malicious intent can utilise GenAI to generate deepfake pictures, voices, videos, texts, etc. for illegal activity. Though many of the larger systems like GPT-4 and Bard have built-in content moderation filters, they can still be coaxed into generating undesirable output[22], such as hate speech or instructions for how to build a bomb. In one recorded instance, researchers instructed ChatGPT to write antisemitic messages in a way that would not be detected and taken down by X (formerly Twitter). The researchers flagged the GenAI's suggestion to "avoid explicit language and instead use stereotypes or tweet support for individuals who are antisemitic". Even after adjustments of the content moderation filter, ChatGPT still responded to similar prompts negatively[23].

Fine-tuning tools in training data (such as protective usage policies) and teams monitoring "bad use cases" to flag undesired responses are demonstrated to reduce harmful content. However, undesired responses cannot be fully eliminated[34]. Insurance might become a viable part of the solution for covering the residual risk, depending on the technical safety strategy used in the GenAI application.

3.6 Other risks including environmental risks

Currently, GenAI models have billions of parameters, which need to be trained to improve the models' performance[24]. As the number of model parameters increases, the energy consumption for one training round rises correspondingly, generating high carbon emissions for a majority of the models. The research shows that the consumption required for training a transformer model using GPUs can lead to more than 626,000 pounds of carbon dioxide, almost five times the lifetime emissions of a car in the US[25]. Also, for water consumption, the training could "cost" 700,000 litres of clear freshwater, equal to the daily drinking water needs of 175,000 people[26]. With GenAI commercial usage growing, training/retraining the models to achieve better performance becomes more frequent. Also, with the model complexity growing, each training/retraining may require more time to converge to a suitable solution for the parameters. The cost could escalate and have a stronger influence on climate and environmental risks.

Recent developments have shown that some GenAI use cases can be ddressed with smaller models achieving a similar performance, such as Meta's LLaMa or DeepMind's Chinchilla. These advancements can help mitigate environmental risk while maintaining model performance and the positive impact GenAI can have on business and society.

Insurance providers could play a role in ensuring responsible model development. By establishing guidelines for balancing training frequency and energy and water consumption, insurance companies could aid in ensuring the right balance between continuous improvement of the GenAI models and an environmentally sustainable development.

4 Mitigation of risks of generative AI

With some of the risks outlined above, it becomes clear why AI has been referred to as a double-edged sword. And "while this can be said of most new technologies, both sides of the AI blade are far sharper, and neither is well understood [by its users]"[27]. Businesses wanting to mitigate the risks associated with AI can look at emerging technical solutions, some of them outlined above, that have been tried and proven valid by researchers in the field.

However, even the most accurate AI will produce wrong or misleading results from time to time, regardless of how extensive its technical improvements are. Munich Re with the aiSure™ products allows businesses to transfer the risk of model failure, model underperformance, discrimination caused by AI models, as well as other AI-related risks, to insurers – insuring GenAI providers and GenAI users against AI "going rogue".

After a quick summary of Munich Re's current AI insurance suite, the focus will be placed on insuring GenAI, and a brief insight into ways to insure the different risks of GenAI applications will be provided. As this is a nascent field surrounded by legal and regulatory uncertainty and corresponding limited loss data, the following section is intended to provide readers with a glimpse of how Munich Re thinks about insuring these risks.

4.1 How does Munich Re insure machine learning models?

Munich Re has been supporting clients in managing novel technology risks as they evolve for decades, ranging from the success of insuring rocket launches to safe storage of cryptocurrencies. This is why Munich Re has supported the AI community for many years by insuring trustworthy and reliable AI solutions in order to ease AI adoption for companies.

- **With aiSure™ – Contractual Liabilities** AI providers can prove the quality of their AI model and assure customers that their AI tool will perform as expected. If the AI does not deliver as promised, Munich Re backs the AI providers' performance guarantee and compensates customers for the losses incurred. This solution allows an AI vendor to e.g. guarantee that their fraud detection model will catch at least 99% of all fraudulent transactions. If the AI underperforms, Munich Re provides a payout amounting to the losses incurred.
  This insurance-backed performance guarantee increases the trust in AI and its adoption, while Munich Re's strong balance sheet carries the risk of the AI models underperforming.

