1/3 AI Governance Alliance
Briefing Paper Series 2024

Presidio AI Framework: Towards Safe Generative AI Models

IN COLLABORATION
WITH IBM CONSULTING

Cover image: MidJourney

Contents

Executive summary

The Presidio AI Framework addresses generative AI risks by promoting safety, ethics, and innovation with early guardrails.

The rise of generative AI presents significant opportunities for positive societal transformations. At the same time, generative AI models add new dimensions to AI risk management, encompassing various risks such as hallucinations, misuse, lack of traceability and harmful output. Therefore, it is essential to balance safety, ethics and innovation.

This briefing paper identifies a list of challenges to achieving this balance in practice, such as lack of a cohesive view of the generative AI model life cycle and ambiguity in terms of the deployment and perceived effectiveness of varying safety guardrails throughout the life cycle. Amid these challenges, there are significant opportunities, including greater standardization through shared terminology and best practices, facilitating a common understanding of the effectiveness of various risk mitigation strategies.

This briefing paper presents the **Presidio AI Framework**, which provides a structured approach to the safe development, deployment and use of generative AI. In doing so, the framework highlights gaps and opportunities in addressing safety concerns, viewed from the perspective of four primary actors: AI model creators, AI model adapters, AI model users, and AI application users. Shared responsibility, early risk identification and proactive risk management through the implementation of appropriate guardrails are emphasized throughout.

The Presidio AI Framework consists of three core components:

1. **Expanded AI life cycle**: This element of the framework establishes a comprehensive end-to-end view of the generative AI life cycle, signifying varying actors and levels of responsibility at each stage.

2. **Expanded risk guardrails**: The framework details robust guardrails to be considered at different steps of the generative AI life cycle, emphasizing prevention rather than mitigation.

3. **Shift-left methodology:** This methodology proposes the implementation of guardrails at the earliest stage possible in the generative AI life cycle. While shift-left is a well-established concept in software engineering, its application in the context of generative AI presents a unique opportunity to promote more widespread adoption.

In conclusion, the paper emphasizes the need for greater multistakeholder collaboration between industry stakeholders, policy-makers and organizations. The Presidio AI Framework promotes shared responsibility, early risk identification and proactive risk management in generative AI development, using guardrails to ensure ethical and responsible deployment. The paper lays the foundation for ongoing safety-related work of the AI Governance Alliance and the Safe Systems and Technologies working group. Future work will expand on the core concepts and components introduced in this paper, including the provision of a more exhaustive list of known and novel guardrails, along with a checklist to operationalize the framework across the generative AI life cycle.

Introduction

The current AI landscape includes both challenges and opportunities for progress towards safe generative AI models.

This briefing paper outlines the Presidio AI Framework, providing a structured approach to addressing both technical and procedural considerations for safe generative artificial intelligence (AI) models. The framework centres on foundation models and incorporates risk-mitigation strategies throughout the entire life cycle, encompassing creation, adaptation and eventual retirement. Informed by thorough research into the current AI landscape and input from a multistakeholder community and practitioners, the framework underscores the importance of established safety guidelines and recommendations viewed through a technical lens. Notable challenges in the existing landscape impacting the development and deployment of safe generative AI include:

- **Fragmentation:** A holistic perspective, which covers the entire life cycle of generative AI models from their initial design to deployment and the continuous stages of adaptation and use, is currently missing. This can lead to fragmented perceptions of the model's creation and the risks associated with its deployment.

- **Vague definitions:** Ambiguity and lack of common understanding of the meaning of safety, risks[1] (e.g. traceability), and general safety measures (e.g. red teaming) at the frontier of model development.

- **Guardrail ambiguity:** While there is agreement on the importance of risk-mitigation strategies – known as guardrails – clarity is lacking regarding accountability, effectiveness, actionability, applicability, limitations and at what stages of the AI design, development and release life cycle varying guardrails should be implemented.

