National Artificial Intelligence Advisory Committee
Briefing Minutes
March 5, 2024

The National Artificial Intelligence Advisory Committee (NAIAC) held a virtual public briefing from 10:00 a.m. to 1:00 p.m. on Tuesday, March 5, 2024. The briefing was recorded and is available online.

NAIAC Committee Members Present

- David Danks
- Victoria Espinel
- Paula Goldman
- Susan Gonzales
- Janet Haven
- Jon Kleinberg
- Ramayya Krishnan
- Christina Montgomery
- Liz O'Sullivan
- Reggie Townsend
- Miriam Vogel (Chair)

NIST NAIAC Staff Members Present

- Cheryl Gendron, Designated Federal Officer (DFO)
- Melissa Taylor, NAIAC Program Manager

Meeting Minutes

Opening Remarks

- Gendron called the meeting to order at 10:05 am EST and confirmed that the committee operates under the Federal Advisory Committee Act. Gendron noted that the meeting is open to the public via live stream and encouraged the public to contact NAIAC by emailing naiac@nist.gov.

- Taylor thanked NAIAC members for their service advising the President and the National AI Initiative Office, thanked members of the public for their participation, and shared the NAIAC email and mailing list.

- Vogel thanked the Safety, Trust, and Rights working group (WG) for coordinating the meeting and invited speakers for sharing their expertise and insights to inform NAIAC’s work.

- Haven thanked the panelists for their participation and noted that the breadth of their experience spans academia, industry, and civil society. Haven outlined the goal of the meeting, which was to explore both conceptual frameworks and practical approaches for advancing AI safety. Haven noted that this discussion might inform ongoing policy work on AI safety, particularly activities conducted at the U.S. AI Safety Institute (US AISI), which is charged with developing standards to ensure the safety of AI systems.

Panel One: Concepts in AI Safety – Scoping "Safety" in AI

Moderator: *Janet Haven,* Co-Chair of the NAIAC's Safety, Trust, and Rights Working Group

Invited Briefers

- **Ms. Inioluwa Deborah Raji,** Fellow, Mozilla and UC Berkeley
- **Dr. Vincent Conitzer,** Professor of Computer Science, Director, Foundations of Cooperative AI Lab (FOCAL), Carnegie Mellon University
- **Dr. Chris Meserole,** Executive Director, Frontier Model Forum
- **Dr. Arvind Narayanan,** Professor of Computer Science, Princeton University
- **Ms. Julia Angwin,** Founder, Proof News
- **The Honorable John C. ("Chris") Inglis,** inaugural U.S. National Cyber Director
- **Dr. Suresh Venkatasubramanian,** Professor of Computer Science, Director, Center for Technological Responsibility, Reimagination, and Redesign, Brown University

Presenter Remarks

- Presenters were invited to give prepared remarks to the Committee. Each presenter's remarks may be viewed in full in the accompanying briefing record.

    - **Inioluwa Deborah Raji,** Fellow, Mozilla and UC Berkeley, introduced the concept of *engineering responsibility*—the obligation engineers have to ensure the safety of the systems they build—and identified strategies for promoting AI safety that link it to responsible engineering practices.

    - **Vincent Conitzer,** Professor of Computer Science and Director of FOCAL, Carnegie Mellon University, listed a range of current AI safety concerns and outlined key factors that impact the safety of an AI system: its use context, access to data and resources, capabilities, and the intentions of its human user.

    - **Chris Meserole,** Executive Director of the Frontier Model Forum, defined the scope of AI safety as encompassing a broad range of (1) risk types, including sociotechnical risks as well as threats to public safety and critical infrastructure, and (2) risk horizons, including both known risks and developing or future risks.

    - **Arvind Narayanan,** Professor of Computer Science, Princeton University, asserted that, because AI models are general-purpose technologies with many potential applications, safety interventions to mitigate the threat of misuse may be most effective if concentrated in the setting of model deployment rather than in the architecture of the models themselves.

    - **Julia Angwin,** Founder of Proof News, called for independent panels of subject matter experts to evaluate context-specific AI model performance; she shared recent work convening a panel of election experts who found that AI models demonstrate significant inaccuracies in answering questions about election processes.

  - **John C. Inglis,** inaugural U.S. National Cyber Director, observed that AI systems comprise technology, human users, and governance frameworks, all of which must work in tandem to ensure AI safety.

  - **Suresh Venkatasubramanian,** Professor of Computer Science and Director, Center for Technological Responsibility, Reimagination, and Redesign, Brown University, advocated for a sociotechnical framing of AI safety that centers human needs and experiences of technology, and uses these to guide technology design and use.

Question and Answer Session

- Haven thanked the presenters and invited NAIAC members to ask follow-up questions.

  - A member asked how AI biases might be evaluated and monitored if AI developers do not provide independent auditors access to the data underlying AI systems.

  - A member asked speakers to outline frameworks for prioritizing the allocation of resources – particularly government resources – to mitigate AI risks.

  - A member noted that some of the risks posed by AI might be entirely new. Others might derive from its magnification of existing risks (e.g., phishing, misinformation, CBRN threats) and its expansion of the set of users who might exploit them. The member asked speakers how approaches to AI safety should identify and address both novel and existing risks.

Panel Two: Operationalizing AI Safety – Methodologies and Organizational Practice

Moderator: *Jon Kleinberg*

Invited Briefers

- **Dr. William Isaac,** Research Scientist, Deep Mind
- **Ms. Miranda Bogen,** Director, AI Governance Lab at the Center for Democracy & Technology
- **Dr. Angela Jiang,** Global Affairs and Product, Open AI
- **Dr. Tamara Kneese,** Project Director, Algorithmic Impact Methods Lab, Data & Society
- **Dr. Joshua A. Kroll,** Assistant Professor of Computer Science, Naval Postgraduate School
- **Ms. Madhulika Srikumar,** Head of Safety, Partnership on AI
- **Dr. Hoda Heidari,** K&L Gates Career Development Assistant Professor in Ethics and Computational Technologies, Carnegie Mellon University
- **Dr. Yejin Choi,** Wissner-Slivka Chair of Computer Science, University of Washington

Presenter Remarks

- Kleinberg thanked the panelists for sharing their perspectives on the operationalization of AI safety and offering approaches that address technical, social, and organizational considerations across the AI development and deployment pipeline.

- Presenters were invited to give prepared remarks to the Committee. Each presenter's remarks may be viewed in full in the accompanying briefing record.

    - **William Isaac,** Research Scientist at Deep Mind, suggested that evaluations of AI safety should address (1) model performance, (2) human interaction with the model, and (3) societal and institutional impacts of widespread model use. Isaac shared recent work that surveys existing generative AI evaluation methods and finds that most address only model performance and limit their scope to text-based models.

    - **Miranda Bogen,** Director, AI Governance Lab at the Center for Democracy & Technology, observed that AI risks could stem from simple systems and the compounding effects of minor system failures as well as from more advanced systems. Bogen outlined organizational and technical infrastructures to address AI risks and called for multidisciplinary approaches to AI risk detection and measurement.

    - **Angela Jiang,** Global Affairs and Product, Open AI, explained that Open AI addresses current, emerging, and longer-term AI safety considerations through distinct teams, and shared several best practices for AI safety: conducting robust evaluations of AI to ground risk mitigation, learning from human feedback, and incorporating public input in the design and use of AI models.

    - **Tamara Kneese,** Project Director, Algorithmic Impact Methods Lab, Data & Society, asserted that AI evaluation requires multidisciplinary and participatory methods that engage stakeholders from communities impacted by AI use; Kneese underscored the importance of conducting privacy and human rights impact assessments in the process of AI evaluation.

    - **Joshua A. Kroll,** Assistant Professor of Computer Science, Naval Postgraduate School, conceptualized AI as existing within a bureaucratic framework of policies, best practices, and training and evaluation processes. Kroll concluded that mitigating AI risk will require safety, oversight, and accountability mechanisms capable of addressing systemic concerns rather than focusing more narrowly on model behavior.

    - **Madhulika Srikumar,** Head of Safety, Partnership on AI, shared three key insights from Partnership on AI's release of its Guidance for Safe Foundation Model Deployment: (1) AI safety practices should be tailored to models' capabilities and release strategies, (2) AI safety standards should address a wide range of potential risks, and (3) advancing safe AI requires the participation and collaboration of diverse stakeholders.

    - **Hoda Heidari,** K&L Gates Career Development Assistant Professor in Ethics and Computational Technologies, Carnegie Mellon University, shared recent work that surveyed recent tech company red-teaming efforts and found a significant lack of consensus on red-teaming scope, structure, and assessment criteria; Heidari then outlined several considerations to incorporate into the planning, execution, and follow-up to red-teaming exercises.

  - **Yejin Choi,** Wissner-Slivka Chair of Computer Science, University of Washington, suggested that aligning AI systems with the plurality of human values, rights, and obligations will require transparency and open science, as well as multidisciplinary collaboration that includes scholars in the humanities and sciences outside of AI.

Question and Answer Session

- Kleinberg thanked the presenters and invited NAIAC members to ask follow-up questions.

  - A member asked participants to share approaches for involving marginalized communities and those impacted by AI use in the AI evaluation process.

  - A member noted that widespread AI literacy is crucial to promote greater awareness of AI's applications and potential impacts – topics of particular importance for groups vulnerable to its potential harms.

  - A member asked how the distribution of AI development and deployment across multiple organizations might affect the evaluation of AI risks and the institution of safety measures.

  - A member asked how the assessment of AI safety should balance its potential benefits and risks.

Briefing Session Conclusion

- Kleinberg recognized the wide range of perspectives presenters shared. He underscored the panel’s consensus that implementing AI safety is a sociotechnical issue that will require a multidisciplinary approach and the coordination of individuals, institutions, technology, and policy.

- Haven thanked the panelists for their remarks, Kleinberg for helping to facilitate the meeting and members of the public for their participation. Haven urged members of the public to contact NAIAC at naiac@nist.gov to share feedback on the topics discussed.

NAIAC Feedback on Executive Order 14110

- An overview of NAIAC members' feedback on Executive Order (EO) 14110 was given as follows:

- There were several themes that characterize NAIAC’s feedback on EO 14110, which underscores the importance of (1) engaging with a wide range of stakeholders in the development of AI systems, (2) developing evidence-based policy and regulations, (3) developing dynamic standards that can adapt to changing technologies, (4) ensuring that individuals maintain meaningful decision-making power within the AI ecosystem, (5) identifying best practices for AI use in specific contexts, and (6) developing internationally interoperable standards and best practices.

- Vogel invited NAIAC members to comment further on these topics. Seeing no further comments, Vogel moved to closing remarks.

Closing Comments

- Vogel thanked Haven and the Safety, Trust, and Rights Working Group for coordinating the briefings and all participants for joining. Vogel noted that video recordings and written summaries of prior NAIAC public briefings are available at ai.gov/naiac.

- Gendron encouraged members of the public to submit comments and questions to the Committee and Subcommittee by emailing naiac@nist.gov. Gendron also directed members of the public to the Committee website, ai.gov/naiac, to view a summary of the meeting and subscribe to the Committee mailing list. Gendron adjourned the meeting at 1:10 pm Eastern Time.

National Artificial Intelligence Advisory Committee
Public Comments
March 5, 2024

The National Artificial Intelligence Advisory Committee (NAIAC) held a virtual public meeting on Tuesday, March 5, 2024. The meeting was recorded and is available online. The following pages are public comments as well as additional speaker comments received connected to this meeting.

From: JB Herrera
To: naiac
Subject: Following up on NAIAC Public Briefing on AI Safety
Date: Tuesday, March 5, 2024 1:14:41 PM
Attachments: image.png

Thank you so much for this briefing.

I am a business mentor to small businesses and Founder/CEO of a consulting firm that uses AI to augment strategy, tactics and implementation.

I'm curious to discover what NAIAC is doing for small businesses? The concept of safety at the global level is daunting, but as a businessman with boots on the ground I can confirm that everyday users need practical guideposts. I'd like to know if there is a small business group I can participate in to help frame the conversation and make suggestions about how to use AI in small business.

Thanks again for the briefing.

Best regards,

JB Herrera
FOUNDER/CEO
INSIGHT DRIVEN BUSINESS
Human Crafted | AI Augmented

JB@INSIGHTDRIVEN.BUSINESS
858.800.2304 - OFFICE
408.505.2532 - WHATSAPP

To conveniently schedule a call, go to the following web page: https://insightdriven.business/schedule/. Thanks and hope to talk with you soon!

Facebook- https://www.facebook.com/jbherrera.2022
LinkedIn- https://www.linkedin.com/in/jbherrera/
X- https://twitter.com/jbherrera
Instagram- https://instagram.com/aidrivenstrategy

From:	Bob Marcus
To:	naiac
Subject:	Operationalizing Safety for Generative AI Software
Date:	Wednesday, March 6, 2024 3:28:49 AM
Attachments:	USAISI Role in Operationalizing AI Safety.pdf

This message is in response to the NAIAC Virtual Meeting today.
NIST's AI Risk Management Framework is a good first step towards operationalizing AI safety within an AI development organization. However due to the wide range of possible applications, the complexity of Generative AI software, and domain-specific extensions; it will not be possible for a development organization to eliminate downstream errors. To operationalize safety, it will be necessary to create well-defined mechanisms and processes for finding and tracking errors after initial releases and including downstream applications.

I believe that there are two essential mechanisms that can increase the safety and trustworthiness of Generative AI software.These are structured vendor-independent External Red Team Testing and a publicly accessible Incident Tracking Database

I have attached a 14 page note describing possible roles for NIST's US AI Safety Institute in these two areas. The purpose of the note is to stimulate discussion on External Red Team Testing and an Incident Tracking Database. There are also links to some related detailed discussions from industry, consortia, and academia.

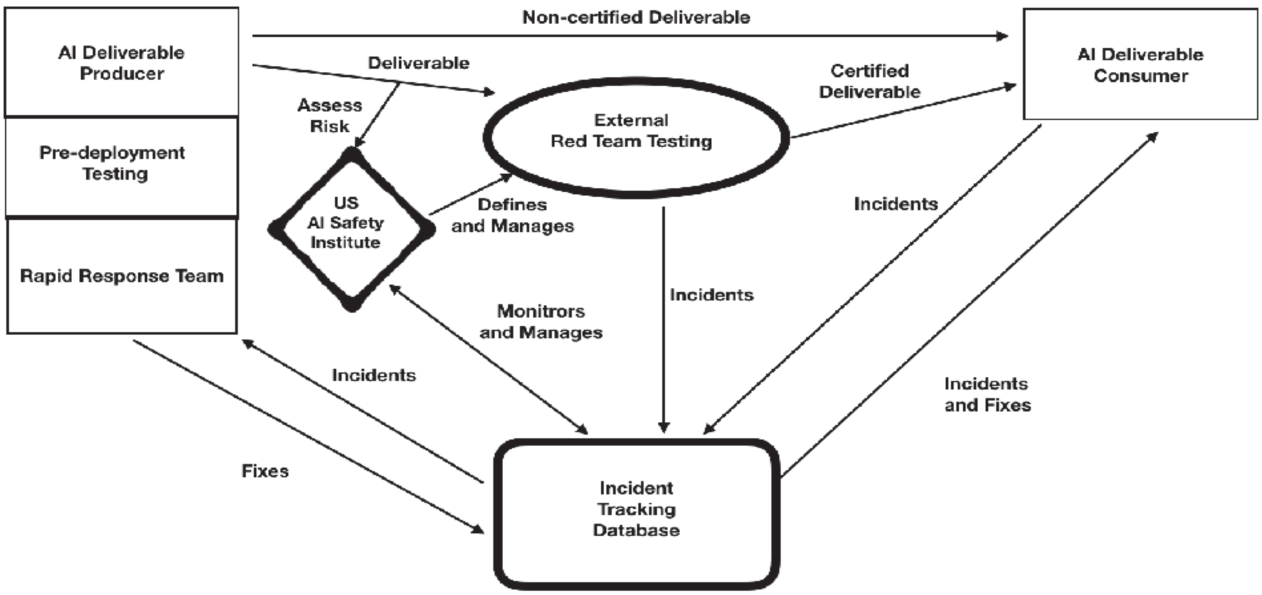
Robert Marcus

Former Co-Chair of NIST Big Data Public Working Group

US AI Safety Institute Role in Operationalizing AI Safety

by Bob Marcus (robert.marcus@gmail.com)

The diagram below illustrates a possible role for the USAISI in enabling trustworthy AI. See Table of Contents on page 2 for a detailed discussion outline.



The US AI Safety Institute should encourage compliance with NIST's AI Risk Management Framework by AI Deliverable Producers. However this will not be sufficient to enable Trustworthy AI due to the complexity and diverse use cases of many Generative AI deliverables. Deliverables can be many outputs in the Generative AI Delivery Process including data sources, foundation models, fine tuned packages, deployed applications, and application output. .

There are two other components that will be essential for increasing reliability in AI applications. These are **External Red Team Testing** for risky deliverables and an **Incident Tracking Database** for problems discovered in testing and use of AI software. Both of these will be necessary for the AI Deliverable Consumers to have confidence that the deliverables have been thoroughly tested and that problems are being addressed and fixed. The US AI Safety Institute can play a role in initiating, monitoring, and managing these components.

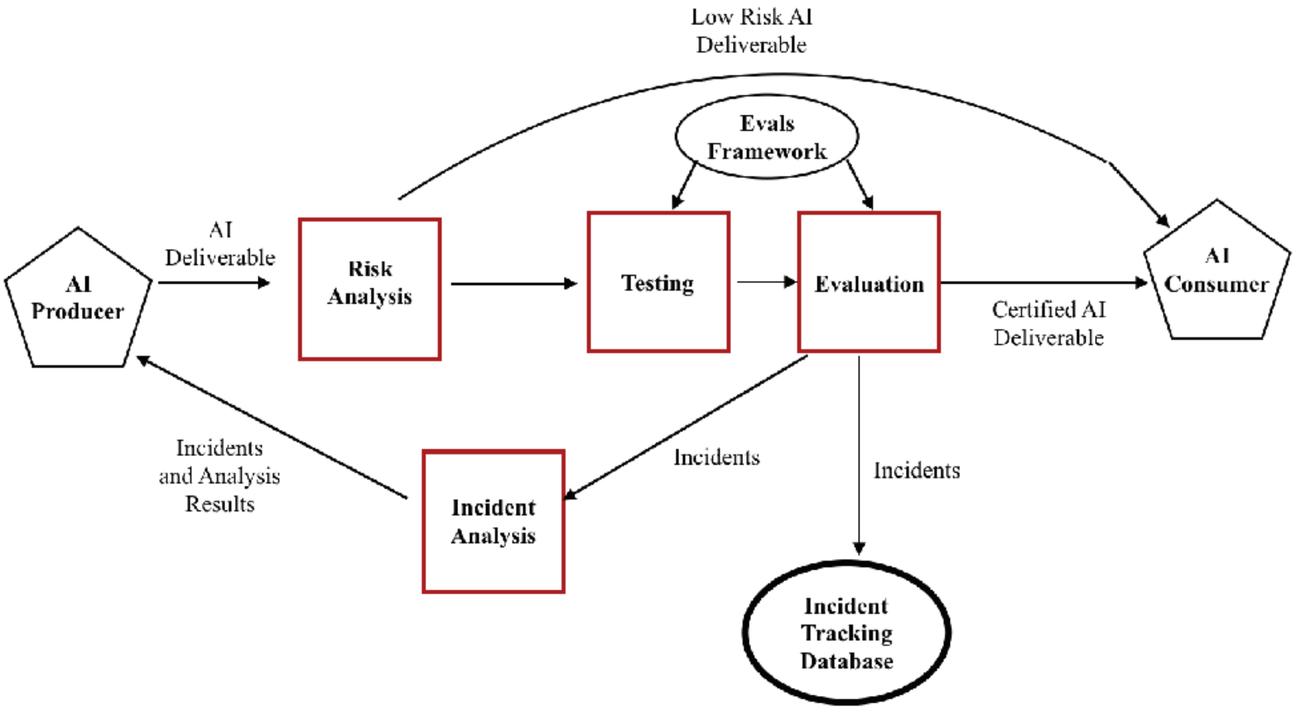
The risk associated with AI deliverables should be evaluated by a risk evaluation team. High risk deliverables should be subjected to External Red Team Testing to uncover possible problems (i.e. incidents). The extent of testing should depend on the risk associated with the deliverable. Deliverables that pass testing can be certified to increase Consumer confidence. Low risk applications can bypass the External Red Team Testing. Incidents discovered by the External Red Team or the AI Consumer should be added to the Incident Tracking Database and reported to the AI Producer. Incident fixes when available should be added to the Incident Tracking Database.

Table of Contents

1. External Red Team Testing 2
1.1 Risk Analysis 3
1.2 Pre-Defined Evals and Testing Frameworks 4
1.3 Lower Risk: Generic Applications and Use Cases 5
1.4 Higher Risk: Domain-specific Applications and Use Cases 8
1.5 Highest Risk: Applications that Change Environment 9
1.6 Incident Analysis and Fixes 10
2. Incident Tracking Database 11
2.1 Incident Tracking Databases 12
2.2 Generative AI Delivery Process 13
2.2 LLM Interfaces to Database 13
2.3 Notifications 14

My comments are in blue.

1. External Red Team Testing



The boxes marked in red in the diagram above are steps in External Red Team Testing. The External Red Teams could generate prompts, evaluate responses, certify, report incidences, and suggest fixes. The individual steps will be discussed below in more detail. The text in italics and quotation marks are from the linked Web site

Generative AI Trust and Governance from Singapore's AI Verify Foundation
https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf

"This discussion paper proposes ideas for senior leaders in government and businesses on building an ecosystem for the trusted and responsible adoption of generative AI.. The practical pathways for governance in this paper seek to advance the global discourse and foster greater collaboration to ensure generative AI is used in a safe and responsible manner, and that the most critical outcome — trust — is sustained"

1.1 Risk Analysis

Open AI Preparedness Framework
https://cdn.openai.com/openai-preparedness-framework-beta.pdf

"We believe the scientific study of catastrophic risks from AI has fallen far short of where we need to be.To help address this gap, we are introducing our Preparedness Framework, a living document describing OpenAI's processes to track, evaluate, forecast, and protect against catastrophic risks posed by increasingly powerful models."

OpenAI Preparedness Team
https://openai.com/safety/preparedness

*"**We will establish a dedicated team to oversee technical work and an operational structure for safety decision-making.** The Preparedness team will drive technical work to examine the limits of frontier models capability, run evaluations, and synthesize reports. This technical work is critical to inform OpenAI's decision-making for safe model development and deployment. We are creating a cross-functional Safety AdvisoryGroup to review all reports"*

"We have several safety and policy teams working together to mitigate risks from AI. Our Safety Systems team focuses on mitigating misuse of current models and products like ChatGPT. Super alignment builds foundations for the safety of super intelligent models that we (hope) to have in a more distant future. The Preparedness team maps out the emerging risks of frontier models, and it connects to Safety Systems, Super alignment and our other safety and policy teams across OpenAI."

Risk Taxonomy, Mitigation, and Assessment Benchmarks of LLM Systems
https://arxiv.org/abs/2401.05778

" In this paper, we delve into four essential modules of an LLM system, including an input module for receiving prompts, a language model trained on extensive corpora, a toolchain module for development and deployment, and an output module for exporting LLM-generated content. Based on this, we propose a comprehensive taxonomy, which systematically analyzes potential risks associated with each module of an LLM system and discusses the corresponding mitigation strategies.

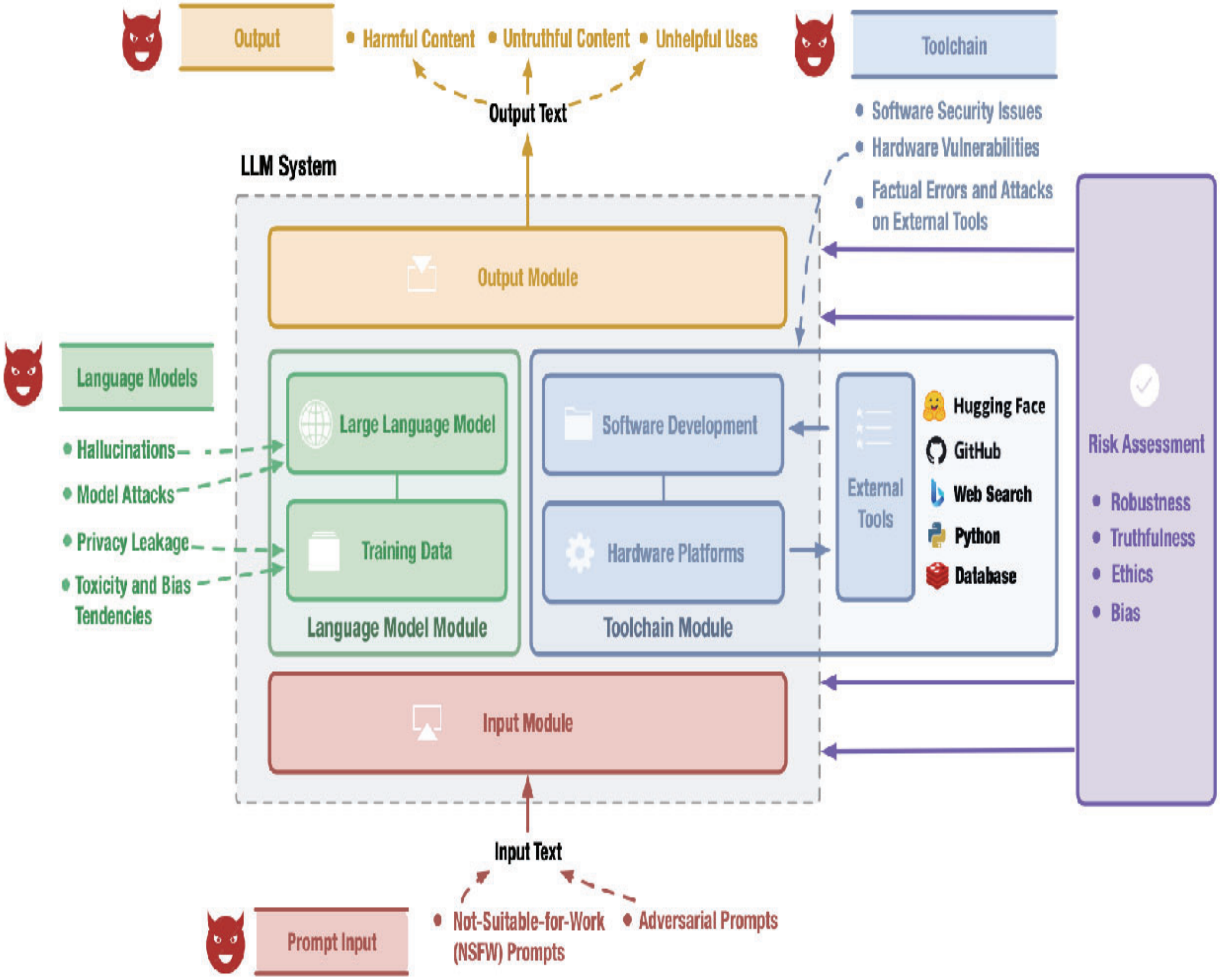


Fig. 2. The overview of an LLM system and the risks associated with each module of the LLM system. With the systematic perspective, we introduce the threat model of LLM systems from five aspects, including prompt input, language models, tools, output, and risk assessment.

1.2 Pre-Defined Evals and Testing Frameworks

AI Verify Foundation from Singapore

https://aiverifyfoundation.sg/

*"A global open-source community that convenes AI owners, solution providers, users, and policymakers, to build trustworthy AI. The aim of AIVF is to harness the collective power and contributions of an international open-source community to develop Artificial Intelligence ("**AI**") testing tools to enable the development and deployment of trustworthy AI. Ai Verify is an AI governance testing framework and software toolkit that help industries be more transparent about their AI to build trust"*

The USAISA could work with Singapore's AI Verify Foundation on testing frameworks

OpenAI Evals

https://portkey.ai/blog/decoding-openai-evals/

*"An **eval** is a task used to measure the quality of output of an LLM or LLM system. Given an input prompt, an output is generated. We evaluate this output with a set of ideal_answers and find the quality of the LLM system. If we do this a bunch of times, we can find the accuracy.*

While we use evals to measure the accuracy of any LLM system, there are 3 key ways they become extremely useful for any app in production.

1. ***As part of the CI/CD Pipeline***
   Given a dataset, we can make evals a part of our CI/CD pipeline to make sure we achieve the desired accuracy before we deploy. This is especially helpful if we've changed models or parameters by mistake or intentionally. We could set the CI/CD block to fail in case the accuracy does not meet our standards on the provided dataset.

2. ***Finding blind-sides of a model in real-time***
   In real-time, we could keep judging the output of models based on real-user input and find areas or use-cases where the model may not be performing well.

3. ***To compare fine-tunes to foundational models***
   We can also use evals to find if the accuracy of the model improves as we fine-tune it with examples. Although, it becomes important to separate out the test & train data so that we don't introduce a bias in our evaluations."

Anthropic Datasets

https://github.com/anthropics/evals?ref=portkey.ai

"This repository includes datasets written by language models, used in our paper on 'Discovering Language Model Behaviors with Model-Written Evaluations.'

'We intend the datasets to be useful to:

1. *Those who are interested in understanding the quality and properties of model-generated data*
2. *Those who wish to use our datasets to evaluate other models for the behaviors we examined in our work (e.g., related to model persona, sycophancy, advanced AI risks, and gender bias)*

The evaluations were generated to be asked to dialogue agents (e.g., a model fine tuned explicitly respond to a user's utterances, or a pre-trained language model prompted to behave like a dialogue agent). However, it is possible to adapt the data to test other kinds of models as well"

Cataloging LLM Evaluations by Singapore's AI Verify

https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf

"In advancing the sciences of LLM evaluations, it is important to first achieve: (i) a common understanding of the current LLM evaluation through a standardized taxonomy; and (ii) a baseline set of pre-deployment safety evaluations for LLMs. A comprehensive taxonomy categorizes and organizes the diverse branches of LLM evaluations, provides a holistic view of LLM performance and safety, and enables the global community to identify gaps and priorities for further research and development in LLM evaluation. A baseline set of evaluations defines a minimal level of LLM safety and trustworthiness before deployment. At this early stage, the proposed baseline in this paper puts forth a starting point for global discussions with the objective of facilitating multi-stakeholder consensus on safety standards for LLMs.

Testing Frameworks for LLMs
https://llmshowto.com/blog/llm-test-frameworks

"An Overview on Testing Frameworks For LLMs. In this edition, I have meticulously documented every testing framework for LLMs that I've come across on the internet and GitHub."

Eleuthera LM Evaluation Harness
https://github.com/EleutherAI/lm-evaluation-harness

“This project provides a unified framework to test generative language models on a large number of different evaluation tasks.

Features:

- *Over 60 standard academic benchmarks for LLMs, with hundreds of subtasks and variants implemented.*
- *Support for models loaded via transformers (including quantization via AutoGPTQ), GPT-NeoX, and Megatron-DeepSpeed, with a flexible tokenization-agnostic interface.*
- *Support for fast and memory-efficient inference with vLLM.*
- *Support for commercial APIs including OpenAI, and TextSynth.*
- *Support for evaluation on adapters (e.g. LoRA) supported in HuggingFace's PEFT library.*
- *Support for local models and benchmarks.*
- *Evaluation with publicly available prompts ensures reproducibility and comparability between papers.*
- *Easy support for custom prompts and evaluation metrics.*

The Language Model Evaluation Harness is the backend for 🤗 Hugging Face's popular Open LLM Leaderboard, has been used in hundreds of papers"

Holistic Evaluation of Language Models (HELM)
https://crfm.stanford.edu/2023/12/19/helm-lite.html

“HELM Lite is inspired by the simplicity of the Open LLM leaderboard (Hugging Face), though at least at this point, we include a broader set of scenarios and also include non-open models. The HELM framework is similar to BIG-bench, EleutherAI’s lm-evaluation-harness, and OpenAI evals, all of which also house a large number of scenarios, but HELM is more modular (e.g., scenarios and metrics are defined separately).’

Holistic Testing
https://static.scale.com/uploads/6019a18f03a4ae003acb1113/test-and-evaluation.pdf

“We introduce a hybrid methodology for the evaluation of large language models (LLMs) that leverages both human expertise and AI assistance. Our hybrid methodology generalizes across both LLM capabilities and safety, accurately identifying areas where AI assistance can be used to automate this evaluation. Similarly, we find that by combining automated evaluations, generalist red teamers, and expert red teamers, we’re able to more efficiently discover new vulnerabilities”

Custom GPTs
https://openai.com/blog/introducing-gpts

"We're rolling out custom versions of ChatGPT that you can create for a specific purpose—called GPTs. GPTs are a new way for anyone to create a tailored version of ChatGPT to be more helpful in their daily life, at specific tasks, at work, or at home—and then share that creation with others. Anyone can easily build their own GPT—no coding is required. You can make them for yourself, just for your company's internal use, or for everyone. Creating one is as easy as starting a conversation, giving it instructions and extra knowledge, and picking what it can do, like searching the web, making images or analyzing data."
A risk evaluation and testing framework is needed for Custom GPTs.

1.3 Lower Risk: Generic Applications and Use Cases (LLM and human testing based on Evals)

Red Teaming Language Models using Language Models

https://arxiv.org/abs/2202.03286

"Overall, LM-based red teaming is one promising tool (among many needed) for finding and fixing diverse, undesirable LM behaviors before impacting users."

Discovering Language Model Behaviors with Model-Written Evaluations

https://arxiv.org/abs/2212.09251

"Prior work creates evaluation datasets manually (Bowman et al., 2015; Rajpurkar et al., 2016, inter alia), which is time-consuming and effortful, limiting the number and diversity of behaviors tested. Other work uses existing data sources to form datasets (Lai et al., 2017, inter alia), but such sources are not always available, especially for novel behaviors. Still other work generates examples with templates (Weston et al., 2016) or programmatically (Johnson et al., 2017), limiting the diversity and customizability of examples. Here, we show it is possible to generate many diverse evaluations with significantly less human effort by using LLMs;"

1.4 Higher Risk: Domain-specific Applications and Use Cases (Fine Tuned Testing LLM + Human Domain Experts)

Large Action Models(LAMs)

http://tinyurl.com/33zwkmbb

"LAMs interact with the real world through integration with external systems, such in IoT devices and others. By connecting to these systems, LAMs can perform physical actions, control devices, retrieve data, or manipulate information"

OpenAI External Red Team

https://openai.com/blog/red-teaming-network

"The OpenAI Red Teaming Network is a community of trusted and experienced experts that can help to inform our risk assessment and mitigation efforts more broadly, rather than one-off engagements and selection processes prior to major model deployments. Members of the network will be called upon based on their expertise to help red team at various stages of the model and product development lifecycle. Not every member will be involved with each new model or product, and time contributions will be determined with each individual member"
A vendor-independent Red Teaming Network of Experts is needed

1.5 Highest Risk: Applications that Change Environment (Simulation or Sandbox Testing)

The environment can be cyber or physical. The changes can be direct or indirect (e.g. code generation, persuasion)

Singapore Generative AI Evaluation Sandbox

https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

"1.The Sandbox will bring global ecosystem players together through concrete use cases, to enable the evaluation of trusted AI products. The Sandbox will make use of a new Evaluation Catalogue, as a shared resource, that sets out common baseline methods and recommendations for Large Language Models (LLM).
2.This is part of the effort to have a common standard approach to assess Gen AI.
3. The Sandbox will provide a baseline by offering a research-based categorization of current evaluation benchmarks and methods. The Catalogue provides an anchor by (a) compiling the existing commonly used technical testing tools and organizing these tests according to what they test for and their methods; and (b) recommending a baseline set of evaluation tests for use in Gen AI products."

Singapore GenAI Sandbox

https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

"Sandbox will offer a common language for evaluation of Gen AI through the Catalogue Sandbox will build up a body of knowledge on how Gen AI products should be tested Sandbox will develop new benchmarks and tests"

Participants in Sandbox include most many AI vendors (not OpenAI)

https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2023/10/generative-ai-evaluation-sandbox/annex-a---list-of-participants-in-sandbox.pdf

USAISI could work with Singapore's AI Verify Foundation on testing sandbox

1.6 Incident Analysis and Fixes

Use Generative AI to automate Incident Analysis

https://www.bigpanda.io/wp-content/uploads/2023/07/bigpanda-generative-ai-datasheet.pdf

"AI-generated summary and title: Identify incidents that require more immediate action by automatically synthesizing complex alert data into clear, crisp incident summaries and titles that can be populated within chat and ITSM tools.

AI-proposed incident impact: Reliably identify the relevancy and impact of incidents across distributed IT systems in clear, natural language within seconds. Easily identify priority actions for ITOps, L2, and L3 response teams across all incidents at scale.

AI-suggested root cause: Automatically surface critical insights and details hidden in lengthy and complex alerts to quickly identify the probable root cause of an incident, as it forms in real-time."

Fixing Hallucinations in LLMs

https://betterprogramming.pub/fixing-hallucinations-in-llms-9ff0fd438e33?gi=a8912d3929dd

"Hallucinations in Large Language Models stem from data compression and inconsistency. Quality assurance is challenging as many datasets might be outdated or unreliable. To mitigate hallucinations:

1. *Adjust the temperature parameter to limit model creativity.*
2. *Pay attention to prompt engineering. Ask the model to think step-by-step and provide facts and references to sources in the response.*
3. *Incorporate external knowledge sources for improved answer verification.*

A combination of these approaches can achieve the best results.

The Rapid Response Team

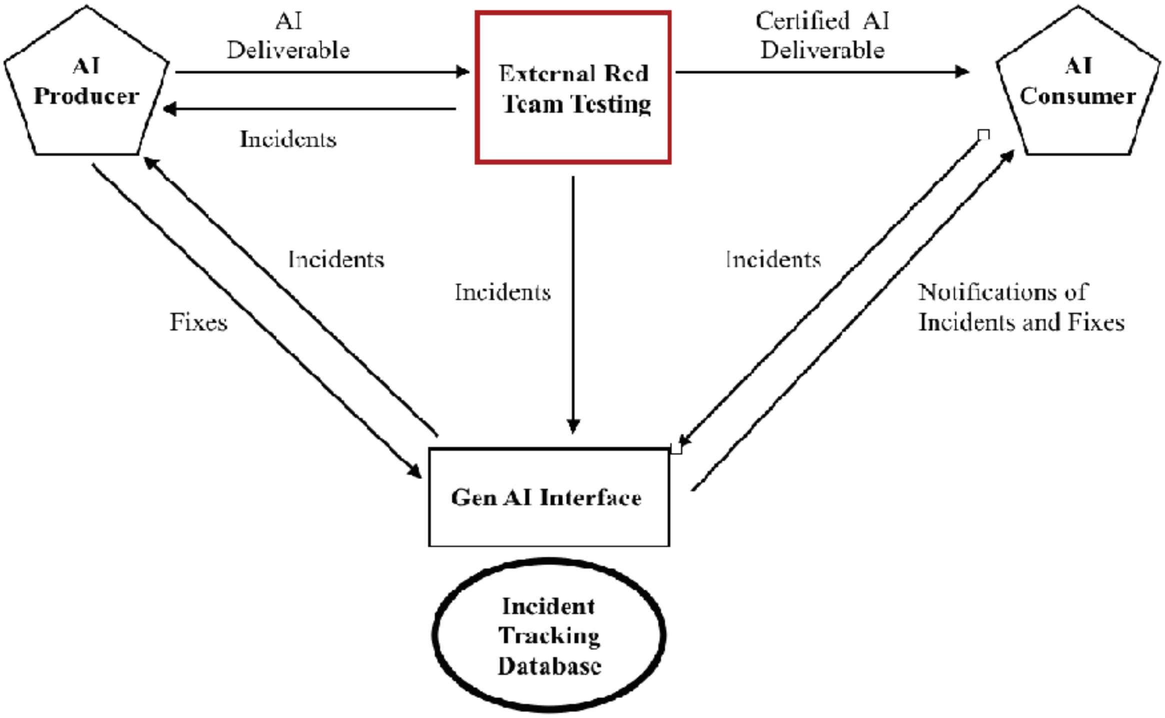
https://www.svpg.com/the-rapid-response-team/

“In these cases, a practice that I have seen make dramatic improvements along both dimensions is to create at least one special dedicated team that we often call the “Rapid Response Team.” This is a dedicated team comprised of a product manager (or at least a part of a product manager), and mainly developers and QA. Usually these teams are not large (2-4 developers is common). This team has the following responsibilities:

- *fix any critical issues that arise for products in the sustaining mode (i.e. products that don’t have their own dedicated team because you’re not investing in them other than to keep it running).*
- *implement minor enhancements and special requests that are high-value yet would significantly disrupt the dedicated team that would normally cover these items.*
- *fix any critical, time-sensitive issues that would normally be covered by the dedicated team, but again would cause a major disruption.”*

The AI Producer should have a Rapid Response team to handle incidents and provide fixes

2. Incident Tracking Database



2.1 Incident Tracking Database

AI Incident Database

https://incidentdatabase.ai/

“The AI Incident Database is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. Like similar databases in aviation and computer security, the AI Incident Database aims to learn from experience so we can prevent or mitigate bad outcomes.

You are invited to submit incident reports, whereupon submissions will be indexed and made discoverable to the world. Artificial intelligence will only be a benefit to people and society if we collectively record and learn from its failings."

Partnership on AI

https://partnershiponai.org/workstream/ai-incidents-database/

“As AI technology is integrated into an increasing number of safety-critical systems — entering domains such as transportation, healthcare, and energy — the potential impact of this technology’s failures similarly grows. The AI Incident Database (AIID) is a tool designed to help us better imagine and anticipate these risks, collecting more than 1,200 reports of intelligent systems causing safety, fairness, or other real-world problems.”

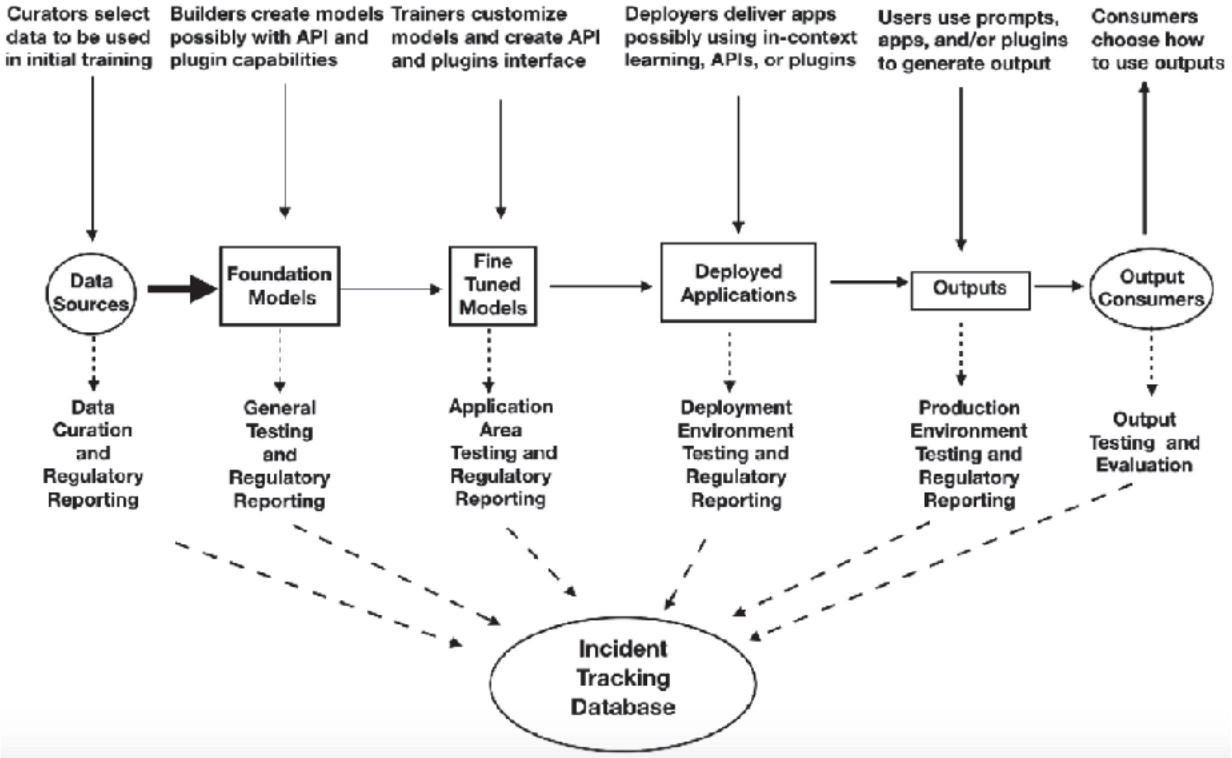
Preventing Repeated Real World AI Failures by Cataloging Incidents

https://arxiv.org/abs/2011.08512

“Mature industrial sectors (e.g., aviation) collect their real world failures in incident databases to inform safety improvements. Intelligent systems currently cause real world harms without a collective memory of their failings. As a result, companies repeatedly make the same mistakes in the design, development, and deployment of intelligent systems. A collection of intelligent system failures experienced in the real world (i.e., incidents) is needed to ensure intelligent systems benefit people and society. The AI Incident Database is an incident collection initiated by an industrial/non-profit cooperative to enable AI incident avoidance and mitigation. The database supports a variety of research and development use cases with faceted and full text search on more than 1,000 incident reports archived to date.”

2.2 Generative AI Delivery Process

AI deliverables can be produced, consumed and generate incidents in many stages of the Generative AI Delivery Process.


Curators select data to be used in initial training
Builders create models possibly with API and plugin capabilities
Trainers customize models and create API and plugins interface
Deployers deliver apps possibly using in-context learning, APIs, or plugins
Users use prompts, apps, and/or plugins to generate output
Consumers choose how to use outputs
Data Sources
Foundation Models
Fine Tuned Models
Deployed Applications
Outputs
Output Consumers
Data Curation and Regulatory Reporting
General Testing and Regulatory Reporting
Application Area Testing and Regulatory Reporting
Deployment Environment Testing and Regulatory Reporting
Production Environment Testing and Regulatory Reporting
Output Testing and Evaluation
Incident Tracking Database

2.3 LLM Interfaces to Database

How LLMs made their way into the modern data stack
https://venturebeat.com/data-infrastructure/how-llms-made-their-way-into-the-modern-data-stack-in-2023/

"The first (and probably the most important) shift with LLMs came when vendors started debuting conversational querying capabilities — i.e. getting answers from structured data (data fitting into rows and columns) by talking with it. This eliminated the hassle of writing complex SQL (structured query language) queries and gave teams, including non-technical users, an easy-to-use text-to-SQL experience, where they could put in natural language prompts and get insights from their data. The LLM being used converted the text into SQL and then ran the query on the targeted dataset to generate answers."

Can LLM Already Serve as A Database Interface
https://typeset.io/questions/can-llm-already-serve-as-a-database-interface-a-big-bench-3gje48fazi

"Large language models (LLMs) have shown impressive results in the task of converting natural language instructions into executable SQL queries, known as Text-to-SQL parsing. However, existing benchmarks like Spider and WikiSQL focus on small-scale databases, leaving a gap between academic study and real-world applications. To address this, the paper "Bird" presents a big benchmark for large-scale databases in the context of text-to-SQL tasks. It contains a large dataset of text-to-SQL pairs and 95 databases spanning various professional domains. The emphasis on database values in Bird highlights the challenges of dirty database contents, external knowledge, and SQL efficiency in the context of massive databases. The experimental results demonstrate the significance of database values in generating accurate text-to-SQL queries for big databases."

Text2SQL
https://medium.com/@changjiang.shi/text2sql-converting-natural-language-to-sql-defa12c2a69f

"Text2SQL is a natural language processing technique aimed at converting natural language expressions into structured query language (SQL) for interaction and querying with databases. This article presents the historical development of Text2SQL, the latest advancements in the era of large language models (LLMs), discusses the major challenges currently faced, and introduces some outstanding products in this field."

2.4 Notifications

Using LLMs for notifications
https://pathway.com/developers/showcases/llm-alert-pathway

"Real-time alerting with Large Language Models (LLMs) like GPT-4 can be useful in many areas such as progress tracking for projects (e.g. notify me when coworkers change requirements), regulations monitoring, or customer support (notify when a resolution is present).

The program that we will create answers questions based on a set of documents. However, after an initial response is provided, the program keeps on monitoring the document sources. It efficiently determines which questions may be affected by a source document change, and alerts the user when a revision - or a new document - significantly changes a previously given answer.

The basic technique of feeding chunks of information from external documents into an LLM and asking it to provide answers based on this information is called RAG - Retrieval Augmented Generations. So, what we are doing here is ***real-time RAG with alerting"***

Gendron, Cheryl L. (Fed)

From: Miranda Bogen <mbogen@cdt.org>
Sent: Thursday, March 7, 2024 3:49 PM
To: naiac
Subject: Re: Thank you + option to submit written comments

Thank you for the opportunity to speak at the panel last week and to submit follow up materials. I'll point you in the direction of two blog posts my team recently published that capture the main thrust of the comments I shared. While they reference NIST's work specifically, the recommendations are more broadly applicable:

- Ensuring NIST's AI Safety Institute Consortium Lives Up to its Potential
- Trustworthy AI Needs Trustworthy Measurements

Best,

Miranda Bogen | Director, AI Governance Lab
Center for Democracy & Technology | **cdt.org**
E: mbogen@cdt.org | **D:** 202.407.8827

---------- Forwarded message ---------
From: **Janet Haven** <janet@datasociety.net>
Date: Wed, Mar 6, 2024 at 6:43 AM
Subject: Thank you + option to submit written comments
To: <chris@chrismeserole.com>, Vincent Conitzer <conitzer@cs.cmu.edu>, deborahraji1@gmail.com <deborahraji1@gmail.com>, inglis6 <inglis6@aol.com>, Arvind Narayanan <arvindn@cs.princeton.edu>, Venkatasubramanian, Suresh <suresh_venkatasubramanian@brown.edu>, <julia@proofnews.org>, <angela@openai.com>, Joshua Kroll <jkroll@jkroll.com>, Madhu Srikumar <madhu@partnershiponai.org>, mirandabogen@gmail.com <mirandabogen@gmail.com>, Tamara Kneese <tkneese@datasociety.net>, William Isaac <williamis@deepmind.com>, Hoda Heidari <hheidari@cmu.edu>, Yejin Choi <yejin@cs.washington.edu>
Cc: Gendron, Cheryl L. (Fed) <cheryl.gendron@nist.gov>, Taylor, Melissa K. (Fed) <melissa.taylor@nist.gov>, Jon Kleinberg <kleinberg@cornell.edu>, Miriam Vogel <miriam.vogel@equalai.org>

Dear all,

Thank you all very much for participating yesterday in NAIAC's public discussion on AI safety. My fellow NAIAC members and I are deeply grateful for your time and for the critical perspective each of you brought to the event. I have heard from many audience members who commented on how much they appreciated hearing a grounded, informed and expert group tackle this issue.

The next step for us on the NAIAC is that we will be writing a finding (based on your comments, others submitted by the public, and additional lit review), and recommendations for the NAIAC to debate, potentially approve and send to the National AI Initiative office and the President. We plan to do this over the next two months, and will let you know when there are further public events or postings related to this topic.

In the meantime, we would absolutely welcome written versions of your comments to supplement your spoken remarks. If you would like to submit your comments in writing, please send them to naiac@nist.gov by March 15th.

Yesterday’s panel was recorded, and video, along with minutes, will be posted on naiac.gov in the coming weeks.

Again, thank you so much for your important contributions yesterday and the work you all do that informed them.
Warm regards,
Janet

Janet Haven
Executive Director
www.datasociety.net | @datasociety

DATA&
SOCIETY

--
Miranda Bogen

I’m the Project Director of Data & Society’s Algorithmic Impact Methods Lab. Previously, I was a Director of Developer Engagement on a software product team at Intel and a professor of media and gender studies, and I am a visiting scholar at UC Berkeley’s Center for Science, Technology, Medicine & Society. To better address existing harms related to AI, at AIMLab we are establishing and experimenting with methods for conducting algorithmic impact assessments in the public interest. Using participatory methods and stakeholder engagement, we advocate for the terms of evaluation to be driven by the communities that are most vulnerable to AI’s impacts. We look at individual harms alongside societal ones. We aim to identify and promote social organizational changes, tools and standards, and regulations to make algorithmic systems more equitable.

Ground up assessments of risk are fundamental to my team’s work—and I believe they’re essential for any effort to promote safe or trustworthy AI. We need to do testing that is both quantitative and qualitative to set rights-protecting standards. And developers need robust, empirical feedback to be able to improve the technologies they are building and deploying, mitigating any potential harms associated with them.

Establishing safety standards requires multidisciplinary expertise, from technical auditing and participatory research methods to public policy and research ethics. Beyond a technical audit, red teaming, or other methods to set and test technical guardrails, it is also important to look at the tech as part of a broader sociotechnical system. In other words, we can’t view AI in isolation; it must be viewed as embedded in larger social structures and power dynamics. When we recognize that AI is a sociotechnical system, we recognize that stakeholder engagement, especially with diverse, marginalized communities, is a crucial part of AI safety.

In some of our pilots, we are engaging community members and having them evaluate systems, too. Red-teaming can be a helpful practice among industry technologists, but we think it’s just as critical to have other forms of accountability based on a community’s own needs and forms of expertise. For example, with one of our partners, the Workers’

Algorithm Observatory, we are using rideshare drivers' own data collection and experiences with algorithmic wage discrimination to provide evidence and develop strategies in collaboration with legal advocates, policy experts, and unions. Algorithmic transparency should lead to action around harm mitigation.

When it comes to evaluating foundation models, we can look to existing frameworks for assessing impacts and risk, including privacy impact assessments and human rights impact assessments. There will always be cutting edge technologies; but just because a technology is new doesn't mean we should abandon the methods and approaches we have now to advance safety and protect rights. We need to have human rights experts in the room. Evaluating general purpose AI requires building on existing templates and incorporating interdisciplinary expertise. For example, in the Biasly project I worked on at Mila AI in Quebec, in order for our model to detect and mitigate subtle forms of misogyny, we needed to have domain experts in linguistics, machine learning, and gender studies weigh in to create inference categories and annotate the dataset. In a workshop AIMLab conducted in partnership with Mozilla and Kwanele, a South African gender based violence advocacy organization, community members' engagement with a general purpose chatbot revealed gaps in South African language recognition and local legal knowledge. Evaluating the potential risks of the model required input from a range of experts beyond machine learning engineers.

Finally, I want to call attention to the importance of user research, UX as an area of specialization, and qualitative lines of inquiry. Even most research-based assessments of foundational models take the shape of checklists regarding bias and risk, and companies and policymakers alike are often quick to say that some models or applications are low risk, without deeply considering the human repercussions along the entire supply chain and life cycle. How can ethnographic knowledge about the more complex aspects of human-computer interaction and the unintended consequences of implementing LLMs be taken up by product and strategy teams and by policymakers? These insights are critical to any kind of AI risk management. It is important to translate the needs and concerns of community members to engineering teams, and to ensure that safety concerns are addressed as soon as possible. Building strong relationships between community organizations and technical teams, for example, is one way to ensure that frameworks are effectively operationalized by companies.

PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY

Members of the National AI Advisory Committee, March 14, 2024

We write to provide a comment that may inform your deliberations on AI safety. This is a revised and elaborated version of remarks by one of us at the NAIAC AI safety panel on March 5. We appreciate the opportunity to provide our input.

Sincerely,

Arvind Narayanan
Professor of Computer Science
Director, Center for Information Technology Policy
Princeton University

Sayash Kapoor
Ph.D. candidate
Center for Information Technology Policy
Princeton University

AI safety is not a model property

The assumption that AI safety is a property of AI models is pervasive in the AI community. It is seen as so obvious that it is hardly ever explicitly stated. Because of this assumption:

- Companies have made big investments in red teaming models before releasing them.
- Researchers are frantically trying to fix the brittleness of model alignment techniques.
- Some AI safety advocates seek to restrict open models given concerns that they might pose unique risks.
- Policymakers are trying to find the training compute threshold above which safety risks become serious enough to justify intervention (and lacking any meaningful basis for picking one, they seem to have converged on 10^{26} rather arbitrarily).

We think these efforts are inherently limited in their effectiveness. That's because AI safety is not a model property. With a few exceptions, AI safety questions cannot be asked and answered at the levels of models alone. Safety depends to a large extent on the context and the environment in which the AI model or AI system is deployed. We have to specify a particular context before we can even meaningfully ask an AI safety question.

As a corollary, fixing AI safety at the model level alone is unlikely to be fruitful. Even if models themselves can somehow be made "safe", they can easily be used for malicious purposes. That's because an adversary can deploy a model without giving it access to the details of the context in which it is deployed. Therefore we cannot delegate safety questions to models — especially questions about misuse. The model will lack information that is necessary to make a correct decision.

Based on this perspective, we make four recommendations for safety and red teaming that would represent a major change to how things are done today.

Safety depends on context: three examples

Consider the concern that LLMs can help hackers generate and send phishing emails to a large number of potential victims. It's true — in our own small-scale tests, we've found that LLMs can generate persuasive phishing emails tailored to a particular individual based on publicly available information about them.

But here's the problem: phishing emails are just regular emails! There is nothing intrinsically malicious about them. A phishing email might tell the recipient that there is an urgent deadline for a project they are working on, and that they need to click on a link or open an attachment to complete some action. What is malicious is the content of the webpage or the attachment. *But the model that's being asked to generate the phishing email is not given access to the content that is potentially malicious*. So the only way to make a model refuse to generate phishing emails is to make it refuse to generate emails. That would affect many non-malicious uses, such as marketing.

We see the same pattern over and over. There has been alarm about LLMs being able to give bioterrorists information on how to create pathogens. But that information is readily available on the internet. The hard parts for would-be bioterrorists are all of the other steps involved: obtaining raw materials, culturing cells in the lab without killing them or infecting oneself, and disseminating the bioweapon to cause harm. AI could potentially aid that work, as it is a general-purpose tool and has some usefulness for almost all knowledge work. Again, this illustrates the limits of attempting to build safety into models: most of the questions the user would ask in this process relate to *synthetic biology in general and not bioweapons in particular*. To be sure that a model couldn't assist bioterrorists, it would have to refuse to assist with any sort of bioengineering.

Or consider the use of LLMs to generate disinformation. Even in the unlikely event that a model could be aligned so that it refuses all requests to generate false information, research has found that true-but-misleading information is far more impactful than false information on social media; 50x more in the case of increasing vaccine hesitancy. So even a hypothetical safe model could be used to aid disinformation efforts: the adversary would use it to generate factual information (e.g. accurately summarizing news stories), with the misleading context added in separately.

In short, trying to make an AI model that can't be misused is like trying to make a computer that can't be used for bad things.

Scope of our claims

Our scope is primarily about misuse, which seems to be the biggest driver of the AI safety worries recently. This includes both malicious misuse, such as the above examples, and nonmalicious misuse, such as students cheating on homework. Here again the model lacks the context to prevent only "bad" uses: it doesn't know whether the task it is given is part of the user's homework.

AI safety encompasses many other types of failures, such as bias and toxicity, accidents, reward hacking, and adversarial inputs (such as prompt injection). These are all different from misuse risks.[1] We think our argument applies in many of these cases, though less strongly. We don't give a full analysis here. In the case of accidents, others have made the point that we have to look at the system and context, rather than the model alone.

Another related failure mode, one that is outside our scope, is overreliance on flawed models for legal or medical advice (whether this falls under AI safety is debatable but tangential to our point). To understand these harms, studying models makes sense: for example, a recent investigation by the AI Democracy Projects found that most models have high rates of incorrect responses to questions about the election.

Even within the category of misuse, there are a few exceptions to the rule that safety is not a model property. Some types of content are intrinsically problematic regardless of what someone does with it, as in the case of AI-generated child sexual abuse material. Aligning AI systems to refuse such requests is important. Outputting memorized copyrighted material is another such category.

In any case, our point is not that red teaming or aligning models is useless, just that safety has to be much broader than looking at models alone.

[1] Prompt injection might enable misuse, but in this document our scope is about misuse that doesn't require violating model safety properties.

Recommendation 1: defenses against misuse must primarily be located outside models

As we've written elsewhere, model alignment can easily be evaded by adversaries. Those evasive techniques, such as jailbreaks, are potentially fixable. Here, we are talking about something more fundamental: misuse that does not require breaching the alignment guarantees in any way, such as writing persuasive emails that can be used for either marketing or phishing.

If model alignment is not the answer, other defenses are sorely needed. As we've consistently argued, defenses should focus on attack surfaces: the downstream sites where attackers use the outputs of AI models for malicious purposes. For example, the best defenses against phishing emails, whether generated by humans or LLMs, are email scanners and URL blacklists – which we've had for a couple of decades and have gradually gotten pretty good, although of course we must continue to improve them.

If we instead keep barking up the tree of model alignment, the fact that the model lacks access to context, and therefore can't make informed safety determinations, will lead to both false positives and false negatives. In other words, it will not only lead to a failure to prevent misuse, but also the opposite problem: refusing innocuous requests like an overzealous censor.

Recommendation 2: assess marginal risk

If safety is not primarily a model property and defenses must reside elsewhere, then there might not be a big difference between the safety implications of open and closed release strategies. In any case, the debate on openness in AI needs a more rigorous risk assessment framework. We were recently part of a large collaboration that presented just such a framework. It enables assessing the *marginal* risk of releasing a model – that is, the additional or incremental risk – compared to the risk from existing models (and non-AI technologies). It takes into account that defenses for some risks might already exist, especially defenses located outside models. Using this framework, we showed that the marginal risk of open models in cybersecurity (specifically, enabling automated vulnerability detection) is low, whereas for the generation of non-consensual intimate imagery, the marginal risk is substantial.

A notable potential safety advantage of closed models is the ability to monitor queries and *retrospectively* identify malicious use. This is a far easier technical problem than building safety into the model itself.[2] Besides, the risk of account suspension or prosecution might exert a deterrent effect on threat actors. In any case, this sort of comparison between open and closed models will have to be made separately for each type of misuse based on empirical evidence. Currently, we don't have reliable evidence of how well monitoring and detection is working because of the lack of transparency by developers. One small but notable exception is a recent blog post on Microsoft's and OpenAI's efforts to detect and disrupt hacking groups' use of LLMs.

Recommendation 3: refocus red teaming toward early warning

We should not expect red teaming to tell us whether or not a model can be misused (the answer is always yes). Instead, we should use red teaming to learn about the advancing frontier of adversary capabilities enabled by state-of-the-art AI models. For example, if AI systems have gotten powerful enough to automate a complex cybersecurity attack chain – scanning social media profiles to gather information, crafting a phishing email, taking over an account, exfiltrating information, and concealing traces of the attack – we need to have early warning of those capabilities so that we can defend appropriately.

To do this, we may need to design better offensive pipelines, such as for hacking. In the case of disinformation, a key offensive capability would be building a bot that can engage in a persuasive conversation on political topics over a long period of time. Building such capabilities raises ethical challenges. The cybersecurity community has long grappled with these challenges, and the general conclusion is that we are better off in a world where everyone has access to offensive capabilities than one in which only attackers do.

The results of red teaming should inform the development of defenses – defenses that almost always will reside *outside* AI models (such as detecting and labeling bot accounts on social media).

[2] There are many reasons why forensics is easier than model alignment. It only requires assessing whether an *account*, which may have made hundreds of queries, has violated usage policies, rather than making assessments query by query, so there is a lot more information to go on. Analysts might also be able to use a materialized attack as evidence of intent: for example, a disinformation campaign on social media that used content generated by a model. Broadly speaking, real-time AI assessment is no match for patient human analysis in making nuanced determinations of malicious use.

Recommendation 4: red teaming should be led by third parties with aligned incentives

The above change in objectives of red teaming leads to a subtle shift in incentives. When red teaming is model focused, developers have an incentive to do a good job. If they find that models produce "dangerous" information, they can fix that behavior, which helps them avoid bad press.

But for the kind of misuse we're talking about, the incentives are reversed. It is not in developers' interest to build the most powerful offensive pipeline possible. If they do, they might find that (for instance) a model can be used for hacking, but they will have no way to prevent this. Thus, they will have to admit that they are knowingly releasing a model that can be used for offensive purposes. It is much better for them to not find out in the first place.

Consider OpenAI's recent study on biological threats from language models. OpenAI evaluated the risk of users gaining access to information using language models compared to the internet (which is much better than previous studies that don't have a baseline at all). But creating a bioweapon requires far more than a few hours of information hunting. A motivated actor needs access to a lab, reagents, and equipment in order to even begin the process. The real question is whether AI can help adversaries acquire these resources. The study does not answer that.

The incentives of third parties are potentially better aligned for a more holistic risk assessment that is less focused on models alone. But here too, there is need for caution. Until recently, much of the evidence for biosecurity risks of language models came from groups funded by a small number of effective altruism organizations. Members of the U.S. House Committee on Science, Space, and Technology recently wrote a letter expressing concern about the lack of transparency in how the National Institute of Standards and Technology and the AI Safety Institute plan to allocate funding for third-party evaluations. They were especially concerned about upholding the standards of scientific research.

Final thoughts: developer responsibility

Why has the myth of safety as a model property persisted? Because it is convenient for everyone! In a world where safety is a model property, companies could confidently determine whether a model is safe enough to release, and AI researchers could apply their arsenal of technical methods toward safety. Most importantly, accountability questions would have relatively clear answers. Companies should have liability for harms if model safety guarantees fail, but not otherwise.

By contrast, accepting that there is no technical fix to misuse risks means that the question of responsibility is extremely messy, and we don't currently have a good understanding of how to allocate liability for misuse. Assuming that retrospective detection is easier (see recommendation 2 above), one low-hanging fruit is to require anyone who hosts a model, whether closed or open, to adhere to certain standards for monitoring and reporting misuse – see our call for generative AI companies to publish transparency reports (and, more generally, the least cost avoider principle). But that won't be enough, and downstream defenses are needed.

Unfortunately, downstream defenses against misuse impose a great cost on the rest of society. For example, the fact that any image or video could be AI-generated means that realism is no longer a marker of authenticity. That means we all need to adapt and change how we assess the veracity of information online. And that's just one of dozens of such adaptations needed.

Morally speaking, developers should bear some of the societal costs of harmful uses of AI, mirroring the fact that they reap profits from beneficial uses of AI. But legally speaking, we have no tool to enforce that. Remedying this situation is the great challenge of AI policy – a point we've made over and over. No amount of "guardrails" will close this gap.

Acknowledgment. We thank Mihir Kshirsagar for feedback on a draft.

Gendron, Cheryl L. (Fed)

From: Dr Cynthia Rudin <cynthia@cs.duke.edu>
Sent: Tuesday, March 12, 2024 8:52 PM
To: naiac
Subject: feedback for AI safety

Hi NAIAC safety team!
I was talking with Cheryl and she suggested I write on this topic to you.

Since we are stuck with deepfakes for the foreseeable future, including very convincing phone scams, we need to figure out ways that people know the sources of information they are receiving. Old people are particularly vulnerable and shouldn't have to suffer with these scams. For phone scams, if we actually knew where the calls we receive originate, it would eliminate most such calls. Yet, we do not know this information. I know that I receive calls that appear to be from Texas and Missouri but are just phone scams from India. This shouldn't happen - each phone number should be associated with a unique source, and no spoofing should be possible. I realize there are some callers who need to call anonymously (e.g., healthcare providers), but they should be able to get permission to receive a "healthcare" code that is visible to the receiver. Given that these annoying phone scams affect *everyone who has a phone,* how are we going to deal with this?

Thanks,
Cynthia