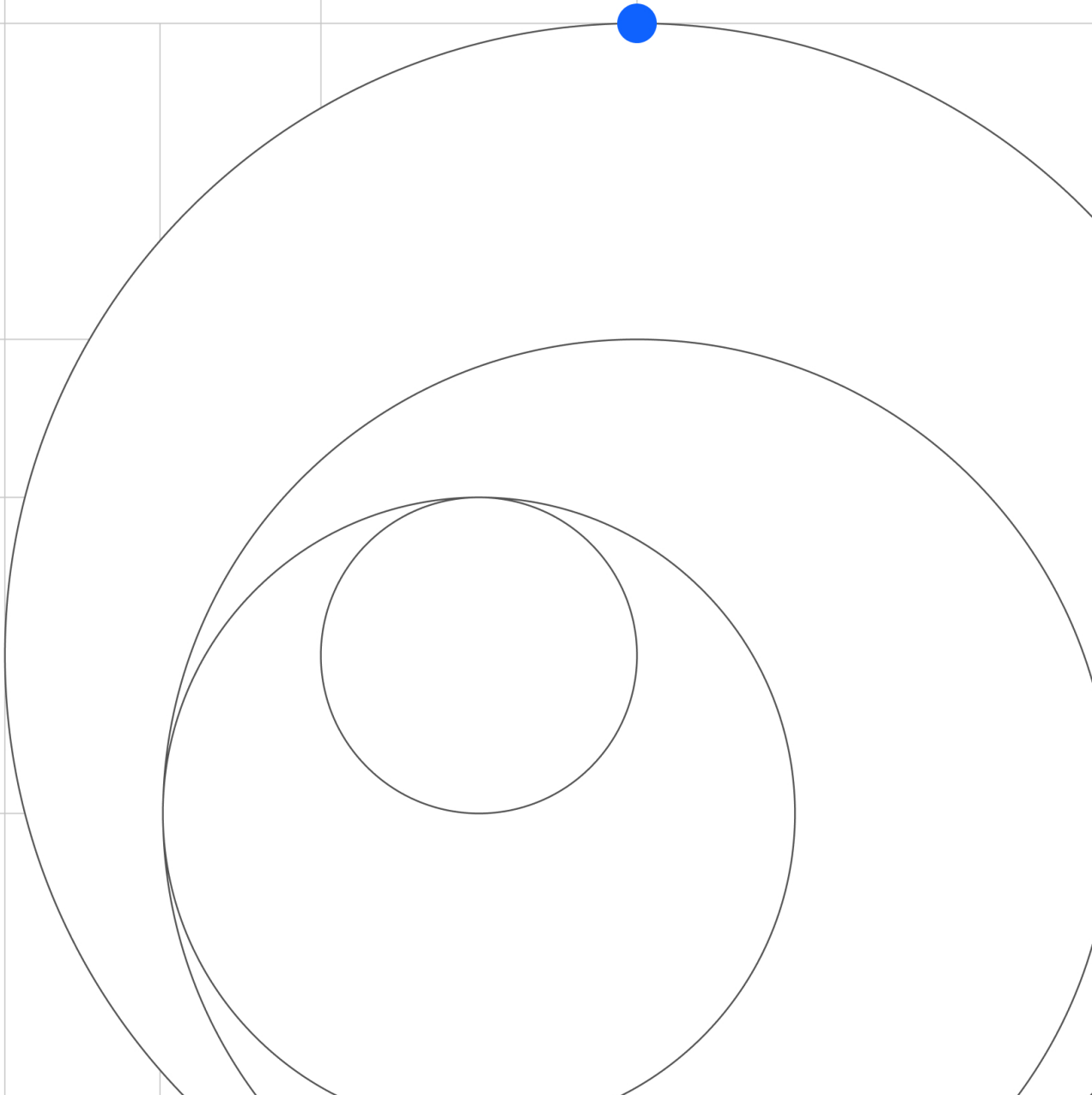
IBM AI Ethics Board

Foundation models: Opportunities, risks and mitigations



IBM

Attribution

With gratitude to the AI Ethics Board workstream's executive sponsors, Christina Montgomery and Francesca Rossi, and the contributions of workstream members Betsy Greytok, Bryan Bortnick, Catherine Quinlan, David Piorkowski, Eniko Rozsa, Heather Domin, Heather Gentile, Jamie VanDodick, Jill Maguire, John McBroom, Joshua New, Justin Weisz, Katherine Fick, Kevin Black, Kush Varshney, Manish Bhide, Manish Goyal, Melis Kiziltay, Michael Epstein, Micheal Hind, Milena Pibric, Phaedra Boinodiris, Rogerio Abreu de Paula, Saishruthi Swaminathan, and Suj Perapa.

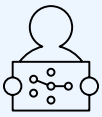
Table of contents

Executive summary

The rise of foundation models offers enterprises exciting new possibilities but also raises new and expanded questions about their ethical design, development, deployment and use. According to a recent IBM Institute for Business Value generative AI survey, organizations are already expressing concerns about trust-related issues—specifically as barriers to investment. Their top concerns are cybersecurity (57%), privacy (51%) and accuracy (47%). Many organizations were taking these concerns seriously before the *consumerization* of generative AI, expressing their intent to invest at least 40% more in AI ethics over the next three years. Awareness about risks and possible ways to mitigate them is the first crucial step toward building trustworthy AI systems.

In this document we:

Explore the benefits of foundation models, including their capability to perform challenging tasks, potential to speed up the adoption of AI, ability to increase productivity and the cost benefits they provide.



Discuss the three categories of risk, including risks known from earlier forms of AI, known risks amplified by foundation models and emerging risks intrinsic to the generative capabilities of foundation models.

Cover the principles, pillars and governance that form the foundation of IBM's AI ethics initiatives and suggest guardrails for risk mitigation.

Introduction

As the use of AI continues to expand, large and complex AI models are delivering promising performance results, as well as solving some of society's most challenging problems. However, building large training data sets and complex models for each AI application can be burdensome for enterprises. Foundation models provide a path to achieve the best of both worlds: build powerful state-of-the-art models and reuse them directly or apply tuning methods to implement a variety of use cases, rather than train new models for each use case. For example, IBM Research® developed foundation models for visual inspection. These foundation models learn the general representation of concrete surfaces and runways and can be further tuned for specific use cases like crack detection or defect inspection with less labeled data.