- **Model access:** An open approach presents significant opportunities for innovation, greater adoption and increased stakeholder population diversity. However, the availability of all the model components (e.g. weights, technical documentation and code) could also amplify risks and reduce guardrails' effectiveness. There is a need for careful analysis of risks and common consensus among the use of guardrails considering the gradient of release;[2] that is, varying levels at which AI models are accessible once released, from fully closed to fully open-sourced.

Simultaneously, there are some identified opportunities for progress towards safety, such as:

- **Standardization:** By linking the technical aspects at each phase of design, development and release with their corresponding risks and mitigations, there is the opportunity for bringing attention to shared terminology and best practices. This may contribute towards greater adoption of necessary safety measures and promote community harmonization across different standards and guidelines.

- **Stakeholder trust and empowerment:** Pursuing clarity and agreement on the expected risk mitigation strategies, where these are most effectively located in the model life cycle and who is accountable for implementation paves the way for stakeholders to implement these proactively. This improves safety, prevents adverse outcomes for individuals and society, and builds trust among all stakeholders.

While this briefing paper details the generative AI model life cycle along with some guardrails, it is by no means exhaustive. Some topics outside this paper's scope include a discussion of current or future government regulations of AI risks and mitigations (this is covered in the Resilient Governance working group briefing paper) or consideration of downstream implementation and use of specific AI applications.

1 Introducing the Presidio AI Framework

A structured approach that emphasizes shared responsibility and proactive risk mitigation by implementing appropriate guardrails early in the generative AI life cycle.

Those releasing, adapting or using foundation models often face challenges in influencing the original model design or setting up the necessary infrastructure for building foundation models. The combined need for regulatory compliance, the significant investments companies are making in AI, and the potential impacts the technology can have on society mean coordination among multiple roles and stakeholders becomes indispensable.

FIGURE 1 **The three elements of the Presidio AI Framework**

The Presidio AI Framework (illustrated in Figure 1) offers a streamlined approach to generative AI development, deployment and use from the perspective of four primary actors: AI model creators, AI model adapters, AI model users and AI application users. This human-centric framework harmonizes the activities of these roles to enable more efficient information transfer between upstream development and downstream applications of foundation models.

AI model creators are responsible for the end-to-end design, development and release of generative AI models. AI model adapters tailor generative AI models to specific generative tasks before integration into AI applications and can provide feedback to the AI model creator. AI model users interact with a generative AI model through an interface provided by the creator. AI application users interact indirectly with the adapted model through an application or application programming interface (API). These actors include secondary groups, for instance, AI model validators and AI model auditors, whose goal is to test and validate against defined metrics, perform safety evaluations or certify the conformity of the AI models pre-release. Validators are internal to AI creator or adapter organizations, while auditors are external entities pursuing model certification.

2 Expanded AI life cycle

The expanded AI life cycle encompasses risks and guardrails with varying safety benefits and challenges throughout each phase.

The expanded AI life cycle synthesizes elements from data management, foundation model design and development, release access, use of generative capabilities and adaptation to a use case. The expanded AI life cycle is introduced in Figure 2.

FIGURE 2 **Presidio AI Framework's expanded AI life cycle**

The **data management phase** describes the data foundations for responsible AI development, including the data access gradient and the catalogue of data source types. The latter aids the AI model creator in navigating various legal implications and challenges, where multiple data source types are typically considered in model creation.

In the **foundation model building phase**, the model moves through various stages from design to internal audit and approval. In contrast, each stage is accompanied by a set of distinct guardrails, detailed in the following section.

The **foundation model release phase** provides responsible model dissemination and risk mitigation, benefiting downstream users and adapters. Foundation models are classified based on how they are released, depending on the level of access granted to downstream actors. This gradient of access spans from fully closed to fully open access; each access type has its own set of norms, standards and release guardrails and has specific benefits and challenges, highlighted in Table 1.