- **With aiSure™ – Own Damages** businesses can insure the performance of their self-built ("home-grown") AI, allowing them to implement AI solutions for critical operational tasks, e.g. in manufacturing or agriculture. This solution insures e.g. a car manufacturer turning to AI for the final quality control before distributing the cars to their final sales location. Insuring the performance of the AI model protects the manufacturer from distributing subpar cars due to the error rate of their AI drifting beyond the desired threshold.
  This insurance solution allows businesses the worry-free implementation of AI models for vital parts of their operations. When their models underperform, businesses know that their financial downside is covered by Munich Re.

– **With aiSure™ – General Liabilities** businesses can protect themselves against damage and financial losses arising from lawsuits alleging that AI-made decisions discriminate against protected groups.
This solution insures e.g. a university that uses black-box AI for college applications against class action suits for alleged discrimination against protected groups. This insurance solution promotes equitable use of AI and shields businesses from expensive and far-reaching lawsuits alleging disparate impact discrimination.

In order to assess the risk of an AI model, Munich Re's AI team follows a proven technical due diligence process. We refer the interested reader to our whitepaper "De-Risking AI Ventures"[28]. With this approach, Munich Re is able to determine the predictive robustness of the AI, quantifying the probability and severity of model underperformance. The insurance premium is calculated depending on the robustness of the AI model.

4.2 How would Munich Re insure the risks introduced by generative AI?

As the first player in the market, Munich Re has been insuring the performance of traditional machine learning models since 2018. As such, the next step is transferring this expertise onto the question of how to best insure GenAI models.

Compared to traditional machine learning models, measuring the performance of GenAI models is less straightforward. Reasons are the differences in training (leading to more complex outcomes), the variety of tasks GenAI models are confronted with, differences in setup (most GenAI models are based on foundation models), and finally the subjectivity of the quality of these outcomes (judging the ground truth). Due to these difficulties, the underperformance of a GenAI application must be defined well in order to capture the essential differences instead of granular variations for a risk transfer to be meaningful.

Now, after contrasting the differences in measuring underperformance, the steps necessary to transfer the risks most akin to the underperformance risks of machine learning models – like hallucinations, false information and harmful content produced by GenAI – will be outlined. In a second step, thoughts around the insurance of discrimination and bias in GenAI will be explored. Finally, the AI risks that Munich Re considers serious, but momentarily surrounded by critical uncertainties – IP infringements, privacy violations and other risks including environmental risks – are listed, together with the market, legal and environmental changes necessary to better measure those risks.

4.2.1 Insuring GenAI against hallucinations, false information and harmful content

Traditional machine learning models are usually trained in a supervised fashion, i.e. minimising the difference between predicted values for given input data and target ground-truth labels. GenAI models follow an unsupervised or semi-supervised training, as generative tasks usually do not have a single "correct" output – instead they respond with "creativity", adding subjectivity and complexity to the output evaluation. Images generated from prompts can be incomplete, not detailed, contain hallucinations in the form of bizarre and fictional additions or simply might not be what the user was looking for. For text generated by AI, the outcomes are similarly hard to assess – and it could be that the hallucinations are not recognised until later.

Furthermore, when testing GenAI model performance in different professional and academic exams, it will excel in certain areas and fail in others when compared to human test takers. While GPT-4 received exceptional scores

in the verbal sections for the GRE and SATs (99th and 93rd percentile of all examinees), its results in quantitative sections were considerably lower and, finally, for competitive programming contests GPT-4 was unable to find solutions to complex problems[29]. As a result, using one GenAI model indiscriminately for a wide area of use cases makes it nearly impossible to accurately determine the constant performance threshold of the models for any given task.

Finally, the performance of GenAI will also depend on underlying model changes. Recent research[30] shows that GPT-4 performance degrades significantly over time, which may also have an impact on its many downstream applications. The degradation can be caused by updates of foundation models or fine-tuning based on new datasets. When and how updates are conducted is often an obscure process for the end-user of downstream applications.

The above-mentioned difficulties when determining the underperformance of GenAI impact the way they can be underwritten and insured. This requires insurance companies to amend their processes to include continuous monitoring and regularly updating insurance policies. Also, in order to be able to accurately and objectively determine underperformance, the threshold of the model underperforming should be set in an abstract way. This abstract threshold is given if the GenAI model hallucinates, spreads false information, or provides users with harmful content.

Hallucinations, false information and harmful content are all part of the performance risk of AI, as they are essentially deviations of model output from expected output. Similar to Munich Re's technical due diligence process for insuring AI models, when insuring GenAI against the above-mentioned risks, the model evaluation pipeline will be co-developed with the GenAI model providers to achieve the insurability of GenAI applications.

Some risk considerations for underperformance are: As GenAI model performance could be different for different tasks, the model evaluation should be tied to a single specified task. The GenAI application developers should define the model input space clearly, such as restricting the topics of queries and specifying the format of input. Furthermore, what is considered false information, hallucination and harmful content will need to be very clearly defined. The expected output does not have to be a single ground-truth output as they are usually not, but could instead be an encoded representation in the latent space of the model.

When it comes to the design testing regime, a complete testing regime should cover the whole input space (or at least have a representative sample of input). This requirement may be difficult for some GenAI applications where data is sensitive or proprietary, such as in medical, insurance or legal domains. In such cases, a continuous testing process that utilises human feedback or equivalent evaluation data could be applied with proper scrutiny. Based on the ground truth defined above and testing data collected in the above-described testing regime, the hallucination, false information and harmful content issues can be transformed to a classical machine learning problem: If damage occurs (legal liabilities, etc.) and the model shows underperformance based on the defined performance metrics and thresholds, the insurance will cover the damage incurred.

Finally, for downstream GenAI applications utilizing the API of foundation models, in order to account for the risks introduced by the fact that the underlying foundation model is not maintained by the GenAI application provider themselves, Munich Re requires the model providers to follow higher standards of continuous monitoring and model improvement. In order to manage the risks, the following modifications to the policies in place can be made:

– Pause guarantee or insurance policy when significant performance degrade is expected or detected in the foundation model.

– Adjust guarantee threshold to keep the risks constant and guarantees in place.

– Adjust guarantee or insurance fees to match the fluctuating risk exposure.

If the above processes and requirements are fulfilled, Munich Re can cover GenAI models against hallucinations, false information and harmful content.

4.2.2 Insuring GenAI against model bias and fairness

Evaluating the risk of GenAI giving biased outcomes comes with its own set of challenges. As pointed out above, many fairness metrics have been established by researchers, some of them yielding mutually exclusive results. Moreover, those fairness metrics used by researchers and computer scientists are not necessarily congruent to the common legal definitions of discrimination of protected groups. This adds the complexity that even if the GenAI model performs well regarding the chosen fairness model, lawsuits alleging discrimination of protected groups can still prove successful. Finally, if a discriminatory model is being used by a variety of companies in a consumer-facing application, its discriminatory decisions can lead to sizeable losses for the insurer, as class action lawsuits for discrimination will be the norm rather than an exception.

Bearing those risks in mind, Munich Re believes that the fairness risk of GenAI models can be evaluated and insured under certain guardrails. Firstly, a fairness metric needs to be determined and agreed upon. This fairness metric should closely align with the goal of the GenAI application and should encompass the sensitive groups or individuals that need to be protected in the domain of application. Secondly, the metric threshold will need to be defined. Here, considerations around the trade-off between fairness and model accuracy, statistical randomness in the data, and impacts of disparity in the context of the use case should be taken into account. The tolerance of disparity will need to be thoroughly assessed and documented.

As litigation progresses, court decisions around the equitability requirement for AI models and the usefulness of the performance in certain fairness metrics as valid evidence for court will aid further in determining the frequency and severity of model bias risk and inform the quantitative assessment of the discrimination risk.

Furthermore, prospective regulations determining the areas in which AI shall not be allowed due to its propensity to discriminate, as well as conditions imposed for safer usage, will aid in developing the insurance space around insuring discrimination of GenAI.

4.2.3 Insuring GenAI against IP and privacy violations

Evaluating the risks of IP infringements by GenAI is even more challenging, due to the current uncertainties in litigation around the proper use of licensed images in training data and the implications for generated images clearly based on these licensed images. In order to measure whether an image is based on a licensed image, their “similarity” can be determined. Researchers use a variety of different ways to measure the “similarity” of two images, detached from the legal interpretations of IP infringement.

A related issue arises when insuring AI models against unlawfully collecting private data and violating privacy laws. The US Department of Labor defines Personal Identifiable Information (PII) as "information that can be used to distinguish or trace an individual's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual" (OMB M-10-23, 2019). While these broad definitions work very well in the court systems, allowing the interpretations to evolve, they make it harder to assign PII infringement probabilities to GenAI models.

When looking to quantify the risks of GenAI models infringing on IP rights and committing data breaches today, the insurer and the insured have to agree on a more narrow definition that describes the infringement or violation compared to the legal definitions above. If this agreement was successful, there are special training techniques (such as Differential Privacy or Near Access-Free) that can make the risks quantifiable in a probabilistic way. If a model is trained in such a way, Munich Re can provide an insurance solution based on the probabilities of violations.

For a more generally trained GenAI model, the evaluations can be made based on the performance of a representative testing sample. For example, to evaluate the privacy infringement risks, one can simulate a large amount of privacy attacks (e.g. membership inference, reconstruction attacks, prompt-injection attacks), and get the success rate of such attacks. This can be used for estimating the loss frequency of privacy infringement. To evaluate the IP infringement risks, one can also check the likelihood of generating output that is similar to the copyrighted data based on a representative sample of inputs. This evaluation, however, depends on the similarity metric and the thresholds that were chosen. Hence, for the given privacy attacks or similarity metrics, it is possible to obtain the frequency of violations and insure the risks.

As more and more court decisions are made concerning AI-induced IP infringements and data breaches, more legal certainty will be created, which will in turn allow for more precise and all-encompassing evaluation methods, including the type of attacks and metrics. Over time, as the uncertainty in law is lifted, Munich Re aims to extend the insurance from a performance-based insurance solution to a full liability insurance solution.

4.2.4 Other risks

As environmental impacts and other risks are still being explored, including their meaning for society, Munich Re is currently not insuring these risks. However, as these risks evolve and the wider GenAI risk landscape becomes clearer, Munich Re will continue to co-develop risk transfer solutions with its clients.

4.2.5 Impact on traditional insurance

Finally, after outlining the novel risks that GenAI introduces, and presenting specialised insurance policies to transfer these risks to insurance companies, it is worth briefly mentioning the impact of GenAI – and more broadly all machine learning models – on traditional insurance policies.

To this day, AI risks are seldom excluded in traditional insurance policies. Therefore, the damage that AI models cause could be covered by traditional insurance policies. Examples that come to mind are AI-based machinery injuring bystanders (could be covered by existing general liability policies), AI models that are hacked (could be covered by existing cyber insurance policies), AI-based cleaning robots that destroy property (could be covered by existing property insurance policies), AI models that make biased employment decisions (could be covered by existing EPLI policies), and many more.

Due to AI impacting all "walks of life", there is a lot of partial coverage from existing insurance policies, making it difficult for both insurer and insured to have full confidence on the extent of the coverage – potentially leading to over- or underinsurance. Furthermore, the lack of conclusiveness on the extent of the coverage could lead to devastating effects for insurers, as the threat of "silent AI" might be underrated, as well as for insureds, as they might be left without financial protection. AI exposures within traditional insurance policies could represent a significant unexpected risk to insurers' portfolios, as the risks of AI underperforming – potentially even systematically – were not considered in the pricing of the insurance.

When using AI, insureds should therefore be aware of potential insurance gaps, leaving them exposed to risks caused by their AI models. Insurers, on the other hand, should be aware of the risks that AI poses to their existing policies and should monitor the risk of silent AI exposure.

5 Outlook

As once famously said by Henry Ford when talking about New York City: "Without insurance we would have no skyscrapers, because no man would dare to work at such heights, at the risk of killing himself and leaving his family destitute. Without insurance, no businessman would invest his millions in constructing a building like this, when a single spark could reduce it to ashes"[31]. We are still at the beginning of fully encompassing all AI risks, their dependence, their systematic nature and their geographical spread[32].

Similar to the quote above, the risks of GenAI as well as their systematic nature could prove devastating for AI providers or AI users who come to face a class action suit in the US with many damaged parties. We at Munich Re believe that insurance could be the right vehicle to mitigate those risks. Pooling the risks of GenAI going wrong enables active innovation and growth by reducing the risk costs for single GenAI applications and allowing companies to focus on further pushing the technological barriers in GenAI without worrying about the residual financial risks. Furthermore, through the risk assessment function of insurance, much-needed industry standards could be developed, making the AI environment safer without unnecessary overregulation. This creates further trust in the market, allowing us to fully harness the power of AI and GenAI for society and businesses.

The emergence of GenAI has undoubtedly revolutionised industries, offering transformative opportunities while introducing novel risks. With clear guardrails set into place for the use cases, continuous performance monitoring and clear metrics for the model to be measured against, insuring GenAI will be possible as it is insuring prediction performance of ML models at Munich Re. As outlined above, some of the risks are more easily insurable (and quantifiable) than others.

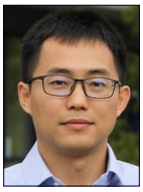
We believe that with the rise of these new and often complex technology risks, a broad demand for insuring GenAI will arise. At Munich Re, we are ready for the demand and are looking forward to exploring the reality of insuring GenAI with corporate, broker and primary insurance partners.

Do you have questions? Please reach out.

6 Contact

Iris Devriese
Underwriter and
Business Development
Manager at Munich Re

Yuanyuan Li
Research Scientist
at Munich Re



Yang Lin
Research Scientist
at HSB

7 References

A
Andersen, McKernan, Ortiz v. Stability AI, 3:23 - cv p 00201 (US District Court Northern District of California January 13, 2023).
Appel, G., Neelbauer, J., & Schweidel, D. A. (2023, April 07). Generative AI Has an Intellectual Property Problem. Harvard Business Review.

B
Bohannon, M. (2023, June 08). Lawyer Used ChatGPT In Court - And Cited Fake Cases. A Judge Is Considering Sanctions. Forbes.
Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., & ... & Liang, P. (2021). On the Opportunities and Risks of Foundation Models. arXiv preprint, arXiv:2108.07258.
Buchanan, B. (2020). The AI Triad and What It Means for National Security Strategy. Washington, D.C.: Center for Security and Emerging Technology.
Bureau of Competition; Office of Technology. (2023, June 29). Generative AI Raises Competition Concerns. Retrieved from Federal Trade Commission: https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns
Burgo, R., & Hughes, W. (2023, August 17). EOC Settles First-Ever AI Discrimination Lawsuit. SHRM.

C
Cheatham, B., Javanmardian, K., & Samandari, H. (2019, April 26). Confronting the risks of artificial intelligence. McKinsey Quarterly.
Chen, M. (2023, January 24). Artists and Illustrators Are Suing Three A.I. Art Generators for Scraping and 'Collaging" Their Work Without Consent. artnet news.
Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharvsky, A., . . . Zemmel, R. (2023, June). The economic potential of generative AI: the next productivity frontier. McKiney Digital.
Civillini, M. (2023, June 08). World Bank set to take on risk of insuring carbon credits amid market upheaval. Climate Home News.

E
Eling, M. (2019, November 7). How insurance can mitigate AI risks. Retrieved from Brookings.edu: https://www.brookings.edu/articles/how-insurance-can-mitigate-ai-risks/

G
Getty Images v. Stability AI , 1:23-cv-00135-UNA (US District Court for the District of Delaware February 03, 2023).
Goldman Sachs. (2023, April 05). Generative AI could raise global GDP by 7%. Retrieved from goldmansachs.com: https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html
Gurman, M. (2023, May 01). Samsung Bans ChatGPT, Google Bard, Other Generative AI Use by Staff After Leak. Bloomberg. Retrieved from https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak

J
Jane Doe v Netflix, Inc., C09 05903 JW PVT (US District Court for the Northern District of California December 17, 2009).
Jimenez, J. H. (2023, May 12). Insurance, the necessary driver for the development of society. Retrieved from mapfre.com: https://www.mapfre.com/en/insights/insurance/insurance-necessary-driver-development-society/

K
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint, arXiv:1609.05807.

L
L., C., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? arXiv, arXiv:2307.09009.
LayerX. (2023). Revealing the true GenAI Data Exposure Risk. LayerX.
Li, P., Yang, J., Islam, M., & Ren, S. (2023). Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. arXiv preprint arXiv:2304.03271.

M
Massachusetts Institute of Technology. (2023). The great acceleration: CIO perspectives on generative AI. MIT Technology Review.
Mata v. Avianca, Inc., 1:22-cv-01461 (District Court, S.D. New York May 25, 2023).
Maxwell, T. (2023, March 28). Before releasing GPT-4, OpenAI's 'red team' asked the ChatGPT model how to murder people, build a bomb, and say antisemitic things. Read the chatbot's shocking answers. Insider.Memarrast, O., Vu, L., & Ziebart, B. D. (2023). Superhuman Fairness. International Conference on Machine Learning, pp. 24420-24435.

N
Nazneen Rajani, H. L. (2023). Responsible Generative AI Tutorial in ICML 2023. Retrieved from https://sites.google.com/view/responsible-gen-ai-tutorial/

O
Office of Public Affairs. (2023, January 09). Justice Department and Meta Platforms Inc. Reach Key Agreement as They Implement Groundbreaking Resolution to Address Discriminatory Delivery of Housing Advertisements. Retrieved from U.S. Department of Justice: https://www.justice.gov/opa/pr/justice-department-and-meta-platforms-inc-reach-key-agreement-they-implement-groundbreaking
OMB M-10-23. (2019). GSA Rules of Behavior for Handling Personaly Identifyable Information. GSA Directive CIO P 2180.2.
OpenAI. (2003, March 27). GPT-4 Technical Report. Retrieved from OpenAI.com: https://arxiv.org/pdf/2303.08774.pdf
OpenAI. (2023, March 23). GPT-4 System Card. Retrieved from OpenAI.com: https://cdn.openai.com/papers/gpt-4-system-card.pdf
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

R
Russel, J. (2023, June 08). @jruss_jruss. Retrieved from twitter.com: https://twitter.com/jruss_jruss/status/1666865275436957696
Ryan-Mosley, T. (2023, May 15). Catching bad content in the age of AI. MIT Technology Review.

S
Singel, R. (2009, December 17). Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims. WIRED.
Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv preprint arXiv:1906.02243.

U
U.S. Equal Employment Opportunity Commission. (2023, 05 18). EEOC Releases New Resource on Artificial Intelligence and Title VII. Retrieved from eeoc.gov: https://www.eeoc.gov/newsroom/eeoc-releases-new-resource-artificial-intelligence-and-title-vii#:~:text=%E2%80%9CAs%20employers%20increasingly%20turn%20to,Burrows.

V
Vyas, N., Kakade, S., & Barak, B. (2023). Provable copyright protection for generative models. arXiv preprint, arXiv:2302.10870.

1 (Massachusetts Institute of Technology, 2023)
2 Chui, et al., 2023)
3 (Goldman Sachs, 2023)
4 (Bureau of Competition; Office of Technology, 2023)
5 (Mata v. Avianca, Inc., 2023).
6 (Bohannon, 2023)
7 (Russel, 2023)
8 (Ouyang, et al., 2022)
9 (Office of Public Affairs, 2023)
10 (Burgo & Hughes, 2023)
11 (Memarrast, Vu, & Ziebart, 2023)
12 (Kleinberg, Mullainathan, & Raghavan, 2016)
13 (U.S. Equal Employment Opportunity Commission, 2023)
14 (LayerX, 2023)
15 (Jane Doe v Netflix, Inc., 2009)
16 (Singel, 2009)
17 (Nazneen Rajani, 2023)
18 (Getty Images v. Stability AI , 2023) (Andersen, McKernan, Ortiz v. Stability AI, 2023) (Chen, 2023)
19 (Appel, Neelbauer, & Schweidel, 2023)
20 (Gurman, 2023)
21 (Vyas, Kakade, & Barak, 2023)
22 (Ryan-Mosley, 2023)
23 (Maxwell, 2023)
24 (Buchanan, 2020)
25 (Strubell, Ganesh, & McCallum, 2019)
26 (Li, Yang, Islam, & Ren, 2023)
27 (Cheatham, Javanmardian, & Samandari, 2019)
28 https://www.munichre.com/en/solutions/for-industry-clients/insure-ai/de-Risking-ai-ventures.html
29 (OpenAI, 2003)
30 (L., Zaharia, & Zou, 2023)
31 (Jimenez, 2023)
32 (Eling, 2019)
33 (Bommasani, et al., 2021)
34 (OpenAI, 2023)

© 2024
Münchener Rückversicherungs-Gesellschaft
Königinstrasse 107, 80802 München, Germany

Picture credits: Munich Re

Münchener Rückversicherungs-Gesellschaft (Munich Reinsurance Company) is a reinsurance company organised under the laws of Germany. In some countries, including in the United States, Munich Reinsurance Company holds the status of an unauthorised reinsurer. Policies are underwritten by Munich Reinsurance Company or its affiliated insurance and reinsurance subsidiaries. Certain coverages are not available in all jurisdictions.

Any description in this document is for general information purposes only and does not constitute an offer to sell or a solicitation of an offer to buy any product.