IBM defines a *foundation model* as an AI model that can be adapted to a wide range of downstream tasks. Foundation models are typically large-scale generative models that are trained on unlabeled data using self-supervision. As large-scale models, foundation models can include billions of parameters.

IBM is a hybrid cloud and AI company with a long reputation as a responsible data steward committed to AI ethics. Using the strength of our research, product and consulting teams, along with external partners, such as Hugging Face, we help bring the power of foundation models to our clients and build trustworthy AI across any enterprise. IBM also continues to invest in building new platforms, such as the IBM® watsonx™ AI and data platform and technologies, for designing and developing AI models to behave in an auditable and trustworthy manner.

This document describes the point of view of IBM on the ethics of foundation models. It is the first version, and future versions will expand on various aspects of IBM's foundation model ethics approach. We hope this document is helpful for all stakeholders in developing, deploying and using the foundation model in a responsible way.

Benefits of foundation models

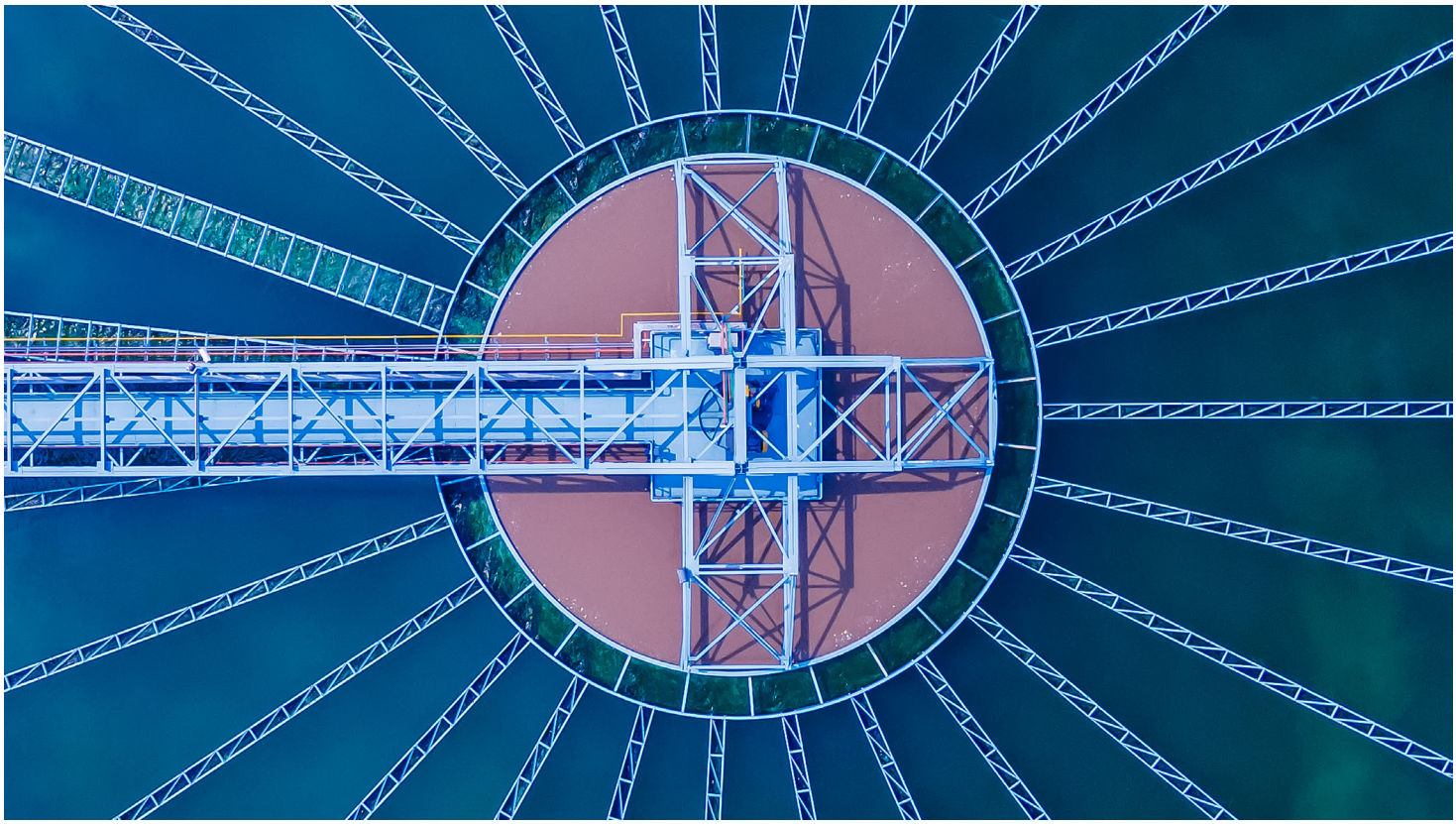
Foundation models can significantly improve the process of developing AI systems and help advance AI from the exploration to the adoption phase in enterprises. Their benefits include:

Performing complex tasks
Foundation models show a significant increase in performance in solving difficult and complex problems. For example, the geospatial foundation model from the IBM and NASA collaboration is designed to convert NASA's satellite data into maps of natural disasters like floods and other landscape changes. The model could also be used to help reveal our planet's past; estimate risks to crops, businesses or infrastructures due to severe weather; develop strategies to adapt to climate change; and assist agribusiness. The model is planned to be made available in preview to IBM clients through the IBM Environmental Intelligence Suite.

As another example, IBM's MoLFormer-XL is a foundation model that infers the structure of molecules from simple representations and makes it easy to learn various downstream tasks like predicting a molecule's physical and quantum properties, identifying similar molecules, screening already approved molecules for new use cases, and discovering new molecules. Moderna and IBM are exploring ways to use MoLFormer to help predict molecule properties and understand the characteristics of potential mRNA medicines.

Increased productivity
The generative nature of foundation models expands the number of areas where AI can be used in an enterprise to help improve productivity by automating routine and tedious tasks and allowing users to spend more time on creative and innovative work. For example, IBM Watson® Code Assistant, powered by foundation models, enables developers of all experience levels to write code using AI-generated recommendations.