In all phases, unexpected model behaviour could harm users and bring reputational risks or legal consequences to the user and the model creator or adapter. However, the chances of misuse – such as plagiarism, intentional non-disclosure, violation of intellectual property (IP) rights, deepfakes, creation of biologically harmful compounds, generation of toxic content, and misinformation generation – may increase if vigilant oversight processes are not adequately implemented going from fully closed to fully open model access.

TABLE 1 **Safety benefits and challenges of release types**

Release type	Safety benefits	Safety challenges
Fully closed	Creators control the model use and can provide safeguards for data privacy and the IP contained in the model. There is more clarity around responsibility and ownership.	Other actors have limited visibility into the model design and development process. Auditability and contributors' diversity are limited. Application users have minimal influence on model outputs.
Hosted	Creators can provide safeguards for model outputs, such as blocking model response for sensitive queries. They can streamline user support. Use can be tracked and used to improve model responses.	Similar challenges as "fully closed". Other actors have little insight into the model, limiting their ability to understand its decisions.
API	Creators retain control over the model while empowering users to adapt the model for specific use cases. They can provide user support. This level of access increases the "researchability" of the model. Increased access allows users to help identify risks and vulnerabilities.	Even though transparency is limited, model details can be inferred by third-party tools or attacks (in case of bad actors).
Downloadable	Along with creators, adapters and users are also empowered through the release of model components. This means more transparency, flexibility for model use and modification of the model.	Lowered barriers for misuse and potential bypassing of guardrails. Model creators have difficulties in tracking and monitoring model use. Users typically have less support when experiencing unexpected undesirable model outputs/outcomes.
Fully open	These models provide the highest levels of auditability and transparency. This level of access increases global participation and contribution to innovation – also in terms of safety and guardrails. Adapters and users are empowered to adapt models that better align with their specific task and improve existing model functionality and safety via fine tuning.	These models present a higher chance of possible misuse. Access to model weights means higher risk of model replication for unintended purposes by bad actors. Ambiguity around accountability and ownership.

The **model adaptation phase** describes several stages, techniques and guardrails for adapting a pre-trained foundation model to perform specific generative tasks. This phase precedes the **model integration phase**, involving the model's integration with an application, including developing APIs to serve downstream AI application users.

In the **model use phase**, users engage with hosted access models using natural language prompts through an interface provided by the model creator or test it for vulnerabilities. This phase highlights the importance of having necessary guardrails during the foundation model building and release phases as users directly interact with the model. In contrast, adapters can add additional guardrails based on the use case.

3 Guardrails across the expanded AI life cycle

Implementation of known and novel guardrails is necessary for safe systems to ensure technical quality, consistency and control.

Guardrails for safe AI systems refer to guidelines, principles and practices that are put in place to ensure the responsible development, deployment and use of generative AI systems and technologies. They are intended to mitigate risks, prevent harm and ensure AI systems operate according to specific standards and ethical and societal values. Guardrails are implemented from the model-building phase and onward throughout the expanded AI life cycle and may be technical or procedural. Technical guardrails involve tools or automated systems and controls, while procedural guardrails rely on human adherence to established processes and guidelines. A combination of both types is often needed to ensure safe systems. Technical guardrails ensure technical quality and consistency, while procedural guardrails provide process consistency and control.

The section below provides a snapshot of selected guardrails applicable at varying phases of the AI life cycle. Due to brevity, only two of the most widely used guardrails are highlighted, along with their phase placement.

TABLE 2 **Highlighted guardrails and their phase placement**

Highlighted guardrails	Phase placement
Red teaming and reinforcement learning from human feedback (RLHF)[3]	Building
Transparent documentation and use restriction	Release
Model drift monitoring and watermarking	Adaptation

3.1 Model building phase

Performing red teaming early, especially during fine-tuning and validation of the building phase, is crucial for preventing adverse outcomes and ensuring model safety. Addressing vulnerabilities and ethical concerns earlier in the life cycle demonstrates a commitment to security and ethics while building trust among stakeholders. For foundation models, tests should cover prompt injection, leaking, jailbreaking, hallucination, IP and personal information (PI) generation, as well as identifying toxic content. While red teaming is effective for known vulnerabilities, it may have limitations in identifying unknown risks, especially before mass release.

Incorporating reinforcement learning from human feedback (RLHF) early on provides a strategic advantage by enabling efficient learning, faster iterations and a strong foundation for subsequent phases, ultimately leading to improved model performance and alignment with human objectives. RLHF may be used here to train a reward model, which is then used to fine-tune the primary model, eliciting more desirable responses. This process ensures the reliability and alignment of the model outputs and improves performance, including an iterative feedback loop between human raters, a trained reward model and the foundation model. Although effective for ongoing improvement, there is a risk of introducing new biases with this method and data privacy and security considerations around the use of generated data.

Novel approaches to implement these guardrails include "red teaming language models with language models" and reinforcement learning from AI feedback (RLAIF).[4] Both techniques employ language models to generate test cases or provide safety-related feedback on the model. The automation significantly reduces the time needed to implement these guardrails. These may also be applied in later phases, but the advantage of using them earlier allows for adjustments to the model hyperparameters to enhance performance. However, they may come with new vulnerabilities that are not yet fully identified.

3.2 Model release phase

Guardrails implemented in the release phase include a combination of approaches designed to empower downstream actors (such as transparent documentation) and protect them (such as use restrictions).

Transparent documentation is a collection of details (decisions, choices and processes) about the AI model, including the data. It mitigates the risk of lack of transparency,[5] and therefore empowers downstream adapters and users to understand the model's limitations, evaluate its impact and make decisions on model use. This guardrail increases the auditability of the model and helps advance policy initiatives. Some best practices include understanding target consumers, their requirements, and expectations, developing persona-based (e.g. business owner, validator and auditors) templates with pre-defined fields and assigning responsibility for gathering information at every phase of the life cycle. Datasheets, data cards, model cards, factsheets and Stanford's foundation model transparency index indicators are a few examples of building templates. Automating fact collection, building documentation and auditing transparency could improve overall efficiency and effectiveness. Limitations include identifying the most useful facts and ambiguity in balancing the disclosure of proprietary and required information.

Use restriction limits the model use beyond intended purposes. It mitigates the risk of model misuse and other unintended harms like generating harmful content and model adaptation for problematic use cases. Some best practices involve using restrictive licences like responsible AI licences (RAIL), setting up model use and user tracking, and providing clear guidelines on allowed use while implementing feedback/incident reporting mechanisms. Additionally, integrating moderation tools to filter or flag undesirable content, disallowing harmful or sensitive prompts and blocking the model from responding to misaligned prompts must be considered. Limitations include having standards for model licences and guidelines and high-quality tools to help restrict the model response.

3.3 Model adaptation phase

A critical goal of the adaptation phase is to ensure that the adapted model remains effective and aligned with the selected use case. Model drift monitoring involves regularly comparing post-deployment metrics to maintain performance in the face of evolving data, adversarial inputs, noise and external factors. The goal is to mitigate the risk of model drift, where the model's output deviates from expectations over time. Best practices include systematically using data, algorithms, and tools for tracking data drift, and defining response protocols and adaptation techniques to sustain model performance and customer trust.

The decision to watermark model outputs depends on the use case, model nature and watermarking goals. Watermarking adds hidden patterns for algorithmic detection, mitigating mass production of misleading content. It aids in identifying AI-generated content for policy enforcement, attribution, legal recourse and deterrence. However, workarounds exist, such as removing watermarks or paraphrasing content. Watermarking can be applied earlier (during model creation for ownership) and adaptation for control over visibility.

4 Shifting left for optimized risk mitigation

The "shift-left" approach involves implementing safety guardrails earlier in the life cycle to mitigate risks and increase efficiency.

The term "shift-left"[6] describes implementing quality assurance and testing measures earlier in a product cycle. The core objective is proactively identifying and managing potential risks, increasing efficiency and cost-effectiveness. This well-established concept applies to various technologies and processes, including software engineering.

In the Presidio AI Framework, the concept of shift-left is extended and applied to generative AI models. It gains a new dimension of importance due to:

- Increased interest in foundation models where model creators are not always the model adapters.
- Increased accessibility of powerful models by users of varying skills and technical backgrounds, raising the demand for model transparency.
- Considerable risk for users using factually incorrect output without validation, model misuse (e.g. in disinformation campaigns) and adversarial attacks on the model (e.g. jailbreaking).

These considerations require understanding and coordination of the activities of different actors (creators, adapters and users) across the AI value chain to avoid significant effort in resolving issues during model adoption and use. For example, data subject rights in some countries allow people to request that their personal information be deleted from the model. The removal can be costly for model creators as they may need to retrain the model. It can also be challenging for adaptors to apply effective guardrails to prevent sensitive information from surfacing in the output.

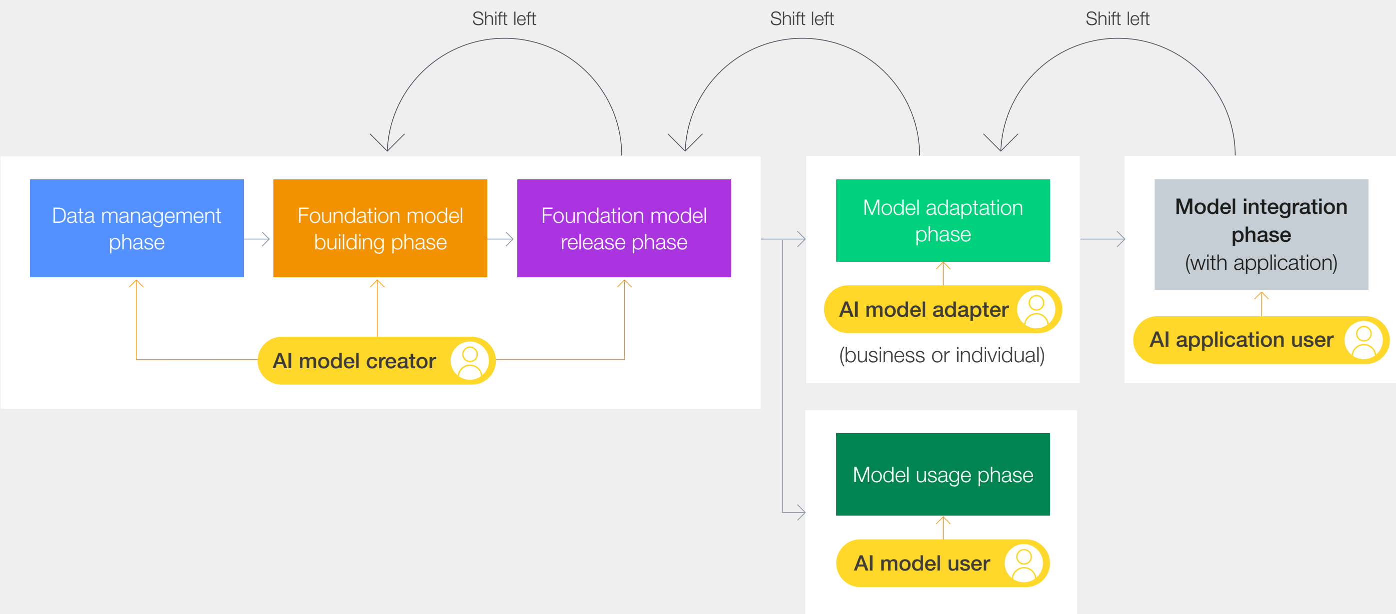
For generative AI, the shift-left methodology proposes guardrails earlier in the life cycle, considering their effectiveness in mitigating risk at a particular phase, along with essential foundation model safety features, the need for balancing safety with model creativity and implementation cost. Based on the model's purpose, there could be a trade-off between guardrail placement and safety dimensions like privacy, fairness, accuracy and transparency.

Figure 3 illustrates three shift-left instances crucial for building safe generative AI models.

- **Release to build shift** occurs when an AI model creator proactively incorporates guardrails throughout the foundation model-building phase and collects necessary data and model facts and transparency surrounding these.
- **Adaptation/use to release shift** occurs during the foundation model release phase. The AI model creator incorporates additional guardrails, establishes norms and standards for use, and creates comprehensive documentation to help downstream actors understand and make informed decisions regarding model use.
- **Application to adaptation shift** occurs when the AI model adapter proactively incorporates guardrails considering the use case and considering the documentation from AI model creators about the foundation model. These would be documented for the downstream application user.

Some organizations have already integrated the shift-left approach into their responsible AI development process. However, it is vital to extend and emphasize the importance of this practice across all expanded phases of the generative AI life cycle and ensure its adoption by all organizations. Those that shift left to implement appropriate safety guardrails where most effective can minimize legal consequences and reputational risk, increase trusted adoption and positively impact society and users.

FIGURE 3 | **Presidio AI Framework with shift-left methodology for generative AI models**



Conclusion

The Presidio AI Framework promotes shared responsibility, early risk identification and proactive risk management in generative AI development, using guardrails to ensure ethical and responsible deployment. The AI Governance Alliance and the Safe Systems and Technologies working group encourage greater information exchange between industry stakeholders, policy-makers and organizations. This collaborative effort aims to increase trust in AI systems, ultimately benefiting society.

In addition to known guardrails, the group will continue to identify novel mechanisms for AI safety, including emerging technical guardrails such as red teaming language models,[7] liquid neural networks (LNN),[8] BarrierNets,[9] causal foundation models[10] and neurosymbolic learning,[11] among others. Additionally, the group will investigate the various guardrail options and introduce a checklist to operationalize the framework to assess AI model risks and guardrails across the generative AI life cycle.

Contributors

This paper is a combined effort based on numerous interviews, discussions, workshops and research. The opinions expressed herein do not necessarily reflect the views of the individuals or organizations involved in the project or listed below. Sincere thanks are extended to those who contributed their insights via interviews and workshops, as well as those not captured below.

World Economic Forum

Benjamin Larsen
Lead, Artificial Intelligence and Machine Learning

Cathy Li
Head, AI, Data and Metaverse; Deputy Head, Centre for the Fourth Industrial Revolution; Member of the Executive Committee

Supheakmungkol Sarin
Head, Data and Artificial Intelligence Ecosystems

AI Governance Alliance Project Fellows

Ravi Kiran Singh Chevvan
AI Strategy & Complex Program Executive, IBM

Jerry Cuomo
Executive Fellow and Vice-President, Technology, IBM

Steven Eliuk
Executive Fellow and Vice-President, AI & Governance, IBM

Jennifer Kirkwood
Executive Fellow, Partner, IBM

Eniko Rozsa
Distinguished Engineer, IBM

Saishruthi Swaminathan
Tech Ethics Program Adviser, IBM

Joseph Washington
Senior Technical Staff Member, IBM

Acknowledgements

Sincere appreciation is extended to the following working group members, who spent numerous hours providing critical input and feedback to the drafts. Their diverse insights are fundamental to the success of this work.

Uthman Ali
Senior Product Analyst, AI Ethics SME, BP

Animashree (Anima) Anandkumar
Bren Professor of Computing and Mathematical Sciences, California Institute of Technology (Caltech)

Amir Banifatemi
Co-Founder and Director, AI Commons

Michael Benton
Director, