Non-Decisional Statement by the National AI Advisory Committee (NAIAC)
Working Group on Generative & NextGen AI: Safety and Assurance

Working Group Members: Jack Clark, David Danks (Vice Chair), Paula Goldman (Chair), Ashley Llorens, Haniyeh Mahmoudian, Swami Sivasubramanian

Disclaimer: This non-decisional document has been drafted by Members of the National Artificial Intelligence Advisory Committee (NAIAC) for the purposes of explaining concepts and does not offer formal recommendations. The opinions discussed in this document do not represent the views of the full Committee, and should not be considered a recommendation by the NAIAC.

FAQs on Foundation Models and Generative AI

Reflecting understanding as of 8/28/2023

Table of Contents

Development and Capabilities of Foundation Models/Generative AI

Use of Foundation Models

Risks of Foundation Models

3. Which of these risks are new and which are amplifications of existing issues?
4. Should we be concerned with existential risks associated with general AI?

Current Technical and Product Related Guardrails for FMs and Applications Built on FMs

1. Are there existing methodologies and benchmarks to evaluate and assess (accuracy, safety etc) FMs?
2. What transparency currently exists in the development and deployment of FMs?
3. How do companies across the stack currently govern their own FMs? What (potential) best practices (could) exist for governance of these technologies?
4. How do app developers currently mitigate risk of adverse outcomes with apps built on FMs?
5. How do FM creators and/or app developers lower the risk of biased or toxic content?
6. Are there current US Federal and/or state regulations that apply to FMs?
7. Can we/how can we verify content that is produced by AI?
8. Are there existing standards that do or could apply to these FMs? Can NIST AI RMF be used to manage, monitor and mitigate FM related risks?
9. What collective efforts exist around creating technical standards and/or probing the capabilities and limitations of large-scale models.
10. What information could be provided to “downstream” developers & users of an FM about the capabilities, limitations, etc. to enable effective & ethical use of the FM? How does this information potentially impact liability and other legal issues?

Development and Capabilities of Foundation Models/Generative AI

1. What is a Foundation Model? How can a foundation model be used?

The term Foundation Model (FM) was introduced by the Stanford Institute for Human-Centered AI (HAI)[1]. They referred to FMs as “any model that is trained on broad data that can be adapted to a wide range of downstream tasks". The term *model* here refers to a complex mathematical function that maps inputs to outputs, trained using machine learning. FMs are built on a massive scale, containing billions of tunable parameters or more, and are trained using data from a broad range of sources often gathered by crawling the web, including internet text, image, and audio. A user can utilize an FM to perform a task by submitting a series of natural language prompts. For example, the user can prompt the model to summarize a draft of legislation and use subsequent prompts to refine the draft. However, this is not only applicable to generating text. For example, a user can prompt an image generation model with a description of a painting and the model will generate visual content.

[1] Stanford CRFM

2. What are different types of Foundation Models?

Different types of foundation models can be categorized based on the type of data they process or tasks they perform. For example:

- **Language Models**: These models, like GPT-3.5 from OpenAI[2] and PALM 2 from Google[3], are trained on large-scale text data and can generate human-like text. By providing specialized content via input prompts, these LLMs can perform specialized tasks such as answering questions about a particular company's products and services using informal retrieval techniques

- **Multimodal Models**: These models are trained on data that includes multiple types of information, such as text and images. They can be used for tasks that require understanding different types of information. For example, models trained on both text and visual information can generate a text description of an image or generate images from text descriptions, like DALL-E from OpenAI and Stable Diffusion[4].

- **Audio Models**: These models are trained on audio data and can be used for tasks like speech recognition, speech synthesis, or music generation[5].

- **Video Models**: These models are trained on video data and can be used for tasks like action recognition or video generation[6].

3. What are Foundation Models capable of doing that earlier generations of AI could not do as well?

What has traditionally been used can be thought of as specialized models, developed using supervised machine learning. Specialized models are typically only used to perform the specific task they were trained for. For example, a specialized model trained using previous customer churn data used to predict future customer churn, or a model trained using previous sales data to predict likely sales trends for new products. In contrast, Foundation Models can be developed using self-supervised learning, and used in a variety of use cases[7]. For instance, a single large language model (LLM) is a type of FM that can be used for multiple tasks such as summarizing texts, translation to other languages, and identifying the topics in the text. FMs have introduced several advancements over earlier generations of AI models. Here are some capabilities of foundation models that set them apart:

1. **Demonstrating linguistic sophistication**: Foundation models excel in processing and generating human-like text. They can process large amounts of text data, allowing them to identify complex language patterns and nuances more effectively than earlier models. In addition, FMs can generate coherent and human-like text, making them useful for tasks like content creation,

[2] GPT-4 (openai.com)
[3] Google AI PaLM 2 – Google AI
[4] Stable Diffusion Online (stablediffusionweb.com)
[5] Meta open sources an AI-powered music generator | TechCrunch
[6] Meta’s new text-to-video AI generator is like DALL-E for video - The Verge
[7] On the Opportunities and Risks of Foundation Models

creative writing, and even storytelling. They can produce entire articles, essays, or narratives based on given prompts or guidelines[8].

2. **Contextual understanding**: Foundation models can leverage context from the prompts to generate more coherent and contextually relevant responses. When interacting with LLMs, the previous turns of dialog and interaction contribute to the prompt for the next turn, and that turn then contributes to future turns. This enables the models to better identify the flow of conversation, which helps them provide more meaningful and accurate information.

3. **Longer contextual memory**: Earlier AI models often struggled with short-term memory due to lower maximum number of tokens accepted as input, but foundation models have significantly improved in this aspect. They can retain information from previous interactions within the same conversation, making it easier to maintain coherent and consistent conversations over extended periods[9].

4. **Larger training datasets**: Foundation models are exposed to a vast amount of information during training. This enables these models to provide plausible generations related to a wide range of topics, given appropriate context. It is important to note that the information and patterns observed during training can change over time, so generations may not always be up-to-date or accurate without augmentation through some form of in-context information retrieval.

5. **In-context learning**: Foundation models can perform reasonably well with very little training data or even without any specific training to perform a particular task. They can generalize learning from the training data to new examples and demonstrate impressive performance on tasks they haven't been explicitly trained on[10].

Overall, foundation models represent a significant advancement in natural language processing, offering improved context awareness, memory, and the ability to perform a wide range of language-based tasks more effectively.

4. Why are FMs suddenly more capable?

Recent advancements in ML (specifically the invention of the *transformer-based*[11] neural network architecture) have led to the rise of large-scale models that contain billions of parameters. A parameter is a configuration variable that is internal to the model and whose value can be tuned to optimize model performance. To give a sense for the change in scale, the largest pre-trained model in 2019 (BERT) was 340M parameters, whereas state-of-the-art FMs introduced in 2022 were on the order of 500B parameters

[8] arxiv.org/pdf/2107.00061.pdf

[9] The Challenges and Opportunities in Long-Term Memory for Language Models - ChatGPT / Bugs - OpenAI Developer Forum

[10] In-Context Learning Approaches in Large Language Models | by Javaid Nabi | Jul, 2023 | Towards Data Science

[11] https://arxiv.org/pdf/1706.03762.pdf

(e.g., Microsoft's Megatron Turing[12] and Google's PALM[13]) – an increase of 1,600x in size. The size of these models plays a big role in what makes them remarkable (though recent research has aimed to reduce the sizes of these models without significant loss of performance).

5. What are the main steps in creating a FM?

Foundation models generally undergo a two-step training process: pre-training and fine-tuning[14]. This process is an active area of research and innovation and comes with considerable variation. During pre-training, the model generally learns from a large corpus of diverse text or other data, developing a general understanding of (in the case of LLMs for example) language patterns and grammar. This pre-training exposes the model to a broad knowledge base. While this can take a number of forms, it generally involves training a model with terabytes of unlabeled text and/or multi-modal data (such as images, audio, video). Sometimes these data are obtained by crawling the Web for publicly available sources (such as Wikipedia)[15]; sometimes the data sets are proprietary; or pre-training may involve a mixture. Again, while processes vary, it is common for much of the data used for training to be unlabeled data, in contrast with labeled data that requires a human task force to create laborious annotations (e.g. explicitly adding the tag "dog" to an image that contains one). During the training process, the model learns to utilize context in a sequence of tokens (e.g., words, parts of an image, etc.) to predict the next token in the sequence. A large model with billions of parameters can better capture this knowledge as it is able to analyze richer and deeper context across large amounts of data in its memory, compared to a smaller model trained on a smaller data set. A pre-trained model is not a database and is not intended to memorize training data but rather to learn relations among input sequences of tokens. Pre-training models of this size requires access to: a) sufficient quantity and quality of training data (this involves collection of the relevant datasets and processing them) and b) large-scale training infrastructure (e.g. GPU chips, which can be expensive).

In foundation models, data featurization (the process of transforming raw data into features that can be used to improve the performance of machine learning algorithms) plays a crucial role in differentiating between contextual understanding and memorization[16]. Here's a short description of how data is featurized for these purposes:

1. **Contextual Featurization**: For context understanding, data is featurized in a way that emphasizes the relationship between words and their surrounding context. The model is trained to capture the sequential dependencies and contextual clues present in the text. This is achieved through "self-attention based featurization" typically refers to a method used in natural language processing (NLP) and other machine learning tasks where the model learns to focus on different parts of the input data to extract meaningful features. The self-attention mechanism allows the model to assign different importance weights to different words in the input sequence, allowing it to capture long-range dependencies and improve its understanding of the context.

[12] Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model - Microsoft Research

[13] Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance – Google Research Blog (googleblog.com)

[14] https://www.datacamp.com/blog/what-are-foundation-models

[15] OpenAI Quietly Unveils Web Crawler to Scrape Data for Its AI Models (aibusiness.com)

[16] Featurization with automated machine learning - Azure Machine Learning | Microsoft Learn

2. **Memorization Avoidance**: To prevent the model from excessively relying on memorization, various techniques are employed during training. One such technique is masking, where certain tokens in the input text are intentionally replaced with special tokens, such as Mask or random tokens. This encourages the model to learn to generate appropriate responses based on the context and surrounding information, rather than simply regurgitating memorized examples.

By carefully featurizing the data and training the model with appropriate techniques, foundation models strike a balance between capturing contextual information and avoiding pure memorization. This enables them to process and generate text that is contextually relevant, coherent, and adaptable to various conversational or task-specific scenarios.

Upon completion of the pre-training step, the resulting model can deliver an impressive out-of-the-box performance on a wide range of tasks across multiple domains. For example, an FM can tackle many diverse tasks such as writing blog posts, summarizing documents, solving math problems, engaging in a chat dialogue, answering questions based on a doc, and even composing poetry. To further refine the model's ability to understand and generate text in specific contexts, the model is trained on specific task data, which provides it with context and prompts relevant to the target application.

6. What is special about transformers and why do they make FMs more powerful?

To understand FMs better, let’s first dive deep into *transformers*, a popular model architecture that led to the rise of FMs. A transformer-based model has an encoder component that converts the input text into *embeddings* (mathematical representations), and a decoder component that consumes these embeddings to emit some output text. Compared to its predecessors like recursive neural nets, transformers are more parallelizable. Rather than process input sequences one token at a time they instead process input sequences all at once capturing sequential information using self-attention and positional encoding. As a result, transformers require significantly less time to train as one can apply more computing power to speed up training.[17]

7. How much does it cost to train FMs?

The cost of training such models can vary greatly based on a multitude of factors, including the size of the model, the resources used for training (e.g., the type and number of GPUs), the duration of training, the cost of the data used for training, and so on. Many technology companies that invest in developing large-scale models do not disclose the exact costs associated with their development. However, it's generally understood that training large-scale models can reach or exceed hundreds of millions of dollars, depending on many different factors. That said, there are many directions of research and innovation aiming to drive training costs down including, for example, data-curation and synthetic data generation, efficient neural architecture design, and efficient acceleration software.[18]

[17] https://arxiv.org/pdf/1706.03762.pdf

[18] https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html

8. Are foundation models an example of AGI?

While there is no agreed-upon definition of Artificial General Intelligence (AGI), many definitions combine some notion of generality (i.e., a single AI system that is competent at performing a broad range of tasks) and level of competency (i.e., performance across a broad range of tasks that meets or exceeds some standard of human-level performance). The underlying cognitive capabilities of such an AI system may include, for example, the ability to master new tasks on the fly, the ability to propose, evaluate and carry out complex courses of action and the ability to apply learned knowledge in novel task domains.

Foundation Models, while offering unprecedented generality and competency, are not widely considered to meet the subject threshold of AGI-level capabilities.[19] This is due, among other things, to limitations in their ability to plan and act in the physical world and inconsistencies in their ability to perform mathematical calculations and reason accurately over bodies of knowledge.

Use of Foundation Models

1. How are FMs being deployed into the world?

As discussed earlier, in the case of specialized AI, a specialized model is trained for each individual use case. However, a single Foundation Model (FM) can serve a wide array of use cases, including image, text, and audio domains. For instance, image-based FMs have the ability to generate images in response to user instructions. Keeping in mind that FMs are still at a nascent stage, below are several domain-specific examples of how FMs are being used in the private and public sectors. These examples should be considered for the value as well as their implications for fairness and bias, continued technical innovations, workforce disruptions and other factors. The examples below should also not be considered as exhaustive, given the rapidly evolving development of the technology as well as innovations by users of the technology in each of the domains.

Private sector

- **Marketing, Sales and Service:** Foundation models (FMs) offer valuable support for content creation across marketing campaigns, social media, and sales and service content. They possess the capability to generate personalized and tailored content that aligns with the specific requirements of prospects and customers. By leveraging FMs, businesses can accelerate content production which can lead to better engagement with customers.

- **Code Generation:** Foundation models (FMs) possess the remarkable ability to generate code and documentation[20], offering a valuable resource for organizations. By leveraging FMs, companies can make it easier for anyone to write code, regardless of expertise, and empower their engineers to concentrate more effectively and efficiently on addressing identified bugs and conducting

[19] Artificial general intelligence - Wikipedia

[20] https://www.forbes.com/sites/janakirammsv/2022/03/14/5-ai-tools-that-can-generate-code-to-help-programmers/?sh=14023d135ee0

thorough system testing. This capability enables a streamlined development process, allowing engineers to optimize their time and expertise for critical tasks

- **Synthetic Data and Data Enrichment:** Data privacy concerns pose significant challenges when it comes to collecting and utilizing certain data sources in AI. However, foundation models (FMs) offer a solution by enabling the creation of synthetic data that closely resembles real-world data. This synthetic data can be leveraged by AI creators to build privacy-enhanced AI systems while maintaining the integrity of the resulting models. Furthermore, FMs can also be utilized to enrich existing data, providing additional context and information to enhance the performance and capabilities of AI systems. By leveraging FMs in data processing, organizations can address privacy concerns and enhance the overall effectiveness and privacy of their AI applications.

- **Healthcare:** By harnessing the power to generate realistic X-rays, MRI, and CT scans from patient data, this technology enhances the precision of specialized AI models in detecting anomalies within medical images. Consequently, it becomes an invaluable tool for physicians in diagnosing various conditions, particularly in critical diseases like cancer that rely on early detection. Additionally, these advancements aid in drug design, leading to significant cost reduction and shorter discovery timelines. Furthermore, generative AI can play a pivotal role in enhancing the quality of prosthetic devices, much like it does in product design. Overall, these advancements in technology have far-reaching implications for improving medical diagnostics, treatment, and patient care.

- **Product Design:** Generative AI holds the potential to expedite the design process for products and user interfaces by swiftly generating initial draft versions of designs. These drafts serve as a starting point for engineers, R&D teams, and design teams, enabling them to collaborate more effectively and enhance the design at a faster pace. By leveraging generative AI, the iterative design cycle becomes more efficient, allowing for rapid improvements and iterative refinements. This collaborative and accelerated approach empowers teams to streamline the design process and ultimately deliver higher-quality products and user interfaces in a more time-efficient manner.

- **Conversational bots:** Utilizing large language models, chatbots have the potential to greatly enhance the customer service and support experience by providing conversational question-and-answer interactions. These enhanced chatbots leverage their extensive language processing capabilities to engage with customers, address their queries, and offer relevant assistance. By employing chatbots based on large language models, businesses can deliver prompt and accurate responses, streamline customer interactions, and ultimately improve the overall customer service and support experience.

- **Law:** Leveraging large language models, organizations can benefit from advanced capabilities in contract analysis, legal case summarization, and identification of relevant cases for lawyers. These models efficiently process vast amounts of legal information, enabling them to provide answers related to contracts and summarize diverse legal cases. By utilizing large language models, legal professionals can significantly reduce research time and costs for clients. This technology empowers lawyers to access relevant information swiftly, enhancing their efficiency and effectiveness in delivering legal services.

Public sector

- **Services:** Generative AI tools offer capabilities to streamline complex customer service cases, whether through chatbots, or case assistance and summarization for service workers. Such tools could help agencies streamline the process by which citizens resolve questions about public services and get access more quickly.
  - **Conversational bots:** Conversational bots can enhance customer support by guiding individuals seeking specific government services in the right direction through improved chatbot and virtual assistant capabilities. Chatbots based on generative AI tools can also help enhance accessibility for vision- and hearing-impaired individuals.
  - **Live translation:** Conversational tools have the potential to enhance and automate translation services, particularly for non-English speakers in need of assistance. These tools enable efficient and accurate communication by facilitating real-time language translation. By leveraging conversational AI, organizations can provide seamless and effective language support, enabling non-English speakers to access the information and services they require with ease. This technology helps bridge language barriers, improves accessibility, and enhances the overall experience for individuals in need of translation services.
  - **Case Management:** Generative AI tools can be used to speed up case management, including claims processing for government agencies by extracting key data from forms, flagging issues for prioritization. FMs can also be used to generate answers to frequently asked questions and to flag questions requiring human support.
- **Press releases and public facing content:** Generative AI tools offer valuable capabilities in generating content assets to effectively communicate government activities through websites and social media platforms. These tools can assist in creating informative materials that keep the public well-informed about government initiatives. Furthermore, these AI models can be utilized to summarize complex legislation, policy papers, and reports, providing concise and easily understandable overviews. By leveraging generative AI, governments can enhance their communication strategies, streamline information dissemination, and facilitate public engagement with important policies and activities.
- **Intelligence agencies:** Text-based foundation models (FMs) tools, such as ChatGPT, can serve as valuable aids to personnel across different intelligence agencies. These tools can support critical intelligence processes. By leveraging FMs, intelligence professionals can access a wealth of information, rapidly process large volumes of text-based data, and gain valuable insights for intelligence analysis and decision-making. The advanced language understanding and contextual capabilities of FMs enable them to assist personnel in tasks such as information extraction, summarization, and contextual understanding, ultimately enhancing the efficiency and effectiveness of intelligence operations.
- **Cyber and Fraud Defense Solutions:** Amid the increasing prevalence of cybersecurity attacks, generative AI tools offer valuable support in the battle against adversarial efforts aimed at

compromising government security. For example, these tools can be utilized to create deceptive traps, known as honeypots, by generating synthetic intellectual property data. By diverting the attention of attackers towards these fake data repositories, the security team can protect actual sensitive information and infrastructure from malicious intent. This proactive approach helps enhance the security posture of the government by mitigating the risk and impact of cybersecurity threats.[21][22][23][24] LLMs are also being used in fraud detection, analyzing large amounts of text in a short period of time to identify anomalies that may require human review.

2. What are the current limitations of FMs (Summer 2023)?

While FMs have made significant advancements and can perform a wide variety of tasks with impressive results, there are still areas where they struggle or fall short. Some of these include:

- **Situational Awareness**: FMs can only utilize context provided in an input prompt combined with contextual understanding gleaned during pre-training to inform their reasoning. While that level of contextual understanding is unprecedented for AI systems, humans utilizing FMs or FM-powered applications often bring a much richer and more nuanced understanding of the state of the world in that moment, of themselves, of the task at hand and the range of acceptable and unacceptable outcomes.

- **Handling Uncertainty**: If a task or question is uncertain or ambiguous, FMs may struggle to generate a sensible response. They don't have the capability to seek clarification like a human would.

- **Fact Checking and Truthfulness**: FMs can't verify the truthfulness or accuracy of their outputs in real-time. They might generate information that seems plausible but is factually incorrect based on their training data. There have, however, been effective instances of people using multiple instances of models to check each other's work. In addition there are emerging examples in which FMs have called on other systems, such as databases, search and others, to improve grounding and factuality.

- **Domain-Specific Tasks**: Without specific fine-tuning, FMs might struggle with very specialized tasks, like providing medical or legal advice.

- **Ethics and Bias**: FMs can unintentionally generate biased or inappropriate content, as they learn from the data they were trained on. If those data contain biases (as almost any large dataset will), the model will learn and potentially propagate those biases. It is worth noting that there has been progress in development of the FMs to enforce no harm and ethical behavior through methods such as constitutional AI[25][26].

[21] https://www.eweek.com/artificial-intelligence/generative-ai-enterprise-use-cases/
[22] https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#business-value
[23] https://research.aimultiple.com/generative-ai-applications/
[24] https://www.gartner.com/en/topics/generative-ai
[25] 2212.08073.pdf (arxiv.org)
[26] https://venturebeat.com/ai/foundation-models-risk-exacerbating-mls-ethical-challenges/

- **Societal Risks**: The Stanford HAI report emphasized the clear and significant societal risks presented by foundation models, such as inequity, misuse, environmental impact, legal frameworks, and economic consequences. These risks could limit their effective application in areas that require careful ethical and societal considerations.

3. What are the technical limitations of FMs (Summer 2023)?

As of the summer of 2023, large-scale foundation models have a few key technical limitations, which has a downstream effect on their use cases:

- **Active learning / online adaptation:** If you want to update a modern foundation model with a large amount of data, you typically need to retrain the model. Models are not able to learn in 'real time' in a meaningful way; e.g, if you write some text to a model, you typically shouldn't expect that model to remember that text the next time you log-in.

- **Context windows:** Foundation models can remember some amount of data (which is typically referred to as a number of 'tokens'). While these amounts of data are quite large (on the order of tens of thousands of words for text models) they aren't infinite. That means that a FM model can only operate over so much data before you need to either finetune or retrain it.

- **Not very portable at the largest scales:** Large-scale foundation models typically require non-trivial hardware to allow sampling from them - usually on the range of tens of high-end computer processors. This means that it is quite hard to miniaturize foundation models and run them locally on phones or computers. (Counterintuitively, image models are *much* cheaper to run locally, so people have been able to get frontier image foundation models running on phones and computers. Currently, text models do not seem to be runnable on phones, but work on minimization is ongoing.)

4. What future advancements in FMs are likely to occur in the near term and what kinds of capabilities will they make possible?

As foundation models grow larger and more complex, researchers are exploring ways to compress them, which allows the deployment of the tools based on FMs on devices such as cell phones. In addition, given the significant impact of these systems on the environment, there is research underway to reduce the energy and carbon footprint on training FMs. Similarly, there are efforts to improve techniques for addressing bias and fairness in FMs, as well as their other limitations, such as factuality and grounding. However, as of now, it's challenging to predict the exact future advancements in foundation models and the specific capabilities they will make possible in the near term. One specific complication for prediction is the increasing integration of FMs with other types of AI systems that are specialized to solve tasks for which FMs typically struggle. These hybrid systems may advance significantly faster than either type alone.

There are a few concerns that have been raised about the current state and direction of foundation models that might shape their future development. For instance, the Stanford HAI institute emphasizes that these models present significant societal risks, including inequity, misuse, environmental impact, legal frameworks, ethics of scale, and economic consequence. They suggest that the current trajectory of foundation models is not inevitable, and significant change is necessary in both model development and the broader ecosystem, such as adopting data practices that respect the rights and dignity of data subjects as opposed to indiscriminate scraping, increasing access to these models, and decreasing centralization of power surrounding these models in large technology companies[27].

Risks of Foundation Models

1. What are the risks associated with input/Model/output?

The AI pipeline can be segmented into three stages, input, model, output. Input consists of the data collection, evaluation, and all the pre-processing done on the data to prepare it for model training. Model refers to the training stage which includes all the decisions made in the design set up of the training process and model performance evaluation. The last stage is output which refers to the model's output. As mentioned above, for FMs the output can be in the form of text, image, or audio. For each stage, the risks associated with FMs differ. Below are some the risks:

- **Hallucination/Confabulation/Accuracy**: Typically, as LLMs get bigger in size, they can produce answers to prompts that are factually incorrect but presented in a way that seems convincing/definitive. This is sometimes called a 'hallucination' or 'confabulation'. Such an outcome is typically a byproduct of the way these FMs represent their inputs, often causing them not to distinguish among different numeric values or names, to "invent" facts to be consistent with the requested output format (e.g., inventing citations and author names when asked to provide evidence to an answer), and to conflate facts that are presented by multiple sources in their input[28].

- **Bias**: The FMs are trained on vast amounts of data crawled from the internet. This data includes text, visual, and audio information that contains bias and stereotyping of different groups. Both predictive AI and FMs are prone to amplifying these biases and stereotypes in their predictions and outputs[29][30][31].

- **Toxic and offensive content**: Similar to the concern on bias, these models are trained on data that are toxic and harmful. Without appropriate guardrails in place, FMs are at risk of generating inappropriate content[32].

- **Privacy**: The FMs are trained on vast amounts of data, some of which could potentially contain sensitive or private information. Even if they don't explicitly memorize this data, they might learn patterns or information structures that could reveal private details. For example, if a model is trained on a dataset that includes private conversations or confidential documents, it could potentially generate outputs that reflect the information from these sources. Further, datasets

[27] Reflections on Foundation Models (stanford.edu)
[28] https://www.techtarget.com/whatis/definition/AI-hallucination
[29] https://arxiv.org/pdf/2209.02965.pdf
[30] https://www.bloomberg.com/graphics/2023-generative-ai-bias/
[31] https://research.ibm.com/blog/debugging-AI-bias
[32] https://calendar.google.com/calendar/u/0/r/week/2022/8/2?pli=1

might be combined, featurized, and analyzed in ways that are more revealing of individuals than would be anticipated by each dataset on its own.

- **Copyright and IP Leakage and "Ownership":**: A number of complex issues arise from the question of intellectual property rights to the inputs and outputs of FMs. On model inputs, the large-scale web crawling approach to collect data for FMs can result in a dataset that includes copyrighted content. If a model is trained on copyrighted texts, for example, it might generate text that is very similar to the original source. This raises concerns about copyright infringement.[33] Accidental IP leakage is another issue that can occur, for example, when employees of a company input proprietary information into publicly available interfaces such as chatGPT.[34]
    - Related, the question of overall ownership for the output of foundation models is a complex and evolving area of law and ethics, and different parties might have different perspectives. Some argue that the user who inputs the prompt owns the output. This is based on the idea that the user's input guides the model's output. Under this view, the output is a kind of collaboration between the user and the model. Alternatively, one might argue that the entity that created and trained the model owns the output, since the model is producing the output based on the patterns it learned during training. In this case, the output might be viewed as a kind of derivative work of the training data. A middle ground might be to view the output as jointly owned by the user and the model creator. This might be appropriate if both the user's input and the model's training are seen as contributing significantly to the output. Lastly, some argue that the outputs of AI models should be considered in the public domain, especially if the training data includes public domain works or contributions from a large number of people. The appropriate view may depend on a variety of factors, including the specific use case, the level of creativity or originality in the user's input, and the nature of the training data. Legal jurisdictions may also have different interpretations of these issues.

- **Security and robustness**: Foundation models, particularly those employed in high-stakes or sensitive applications, can pose significant security and robustness risks. These models are susceptible to adversarial attacks, where slight modifications to the input data can lead to drastically different and potentially harmful outputs. Similarly, codes generated by language models may contain security vulnerabilities which may result in system/software vulnerabilities if not reviewed by humans. Robustness in the context of AI models refers to their ability to perform well and maintain accuracy when faced with various types of inputs, including noisy or adversarial inputs. This is a significant risk as the model can fail or behave unpredictably when confronted with less represented (e.g. languages spoken by small communities), unfamiliar, or intentionally misleading inputs. In addition, even when guardrails are in place, adversaries can use jailbreaking to trick the system to bypass the guardrails to generate harmful outputs or share sensitive information.

- **Accountability in complex value-chains**: There are two levels of consideration in FMs. The first level is where organizations are creating foundation models, where risks are associated with data

[33] Lawsuit Takes Aim at the Way A.I. Is Built - The New York Times (nytimes.com)
[34] Samsung Fab Workers Using ChatGPT Accidentally Leak Confidential Information | Extremetech

and trained models. The second level is at the application level –for example when LLMs used to create a chatbot or image model used to generate marketing images. Depending on the application, use of FMs can pose risks specific to that use case. Further value-chain concerns can arise when FMs are developed, fine-tuned by others in the value chain, or when applications are built by others using APIs to FMs created by others, or in the case of open-sources FMs where the FM is modified by others, as well as through misapplication and misuse of FMs (or their derivative products) by users - individuals or organizations. Thus, defining responsible parties and accountability can be challenging.

2. What are the potential Societal and Economic Impacts of FMs ?

There are several concerns regarding potential negative social and economic impacts of foundation models. Various researchers have highlighted these potential risks in some detail, including a study by Google DeepMind researchers that identified 21 in 5 different categories.[35]

Below are some of the documented impacts:

- **Labor market implications**: Some studies conclude that FMs will create significant job losses, leading to major labor market disruptions[36]. Other studies argue that in the medium to long term AI will generate new jobs and industries and be net positive for jobs[37]. Many of the same studies and others also suggest that in addition to potential job losses and job gains, many jobs will change as some aspects and components of jobs are complemented by AI. One general place in consensus is the need to concentrate on disruptions and shifts that occur as a result of AI.

- **Denial of consequential services/rights/biased inputs to decision making**: When FMs are deployed as decision inputs to high stakes decisions they could result in life altering consequences if inaccuracy or bias is present in the model's outputs. This may be particularly true when deployed in high stakes settings such as medical, legal, employment, energy, law enforcement, finance and other sensitive domains.

- **Mis/disinformation at scale**: With the advancement in the generated content by the FMs, it has become harder to distinguish between AI and human generated content. This opens the door for bad actors to generate and spread mis/disinformation at scale which can have significant impacts[38].

- **Environmental Impact**: To process the large-scale data, train FMs, and content generation by FMs, staggering computing power is required to operate the process. Generally, larger FMs require more energy for training and thus higher carbon footprint. It is worth noting that these models need continual updates and retraining to incorporate the most recent information in the training process. This results in continual impact of the technology on the environment. Google and UC Berkeley researchers have estimated that GPT-3 has emitted close to 552 tCO_2e in CO_2

[35] [https://arxiv.org/abs/2112.04359

[36] https://www.oecd.org/employment-outlook/2023/#ai-jobs

[37] https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america

[38] https://www.vice.com/en/article/xgwgn4/researchers-demonstrate-ai-supply-chain-disinfo-attack-with-poisongpt

or 1,287 MWh in energy consumption during training[39] which is similar to electricity consumed by 121 U.S. households in an entire year.[40]

- **Economic Concentration**: Given the current state of the technology, developing FMs are costly and require significant funding that notably limits the prospect of partaking in the development of FMs to a small number of organizations. Some have voiced concerns about possible economic concentration accruing to organizations that can afford to build FMs.

3. Which of these risks are new and which are amplifications of existing issues?

Risks such as bias and fairness and stereotyping, mis/disinformation being promoted on social media are among the existing issues with AI. However, the *widespread* generation of toxic and harmful content and mis/disinformation that is now significantly easier due to FMs. Similarly, with the general purpose nature of these models, bias and discrimination can be amplified by FMs in a variety of forms based on the use case.

4. Should we be concerned with existential risks associated with general AI?

There are a wide range of views on this topic, often provoking contentious debate. At a high level, the different positions are characterized by (i) the likelihood of human extinction or a global-scale catastrophic event due to rapid advances in AI technology; (ii) the timescale for such risks (imminent vs. within the next decade vs. farther into the future); and (iii) the energy and resources that should be devoted to such risks, relative to non-existential risks.

For example, some have argued that AI technology could advance in such a way that it cannot be governed in accordance with human values. For example, the technology may become too complex for explicit encoding of human values, the system might reject such controls, or the speed of advanced AI development may take society by surprise, not giving enough time to develop sufficient controls. If the AI system had significant capabilities or influence, then it could pose a significant risk of existential threat, perhaps even in just a few years. A different position argues that this possibility is quite unlikely, whether because of proposed technical limitations or existing, well documented issues with artificial intelligence – such as bias or ‘hallucinations’. These arguments also sometimes contend that existential risks might be possible, but only very far into the future.

In addition to disagreements about the likelihood and timescale of the like are also difficult questions about how to compare risks of existential threat against currently-known risks and harms. Some positions hold that *any* existential risk warrants extreme responses now; others hold that existing, established harms should receive more attention than hypothetical scenarios. Regardless of one’s position on this point, there is agreement about the importance of building institutional knowledge and processes to address future rapid advances in AI technology.

[39] 2104.10350.pdf (arxiv.org)
[40] Generative AI's Hidden Cost: Its Impact on the Environment | Nasdaq

Current Technical And Product Related Guardrails for FMs and Applications Built on FMs

1. Are there existing methodologies or benchmarks to evaluate and assess (accuracy, safety etc) FMs?

There are a number of metrics to compare performance across FMs including:

- Stanford HELM benchmark[41]
- Open LLM leaderboard[42]
- EluetherAI Harness framework to test generative models on a large number of tasks.[43]

However, there is a strong need for standardization and additional precision in the benchmarks alongside greater discussion of tradeoffs that sometimes occur on performance between benchmarks. It is also worth noting that there is an open need to understand the capabilities of limitations of AI/ML models in the context of AI systems and applications (currently benchmarks focus on the inherent capabilities of the models themselves).

2. What transparency currently exists in the development and deployment of FMs?

Various FMs providers are at various points in the process of articulating how these models are developed and implemented. There exist certain pressures related to safeguarding their intellectual property, a reasonable concern considering the significant expense associated with building these models. In response to this, many providers disclose information about their data resources and offer evaluations detailing the specific use cases for which the models will be utilized, along with their performance metrics (akin to a model evaluation card tailored to specific use cases).

[41] Holistic Evaluation of Language Models (HELM) (stanford.edu)
[42] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
[43] https://github.com/EleutherAI/lm-evaluation-harness

3. How do companies across the stack currently govern their own FMs? What (potential) best practices (could) exist for governance of these technologies?

There are methods such as reinforcement learning from human feedback (RLHF) which incorporates human feedback in the training of the models. In addition, there have been efforts by FM providers to red-team at scale and evaluate and improve the performance and behavior of the FMs before release. To ensure the FMs do not produce toxic, harmful, and biased content, organizations have incorporated technical and rule based safeguards to prevent models from generating inappropriate content. For example, Microsoft Azure provides OpenAI Transparency Note: Transparency Note for Azure OpenAI - Azure Cognitive Services | Microsoft Learn.[44] Similarly, OpenAI has put forth safety standards and best practices on the use of their FMs.[45] Recent voluntary AI governance commitments announced by the White House[46] include external red-teaming, bug-bounty programs (and other similar external efforts to encourage communities to identify safety issues in models), and efforts to identify watermarks and other technology to enable users to recognize AI-generated content.

4. How do app developers currently mitigate risk of adverse outcomes with apps built on FMs?

Many methodologies exist to improve accuracy and minimize adverse outcomes. For example, grounding and fine-tuning a model by utilizing a particular body of data enables the model to be more factual and reduce hallucination. Furthermore, enforcing the FMs to provide confidence scores and citations for answers can help with identifying hallucination by the models. In addition, developers can screen prompts by the users and responses from the FMs for identifying bias, toxicity, malicious, criminal content. There are also evolving methods to restrict or narrow the types of prompts users can input and “system prompts” which guide the system on how to respond to the prompts.

5. How do FM creators and/or app developers lower the risk of biased or toxic content?

A combination of methods are applied to lower the risk of toxic content such as simple search of the content for inappropriate words a.k.a blocklisting, utilizations of models specifically trained to identify toxic content, adversarial testing/red-teaming, and reinforcement learning. In many cases models are instructed not to respond to prompts containing or seeking toxic content. Bias mitigation follows a similar methodology in many cases however often requires even more specific attention to the context of apps built on top of a model – and additional safeguards to guard bias based on that context, including post-processing techniques. It is currently unknown exactly what procedures are used to reduce the presence or impact of toxic content in the input/training data.

[44] https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note?tabs=text

[45] Safety standards (openai.com)

[46] https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

6. Are there current US Federal and/or state regulations that apply to FMs?

In most cases FMs are being provided or deployed with significant caveats, disclaimers, and product use limitations. As a result it can be hard to decipher which regulations constrain use of FMs in a substantive way.

That said, it is important to note that laws do not exempt AI, let alone foundation models. Indeed, numerous existing regulatory protections apply to the use of FMs. This is particularly the case for civil rights protection in the use of AI for decision-making in sensitive realms such as employment, creditworthiness and the like. Privacy regulations (e.g HIPAA, FERPA, FCRA) apply to the input data, though the exact data sources are typically not known. "Truth in advertising" labeling may place constraints on how FMs are represented. (For a fuller list of how existing regulations apply, see the following NAIAC publication: “Rationales, Mechanisms, and Challenges to Regulating AI: A Concise Guide and Explanation[47].”)

Globally, there are several efforts to govern FMs including in the European Union, United Kingdom, and China not to mention multilateral efforts at the OECD and G7.

7. Can we/how can we verify content that is produced by AI?

There are existing efforts to detect AI generated content using classification methods. OpenAI and others have launched classifiers to identify AI-written content and distinguish them from Human written content.[48][49][50] Similar concepts have been used to develop tools to detect AI generated images[51] and speech.[52] It is worth mentioning that these tools are prone to mistakes. On the developers’ side, there are nascent efforts to verify the factuality of FM outputs.

8. Are there existing standards that do or could apply to these FMs? Can NIST AI RMF be used to manage, monitor and mitigate FM related risks?

The White House’s Blueprint for an AI bill of rights establishes principles to provide guidance on how to implement and deploy AI in a responsible way. Similarly, NIST has released its AI risk management framework which is intended to help improve the ability for organizations to incorporate trustworthiness considerations into the design, development, and deployment of AI systems. The principles and the AI RMF provided by NIST can be applied to FMs, however, it requires further evaluation to ensure the framework addresses the risks uniquely relevant to FMs.

[47] https://www.ai.gov/wp-content/uploads/2023/07/Rationales-Mechanisms-Challenges-Regulating-AI-NAIAC-Non-Decisional.pdf

[48] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

[49] https://copyleaks.com/ai-content-detector?fbclid=IwAR2Zip7fBI4ZRrw0dyWQvaMrN846Tvzx713eNslnUVkSoWbjgCCv8_FIqpw

[50] https://writer.com/ai-content-detector/

[51] https://huggingface.co/spaces/umm-maybe/AI-image-detector

[52] https://arxiv.org/pdf/2209.03143.pdf

9. What collective efforts exist around creating technical standards and/or probing the capabilities and limitations of large-scale models.

One recent effort in this vein happened under the aegis of DEF CON – under which thousands of hackers were invited to identify bugs and biases in LLMs.[53] Additionally, IEEE Standard Association offers standards, training and education, certification programs, and more, to empower stakeholders designing, developing, and using Autonomous Intelligent Systems (AIS)[54]. A National Academies Study is currently exploring the trustworthiness of machine learning, especially very large or complex models, in safety-critical applications.[55]

10. What information could be provided to "downstream" developers & users of an FM about the capabilities, limitations, etc. to enable effective & ethical use of the FM? How does this information potentially impact liability and other legal issues?

One possibility would be the use of model cards or similar summaries of the inputs and outputs of a complex model. However, these types of summaries have historically been used for more specialized or single-purpose models, and so may require significant modification to be usable for more general-purpose models such as FMs. Alternatively, the creator of the FMs could be required or expected to provide clear descriptions of the permissible uses or contexts as part of their agreements with the "downstream" developers.

Acknowledgements

With thanks to NAIAC members:

Daniel Ho
Christina Montgomery
Fred Oswald
Victoria Espinel
Liz O'Sullivan

Thank you to the following individuals for their research and support in creating this document. All contributions made by non-Members have been performed under the supervision of a NAIAC Member.

Christina Zhang (Paula Goldman)
Danielle Gilliam Moore (Paula Goldman)

[53] https://www.theregister.com/2023/05/06/ai_hacking_defcon/
[54] https://standards.ieee.org/initiatives/autonomous-intelligence-systems/
[55] Using Machine Learning in Safety-Critical Applications Setting a Research Agenda | National Academies