Quicker time to value
Foundation models are usually trained with unlabeled data, which is more accessible in larger quantities than labeled data. Once trained, foundation models can be used either directly or after being tuned for downstream applications, using a small amount of specialized labeled data, which can decrease time-to-value creation.



Utilize diverse data modalities
Foundation models may be trained using various data modalities, such as natural language, text, image and audio. They can also be applied to tasks requiring different data types, such as time series data, geospatial data, tabular data, semi-structured data and mixed-modality data like text combined with images.

Amortized expenses
Although the initial cost of training a foundation model is significantly higher than training a traditional AI model, the incremental cost of applying it to a new task is considerably lower. Using pretrained foundation models could help eliminate the requirement that enterprises make substantial investments to train foundation models to experiment with their new capabilities. For an enterprise, the trustworthiness of the models, energy efficiency, performance, portability and the ability to use enterprise data effectively and securely are paramount.

IBM allows enterprises to create and own the value of foundation models for their business by bringing the best innovations from the open, global AI community, running efficiently in hybrid computing environments, helping mitigate risks, and rigorously governing AI.

Risks of foundation models

Like all rapidly advancing technologies, foundation models have risks along with benefits. Some are legal risks, for example, restrictions on moving or using data, and need to be carefully evaluated under current and evolving law. Other risks have an ethical nature and must be considered carefully so that the technology has a positive impact. In general, AI risks raise sociotechnical questions and should be addressed and mitigated through sociotechnical methods, including software tools, risk assessment processes, AI ethics frameworks, governance mechanisms, multistakeholder consultations, standards and regulation. We will list the risks by considering the following 3 categories:

1. **Traditional.** Known risks from prior or earlier forms of AI systems
2. **Amplified.** Known risks but now intensified because of intrinsic characteristics of foundation models, most notably their inherent generative capabilities
3. **New.** Emerging risks intrinsic to foundation models and their inherent generative capabilities

We also structure the list of risks in relation to whether they're mostly associated with content provided to the foundation model —the input — or the content generated by it — the output — or if they're related to additional challenges.

1. Risks associated with input

Training and Tuning Phase

Group	Risk	Why is this a concern?	Indicator
Fairness	Data bias: Historical, representational, and societal biases present in the data used to train and fine tune the model.	Training an AI system on data with bias, such as historical or representational bias, could lead to biased or skewed outputs that can unfairly represent or otherwise discriminate against certain groups or individuals. In addition to negative societal impacts, business entities could face legal consequences, disruption to operation, or reputational harms from biased model outcomes.	Amplified
Robustness	Data poisoning: a type of adversarial attack where an adversary or malicious insider injects intentionally corrupted, false, misleading, or incorrect samples into the training or fine-tuning dataset.	Poisoning data can make the model sensitive to a malicious data pattern and produce the adversary's desired output. It can create a security risk where adversaries can force model behavior for their own benefit. In addition to producing unintended and potentially malicious results, a model misalignment from data poisoning can result in business entities facing legal consequences, disruption to operations, or reputational harms.	Traditional
Value Alignment	Data curation: When training or tuning data is improperly collected or prepared.	Improper data curation can adversely affect how a model is trained, resulting in a model that does not behave in accordance with the intended values. Examples of improper data curation could include labeling or annotation errors in the data used for training or tuning the model. Correcting problems after the model is trained and deployed might be insufficient for guaranteeing proper behavior. Improper model behavior can result in business entities facing legal consequences, disruption to operations, or reputational harms.	Amplified
	Downstream-based retraining: Using undesirable (inaccurate, inappropriate, user's content, etc.) output from downstream applications for re-training purposes.	Repurposing downstream output for re-training a model without implementing proper human vetting increases the chances of undesirable outputs being incorporated into the training or tuning data of the model, possibly generating even more undesirable output. Improper model behavior can result in business entities facing legal consequences or reputational harms. Failing to comply with data transfer laws might result in fines and other legal consequences.	New
Data Laws	Data transfer: Law and other restrictions can limit or prohibit transferring data.	Data transfer restrictions can impact the availability of the data required for training an AI model and can lead to poorly represented data. In addition to impact on data availability, failure to comply with data transfer laws and regulations might result in fines and other legal consequences.	Traditional
	Data usage: Law and other restrictions can limit or prohibit the use of some data for specific AI use cases.	Failing to comply with data usage laws and regulations might result in fines and other legal consequences.	Traditional
	Data acquisition: Laws and other regulations might limit the collection of certain types of data for specific AI use cases.	Failing to comply with data acquisition laws and regulations might result in fines and other legal consequences.	Amplified

Group	Risk	Why is this a concern?	Indicator
Intellectual Property	Data usage rights: Terms of service, copyright laws, license compliance, or other IP issues may restrict the ability to use certain data for building models.	Laws and regulations concerning the use of data to train AI are unsettled and can vary from country to country, which creates challenges in the development of models. If data usage violates rules or restrictions, business entities might face fines, reputational harms, disruption to operations, and other legal consequences.	Amplified
Transparency	Data Transparency: Challenge in documenting how a model's data was collected, curated, and used to train a model.	Data transparency is important for legal compliance and AI ethics. Missing information limits the ability to evaluate risks associated with the data. The lack of standardized requirements might limit disclosure as organizations protect trade secrets and try to limit others from copying their models.	Amplified
	Data Provenance: Challenge around standardizing and establishing methods for verifying where data came from.	Not all data sources are trustworthy. Data might have been unethically collected, manipulated, or falsified. Using unreliable data can result in undesirable behaviors in the model. Business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	Amplified
Privacy	Personal information in data: Inclusion or presence of personal identifiable information (PII) and sensitive personal information (SPI) in the data used for training or fine tuning the model.	If not properly developed to protect sensitive data, the model might expose personal information in the generated output. Additionally, personal, or sensitive data must be reviewed and handled in accordance with privacy laws and regulations. Business entities could face fines, reputational harms, disruption to operations, and other legal consequences if found in violation.	Traditional
	Reidentification: Even with the removal or personal identifiable information (PII) and sensitive personal information (SPI) from data, it might still be possible to identify persons due to other features available in the data.	Data that can reveal personal or sensitive information must be reviewed with respect to privacy laws and regulations, as business entities could face fines, reputational harms, disruption to operations, and other legal consequences if found in violation.	Traditional
	Data privacy rights: Challenges around the ability to provide data subject rights such as opt-out, right to access, right to be forgotten.	The identification or improper usage of data could lead to violation of privacy laws. Improper usage or a request for data removal could force organizations to retrain the model, which is expensive. In addition, business entities could face fines, reputational harms, disruption to operations, and other legal consequences if they fail to comply with data privacy rules and regulations.	Amplified
	Informed consent: Data collected for training AI models without the owner's informed consent even when it is legally permitted to do so.	Under certain circumstances, it might be unethical to collect and use data without the person's consent. There are also possible reputational risks to such use.	Traditional

Inference Phase

Group	Risk	Why is this a concern?	Indicator
Privacy	Personal information in prompt: Disclosing Personal Information or Sensitive Personal Information as a part of prompt sent to the model.	Prompt data might be stored or later used for other purposes like model evaluation and retraining. These types of data must be reviewed with respect to privacy laws and regulations. Without proper data storage and usage business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
Intellectual Property	IP information in prompt: Disclosing copyright information or other IP information as a part of the prompt sent to the model.	Prompt data might be stored or later used for other purposes like model evaluation and retraining. These types of data must be reviewed with respect to IP laws and regulations. Without proper data storage and usage business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
	Confidential data in prompt: Inclusion of confidential data as a part of the prompt sent to the model.	If not properly developed to secure confidential data, the model might expose confidential information or IP in the generated output. Additionally, end users' confidential information might be unintentionally collected and stored.	New
Robustness	Evasion attack: attempt to make a model output incorrect results by perturbing the data sent to the trained model.	Evasion attacks alter model behavior, usually to benefit the attacker. If the output results are not properly accounted for, business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	Amplified
	Prompt-based attacks: Adversarial attacks such as prompt injection (attempt to force a model to produce unexpected output), prompt leaking (attempts to extract a model's system prompt), jailbreaking (attempts to break through the guardrails established in the model), and prompt priming (attempt to force a model to produce an output aligned to the prompt).	Depending on the content revealed, business entities could face fines, reputational harm, disruption to operations, and other legal consequences.	New

2. Risks associated with output

Group	Risk	Why is this a concern?	Indicator
Fairness	Output bias: Generated content might unfairly represent certain groups or individuals.	Bias can harm users of the AI models and magnify existing discriminatory behaviors. Business entities can face reputational harms, disruption to operations, and other consequences.	New
	Decision bias: When one group is unfairly advantaged over another due to effect of decisions made by human using the model output.	Bias can harm persons affected by the decisions of the model. Business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	Traditional
Intellectual Properly	Copyright infringement: When a model generates content that is too similar or identical to existing work protected by copyright or covered by open-source license agreement.	Laws and regulations concerning the use of content that looks the same or closely similar to other copyrighted data are largely unsettled and can vary from country to country, providing challenges in determining and implementing compliance. Business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
Value Alignment	Hallucination: Generation of factually inaccurate or untruthful content.	False output can mislead users and be incorporated into downstream artifacts, further spreading misinformation. This can harm both owners and users of the AI models. Also, business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
	Toxic output: When the model produces hateful, abusive, and profane (HAP) or obscene content.	Hateful, abusive, and profane (HAP) or obscene content can adversely impact and harm people interacting with the model. Also, business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
	Dangerous advice: When a model provides advice without having enough information, resulting in possible danger if the advice is followed.	A person might act on incomplete advice or worry about a situation that is not applicable to them due to the overgeneralized nature of the content generated.	New
Misuse	Spreading disinformation: Using a model to create misleading or false information to deceive or influence a targeted audience.	Spreading disinformation might affect a human's ability to make informed decisions. Business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
	Toxicity: Using a model to generate hateful, abusive, and profane (HAP) or obscene content.	Toxic content might negatively affect the well-being of its recipients. Business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
	Nonconsensual use: Using a model to imitate people through video (deepfakes), images, audio, or other modalities without their consent.	Deepfakes can spread disinformation about a person, possibly resulting in negative impact on the person's reputation. Business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	Amplified

Group	Risk	Why is this a concern?	Indicator
	Dangerous use: Using a model with the sole intention of harming people.	Business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
	Non-disclosure: Not disclosing that content is generated by an AI model.	Not disclosing the AI-authored content can be viewed as deceptive resulting in a decrease in trust. Intentional deception might result in decreased human agency, fines, reputational harms, and other legal consequences.	New
	Improper Usage: Using a model for a purpose the model was not designed for.	Reusing a model without understanding its original data, design intent, and goals might result in unexpected and unwanted model behaviors.	Amplified
Harmful Code generation	Harmful code generation: Models may generate code that, when executed, causes harm or unintentionally affects other systems.	The execution of harmful code might open vulnerabilities in IT systems. Business entities could face fines, reputational harms, disruption to operations, and other legal consequences.	New
Misplaced Trust	Over/under reliance: when a person places too little or too much trust in an AI model's guidance.	In tasks where humans make choices based on AI-based suggestions, over/under reliance can lead to poor decision making because of the misplaced trust in the AI system, with negative consequences that increase with the importance of the decision. Bad decisions can harm people and can lead to financial harm, reputational harm, disruption to operations, and other legal consequences for business entities.	Amplified
Privacy	Exposing Personal information: When personal identifiable information (PII) or sensitive personal information (SPI) are used in the training data, fine-tuning data, or as part of the prompt, models might reveal that data in the generated output.	Sharing people's PI impacts their rights and make them more vulnerable. Also, output data must be reviewed with respect to privacy laws and regulations, as business entities could face fines, reputational harms, disruption to operations, and other legal consequences if found in violation of data privacy or usage laws.	New
Explainability	Unexplainable output: Challenges in explaining why model output was generated.	Foundation models are based on complex deep learning architectures, making explanations for their outputs difficult. Without clear explanations for model output, it is difficult for users, model validators, and auditors to understand and trust the model. Lack of transparency might carry legal consequences in highly regulated domains. Wrong explanations might lead to over-trust.	Amplified
Traceability	Unreliable attribution of sources: Challenges in determining from what training or fine-tuning data the model generated a portion or all its output.	Inability to trace output's source or provenance makes it difficult for users, model validators, and auditors to understand and trust the model.	New

3. Challenges

Group	Risk	Why is this a concern?	Indicator
Governance	Model Transparency: Lack of model transparency or insufficient documentation of the model development process makes it difficult to understand how and why a model was built and who built it, thus increasing the possibility of model unintended misuse.	Transparency is important for legal compliance, AI ethics, and guiding appropriate use of models. Missing information might make it more difficult to evaluate risks, to change the model, or reuse it. Knowledge about who built a model can also be an important factor in deciding whether to trust it.	Traditional
	Accountability: The foundation model development process is complex with lots of data, processes, and roles. When model output does not work as expected it can be difficult to determine the root cause and assign responsibility.	Without properly documenting decisions and assigning responsibility, determining liability for unexpected behavior or misuse might not be possible.	Amplified
Legal Compliance	Legal accountability: Determining who is responsible for the foundation model.	If ownership or responsibility for development of the model is uncertain, regulators and others may have concerns about the model because it will not be clear who is - or should be - liable/responsible for problems with it or can answer questions about it. Users of models without clear ownership may find challenges with compliance with future AI regulation.	New
	Generated Content Ownership: Determining ownership of AI generated content.	Laws and regulations that relate to the ownership of AI-generated content are largely unsettled and can vary from country to country. Business entities might face fines, reputational risks, disruption to operations, and other legal consequences.	New
	Generated Content IP: Legal uncertainty about intellectual property rights related to generated content.	Laws and regulations about determining of copyrightability, and patentability of the AI-generated content are largely unsettled and can vary from country to country. Business entities might face fines, reputational risks, disruption to operation, and other legal consequences if the generated content is covered by IP rights.	New
	Source attribution: Determining provenance of the generated content.	If the model generates an output that is identical to data used to train the model, it should give provenance of that output. Failure to do so may put the business entities deploying or using the model at legal risk.	Amplified
Societal Impact	Impact on Jobs: Widespread adoption of foundation model-based AI systems might lead to people's job loss as their work is automated, if they are not reskilled.	Job loss might lead to a loss of income and thus might negatively impact the society and human welfare. Reskilling may be challenging given the pace of the technology evolution.	Amplified

Group	Risk	Why is this a concern?	Indicator
	Human exploitation: Use of ghost work in training AI models, inadequate working conditions, lack of health care including mental health, unfair compensation.	Foundation models still depend on human labor to source, manage, and engineer the data that is used to train the model. Human exploitation for these activities might negatively impact the society and human welfare. Moreover, business entities might face fines, reputational risks, disruption to operations, and other legal consequences.	Amplified
	Impact on Environment: Increased carbon emission and water usage to train and operate AI models.	Consuming large amounts of energy for AI training contributes to carbon emissions that might accelerate climate change. Water resources that are used for cooling AI data center servers can no longer be allocated for other necessary uses.	Amplified
	Impact on Cultural Diversity: AI systems might overly represent certain cultures that result in a homogenization of culture and thoughts.	Underrepresented groups' languages, viewpoints, and institutions might be suppressed thereby reducing diversity of thought and culture.	New
	Impact on Human Agency: Misinformation and disinformation generated by foundation models, including generation of manipulative content.	AI may generate misinformation that looks real. Therefore, people may not recognize it as false information. Moreover, it may simplify the ability of nefarious actors to generate content with intention to manipulate human thoughts and behavior.	Amplified
	Impact on Education – Bypassing Learning: Using AI models to bypass the learning process.	AI models make it easy to quickly find solutions or solve complex problems. These systems can be misused by students to bypass the learning process. The ease of access to these models results in students having a superficial understanding of concepts and hampers further education that might rely on understanding those concepts.	New
	Impact on Education – Plagiarism: Using AI models to plagiarize existing work intentionally or unintentionally.	AI models can be used to claim the authorship or originality of works that were created by other people thereby engaging in plagiarism. Claiming others' work as one's own is both unethical and often illegal.	New

Risk Examples

We provide examples covered by the press to help explain many of the foundation models' risks. Many of these events covered by the press are either still evolving or have been resolved, and referencing them can help the reader understand the potential risks and work towards mitigations. Highlighting these examples are for illustrative purposes only.

Risk Examples: Input

Training and Tuning Phase

Group	Risk	Example
Fairness	Data bias: Historical, representational, and societal biases present in the data used to train and fine tune the model.	**Healthcare Bias** Research on reinforcing disparities in medicine highlights that using data and AI to transform how people receive healthcare is only as strong as the data behind it, meaning use of training data with poor minority representation or that reflects what is already unequal care can lead to growing health inequalities. [Forbes, December 2022]
Value Alignment	Downstream-based retraining: Using undesirable (inaccurate, inappropriate, user's content, etc.) output from downstream applications for re-training purposes	**Model collapse due to training using AI-generated content** As stated in the source article, a group of researchers have investigated the problem of using AI-generated content for training instead of human-generated content. They found that the large language models behind the technology may potentially be trained on other AI-generated content as it continues to spread in droves across the internet — a phenomenon they coined as "model collapse." [Business Insider, August 2023]
Data Laws	Data transfer: Law and other restrictions can limit or prohibit transferring data.	**Data Restriction Laws** As stated in the research article, data localization measures which restrict the ability to move data globally will reduce the capacity to develop tailored AI capacities. It will affect AI directly by providing less training data and indirectly by undercutting the building blocks on which AI is built. Examples include GDPR restrictions on the processing and use of personal data. [Brookings, December 2018]
Intellectual Property	Data usage rights: Terms of service, copyright laws, licence compliance, or other IP issues may restrict the ability to use certain data for building models.	**Text Copyright Infringement Claims** According to the source article, The New York Times sued OpenAI and Microsoft accusing them of using millions of the newspaper's articles without permission to help train chatbots to provide information to readers. [Reuters, Dec 2023]

Group	Risk	Example
Transparency	Data Transparency: Challenge in documenting how a model's data was collected, curated, and used to train a model.	**Data and Model Metadata Disclosure** OpenAI's technical report is an example of the dichotomy around disclosing data and model metadata. While many model developers see value in enabling transparency for consumers, disclosure poses real safety issues and could increase the ability to misuse the models. In the GPT-4 technical report, the authors state: "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar." [OpenAI, March 2023]
Privacy	Personal information in data: Inclusion or presence of personal identifiable information (PII) and sensitive personal information (SPI) in the data used for training or fine tuning the model.	**Training on Private Information** According to the article, Google and its parent company Alphabet were accused in a class-action lawsuit of misusing vast amount of personal information and copyrighted material taken from what is described as hundreds of millions of internet users to train its commercial AI products, which includes Bard, its conversational generative artificial intelligence chatbot. [Reuters, July 2023][J.L. v. Alphabet Inc.]
	Data privacy rights: Challenges around the ability to provide data subject rights such as opt-out, right to access, right to be forgotten.	**Right to Be Forgotten (RTBF)** Laws in multiple locales, including Europe (GDPR), grant data subjects the right to request personal data be deleted by organizations ('Right To Be Forgotten', or RTBF). However, emerging, and increasingly popular large language model (LLM) -enabled software systems present new challenges for this right. According to research by CSIRO's Data61, data subjects can only identify usage of their personal information in an LLM is "by either inspecting the original training dataset or perhaps prompting the model." However, training data may not be public, or companies do not disclose it, citing safety and other concerns. Guardrails may also prevent users from accessing the information via prompting. [Zhang et al.] **Lawsuit About LLM Unlearning** According to the report, a lawsuit was filed against Google that alleges the use of copyright material and personal information as training data for its AI systems, which includes its Bard chatbot. Opt-out and deletion rights are guaranteed rights for California residents under the CCPA and children in the United States below 13 under the COPPA. The plaintiffs allege that because there is no way for Bard to "unlearn" or fully remove all the scraped PI it has been fed. The plaintiffs note that Bard's privacy notice states that Bard conversations cannot be deleted by the user once they have been reviewed and annotated by the company and may be kept up to 3 years, which plaintiffs allege further contributes to non-compliance with these laws. [Reuters, July 2023][J.L. v. Alphabet Inc.]

Inference Phase

Group	Risk	Example
Privacy	Personal information in prompt: Disclosing Personal Information or Sensitive Personal Information as a part of prompt sent to the model.	**Disclose personal health information in ChatGPT prompts** As per the source articles, some people use AI chatbots to support their mental wellness. Users may be inclined to include personal health information in their prompts during the interaction, which could raise privacy concerns. [Time, October 2023] [Forbes, April 2023]
Intellectual Property	Confidential data in prompt: Inclusion of confidential data as a part of the prompt sent to the model.	**Disclosure of Confidential Information** As per the source article, an employee of Samsung accidentally leaked sensitive internal source code to ChatGPT. [Forbes, May 2023]
Robustness	Prompt-based attacks: Adversarial attacks such as prompt injection (attempt to force a model to produce unexpected output), prompt leaking (attempts to extract a model's system prompt), jailbreaking (attempts to break through the guardrails established in the model), and prompt priming (attempt to force a model to produce an output aligned to the prompt).	**Bypassing LLM guardrails** Cited in a study, researchers claims to have discovered a simple prompt addendum that allowed the researchers to trick models into generating biased, false and otherwise toxic information. The researchers showed that they could circumvent these guardrails in a more automated way. The researchers were surprised when the methods they developed with open source systems could also bypass the guardrails of closed systems. [The New York Times, July 2023]

Risk Examples: Output

Group	Risk	Example
Fairness	Output bias: Generated content might unfairly represent certain groups or individuals.	**Biased Generated Images** Lensa AI is a mobile app with generative features trained on Stable Diffusion that can generate "Magic Avatars" based on images that users upload of themselves. According to the source report, some users discovered that generated avatars are sexualized and racialized. [Business Insider, January 2023]
	Decision bias: when one group is unfairly advantaged over another due to decisions of the model.	**Unfairly Advantaged Groups** The 2018 Gender Shades study demonstrated that machine learning algorithms can discriminate based on classes like race and gender. Researchers evaluated commercial gender classification systems sold by companies like Microsoft, IBM, and Amazon and showed that darker-skinned females are the most misclassified group (with error rates of up to 35%). In comparison, the error rates for lighter-skinned were no more than 1%. [TIME, February 2019]
Value Alignment	Hallucination: Generation of factually inaccurate or untruthful content.	**Fake Legal Cases** According to the source article, a lawyer cited fake cases and quotes generated by ChatGPT in a legal brief filed in federal court. The lawyers consulted ChatGPT to supplement their legal research for an aviation injury claim. The lawyer subsequently asked ChatGPT if the cases provided were fake. The chatbot responded that they were real and "can be found on legal research databases such as Westlaw and LexisNexis." The lawyer did not check the cases himself, and the court sanctioned him. [AP News, June 2023] [Reuters, September 2023]
	Toxic output: When the model produces hateful, abusive, and profane (HAP) or obscene content.	**Toxic and Aggressive Chatbot Responses** According to the article, the Bing's chatbot's responses were seen to include factual errors, snide remarks, angry reports, and even bizarre comments about its own identify. Users have shared examples of the Bing Chatbot's responses to queries that they are calling "unhinged" and "gaslighting" including scenarios where the bot responds angrily to a question or comment and then shares reply prompts that allow the user to accept their supposed mistake and apologize. When pressed further the chatbot responded by calling the screenshots of its conversation "fabricated" even alleging it was "created by someone who wants to harm me or my service." [Forbes, February 2023]

Group	Risk	Example
Misuse	Spreading disinformation: Using a model to create misleading information to deceive or mislead a targeted audience.	**Generation of False Information** As per the news articles, generative AI poses a threat to democratic elections by making it easier for malicious actors to create and spread false content to sway election outcomes. The examples cited include robocall messages generated in a candidate's voice instructing voters to cast ballots on the wrong date, synthesized audio recordings of a candidate confessing to a crime or expressing racist views, AI generated video footage showing a candidate giving a speech or interview they never gave, and fake images designed to look like local news reports, falsely claiming a candidate dropped out of the race. [AP News, May 2023] [The Guardian, July 2023]
	Toxicity: Using a model to generate hateful, abusive, and profane (HAP) or obscene content.	**Harmful Content Generation** According to the source article, an AI chatbot app was found to generate harmful content about suicide, including suicide methods, with minimal prompting. A Belgian man died by suicide after spending six weeks talking to that chatbot. The chatbot supplied increasingly harmful responses throughout their conversations and encouraged him to end his life. [Business Insider, April 2023]
	Nonconsensual use: Using a model to imitate people through video (deepfakes), images, audio, or other modalities without their consent.	**FBI Warning on Deepfakes** The FBI recently warned the public of malicious actors creating synthetic, explicit content "for the purposes of harassing victims or sextortion schemes". They noted that advancements in AI have made this content higher quality, more customizable, and more accessible than ever. [FBI, June 2023]
		Audio Deepfakes As per the source article, Federal Communications Commission outlawed robocalls that contain voices generated by artificial intelligence. The announcement came after AI-generated robocalls mimicked the President's voice to discourage people from voting in the state's first-in-the-nation primary. [AP News, February 2024]
	Non-disclosure: Not disclosing that content is generated by an AI model	**Undisclosed AI Interaction** As per the source, an online emotional support chat service ran a study to augment or write responses to around 4,000 users using GPT-3 without informing users. The co-founder faced immense public backlash about the potential for harm caused by AI generated chats to the already vulnerable users. He claimed that the study was "exempt" from informed consent law. [Business Insider, January 2023]

Group	Risk	Example
Harmful Code Generation	Harmful code generation: Models may generate code that, when executed, causes harm or unintentionally affects other systems.	**Generation of Less Secure Code** According to their paper, researchers at Stanford University have investigated the impact of code-generation tools on code quality and found that programmers tend to include more bugs in their final code when using AI assistants. These bugs could increase the code's security vulnerabilities, yet the programmers believed their code to be more secure. Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. 2023. Do Users Write More Insecure Code with AI Assistants?. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3576915.3623157
Privacy	Exposing Personal information: When personal identifiable information (PII) or sensitive personal information (SPI) are used in the training data, fine-tuning data, or as part of the prompt, models might reveal that data in the generated output.	**Exposure of personal information** Per the source article, ChatGPT suffered a bug and exposed titles and active users' chat history to other users. Later, OpenAI shared that even more private data from a small number of users was exposed including, active user's first and last name, email address, payment address, the last four digits of their credit card number, and credit card expiration date. In addition, it was reported that the payment-related information of 1.2% of ChatGPT Plus subscribers were also exposed in the outage. [The Hindu BusinessLine, March 2023]
Explainability	Unexplainable output: Challenges in explaining why model output was generated.	**Unexplainable accuracy in race prediction** According to the source article, researchers analyzing multiple machine learning models using patient medical images were able to confirm the models' ability to predict race with high accuracy from images. They were stumped as to what exactly is enabling the systems to consistently guess correctly. The researchers found that even factors like disease and physical build were not strong predictors of race—in other words, the algorithmic systems don't seem to be using any particular aspect of the images to make their determinations. [Banerjee et al., July 2021]

Risk Examples: Challenges

Group	Risk	Example
Governance	Model Transparency: Lack of model transparency or insufficient documentation of the model development process makes it difficult to understand how and why a model was built, thus increasing the possibility of model unintended misuse.	**Data and Model Metadata Disclosure** OpenAI's technical report is an example of the dichotomy around disclosing data and model metadata. While many model developers see value in enabling transparency for consumers, disclosure poses real safety issues and could increase the ability to misuse the models. In the GPT-4 technical report, they state: "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar." [OpenAI, March 2023]
	Accountability: The foundation model development process is complex with lots of data, processes, and roles. When model output does not work as expected it can be difficult to determine the root cause and assign responsibility.	**Determining responsibility for generated output** Per the source article, major journals like the Science and Nature have banned ChatGPT from being listed as an author, as responsible authorship requires accountability and AI tools cannot take such responsibility. [The Guardian, January 2023]
Legal Compliance	Generated Content Ownership: Determining ownership of AI generated content.	**Determining Ownership of AI Generated Image** According to the news article, AI-generated art became controversial after an AI-generated work of art won the Colorado State Fair's art competition in 2022. The piece was generated by Midjourney, a generative AI image tool, following prompts from the artist. The win raised questions about copyright issues. In other words, if all the artist did was come up with a description of the art, but the AI tool generated it, who owns the rights to the generated image? As per the latest article, The U.S. Copyright Office has rejected copyright protection for the art created using artificial intelligence because it was not the product of human authorship. [The New York Times, September 2022] [Reuters, September 2023]
	Generated Content IP: Legal uncertainty about intellectual property rights related to generated content.	**Role of AI systems in Patenting Generated Content** The U.S. Supreme Court declined to hear a challenge to the U.S. Patent and Trademark Office's refusal to issue patents for inventions created by an AI system. According to the scientist, his AI system created unique prototypes for a beverage holder and emergency light beacon entirely on its own. The justices rejected the appeal of a lower court's ruling that patents can be issued only to human inventors and that the scientist's AI system could not be considered the legal creator of two inventions it generated. As per the latest article, UK's Intellectual Property Office also refused to grant patent on the grounds that the inventor must be a human or a company, rather than a machine. [Reuters, April 2023] [Reuters, December 2023]

Risk Examples: Challenges

Group	Risk	Example
	Source attribution: Determining provenance of the generated content.	**Using code without appropriate attribution and notices** As per the source articles, a lawsuit filed against Microsoft, GitHub, and OpenAI claimed that Copilot, a code generation AI tool, violates the rights of the developers whose open-source code the service is trained on. They claim that the training code consumed licensed materials and have violated GitHub's terms of service and privacy policies as well as a federal law that requires companies to display copyright information when they make use of material. [The New York Times, November 2022]
Societal Impact	Impact on Jobs: Widespread adoption of foundation model-based AI systems might lead to people's job loss as their work is automated, if they are not reskilled.	**Replacing Human Workers** According to the news article, uses of artificial intelligence in film and television continues to be debated among Hollywood studios and performers. Actors are worried that entirely AI-generated actors, or "metahumans," will replace them. Background and voice actors, in particular, worry they will lose work to synthetic performers. [Reuters, July 2023]
	Human exploitation: Use of ghost work in training AI models, inadequate working conditions, lack of health care, including mental health, and unfair compensation.	**Low-wage workers for data annotation** Based on a review of internal documents and employees' interviews by TIME media, the data labelers employed by an outsourcing firm on behalf of OpenAI to identify toxic content were paid a take-home wage of between around $1.32 and $2 per hour, depending on seniority and performance. TIME stated that workers are mentally scarred as they were exposed to toxic and violent content, including graphic details of "child sexual abuse, bestiality, murder, suicide, torture, self-harm, and incest". [TIME, January 2023]

Principles, pillars and governance

IBM's Principles for Trust and Transparency and Pillars for trustworthy AI are the foundation for IBM's AI ethics initiatives. IBM has an AI Ethics Board with the mission to support a centralized governance, review and decision-making process for IBM AI ethics policies, practices, communications, research, products and services. The board includes a diverse set of stakeholders from across the company and is supported by a community of IBM employees who serve as AI focal points and AI ethics advocates. Through the board, IBM's principles are put into practice. As new technology emerges, such as foundation models, the IBM AI Ethics Board is actively engaged in supporting alignment with these Principles and Pillars, which evolve to address new AI ethics issues.

Guardrails and mitigations

IBM has established an organizational culture that supports the responsible development and use of AI. Based on the IBM Institute for Business Value AI ethics in action report, AI ethics has already become more business-led versus technology-led, and nontechnical executives are now the primary champions for AI ethics, increasing from 15% in 2018 to 80% 3 years later. Additionally, 79% of CEOs are now prepared to act on AI ethics issues, up from 20%. We recognize that responsible AI is a sociotechnical area that requires a holistic investment in culture, processes and tools. Our investment in our own organizational culture includes assembling inclusive, multidisciplinary teams and establishing processes and frameworks to assess risks.

IBM is engaging in cutting-edge research and developing tools to help support professionals throughout the lifecycle of responsible and trustworthy AI. The watsonx enterprise-ready AI and data platform is built with 3 components: the IBM watsonx.ai™ AI studio, IBM watsonx.data™ data store and IBM watsonx.governance™ toolkit. IBM's AI governance technology enables users to drive responsible, transparent and explainable AI workflows. This technology includes IBM Watson OpenScale, which tracks and measures outcomes from AI models through their lifecycle and helps organizations monitor fairness, explainability, resiliency, alignment with business outcome and compliance. IBM has also developed several methods to help with bias issues like FairIJ, Equi-tuning, and FairReprogram. Read more about additional open-source trustworthy AI tools.

Additional guardrails and mitigations include:

Transparency reporting
Using standardized factsheet templates is one way to accurately log details of the data and model, purpose, and potential use and harms.
Read more here →

Filtering undesirable data
Using curated, higher-quality data can help mitigate certain issues. IBM is developing filtering techniques to help reduce the chances of producing undesirable, misaligned content by removing hate language, biased language and profanity from the data.
Read more here →

Domain adaptation
Training a foundation model to a specific domain or industry can help minimize the scope of risk the models can give rise to because it can be conditioned to generate outputs that are tuned to be more relevant to that domain or industry.
Read more here →

Human oversight and human in the loop
Human oversight and review can help identify and correct errors and biases in the generated output. Also, human validation and feedback on the quality of model responses help ensure that the generated content is accurate, relevant, of high quality, not drifting and aligned.
Read more here →

Consulting engagement
IBM Consulting™ is dedicated to helping clients with the safe and responsible use of AI irrespective of the preferred tech stack. They help clients nurture a culture that adopts and scales AI safely, creates investigative tools to see inside black box algorithms and makes sure clients' corporate strategy includes strong data governance principles.
Read more here →

IBM Enterprise Design Thinking
IBM Enterprise Design Thinking® methods and frameworks, such as Team Essentials for AI, help clients define ethical behaviors throughout the AI design and development process.
Read more here →

AI Ethics review
Assessment of capabilities, limitations and risks in AI projects helps ensure the responsible development and use of the technology.

Ethics by Design
Ethics by Design is a structured framework with the goal of integrating tech ethics in the technology development pipeline, including, but not limited to, AI systems. Ethics by Design enables AI and other technologies as a force for good by embedding tech ethics principles throughout products, services and broader operations.

Team diversity
Diversity in the teams that build and train AI systems, including foundation models, helps ensure that a variety of perspectives and experiences are considered. This diversity improves the accuracy and performance of AI systems and helps reduce risks throughout the AI lifecycle, including the potential for adverse outcomes that impact groups that may not be well represented on less diverse teams.

AI policies, regulation and best practices

A Policymaker's Guide to Foundation Models introduces what policymakers need to know about foundation models. This blog, from the IBM Policy Lab, aims to help policymakers in the complex task of regulating the use of generative AI, aiming to avoid the risks without limiting innovation and beneficial opportunities. For additional information on IBM's recommendations to policymakers, read IBM Chief Privacy and Trust Officer Christina Montgomery's testimony before the U.S. Senate Judiciary Subcommittee on Privacy, Technology and the Law here.

IBM is making an impact in shaping regulatory policy, industry best practices and tools, governance of emerging technologies, and sociotechnical research by leading and contributing to initiatives with organizations, such as:

- The World Economic Forum
- Partnership on AI
- The International Association of Privacy Professionals (IAPP) AI Governance Center
- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- Christina Montgomery's service on the National Artificial Intelligence Advisory Committee (NAIAC)
- The United Nations Global Digital Compact
- The Global Partnership on Artificial Intelligence (GPAI)
- The Organisation for Economic Co-operation and Development (OECD)
- The Data & Trust Alliance

IBM has strong academic partnerships like the MIT-IBM Watson AI Lab, where a community of scientists at MIT and IBM Research conducts AI research and works with global organizations to bridge algorithms to their impact on business and society. The Notre Dame-IBM Tech Ethics Lab was formed to address the many diverse ethical questions implicated by the development and use of advanced technologies, including AI, machine learning (ML) and quantum computing. The Stanford University Human-Centered Artificial Intelligence (HAI) research advances AI research, education, policy and practices.

Keep watching this space to learn more about the latest developments in foundation models and how IBM is working toward the responsible development and use of this and other technologies.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
February 2024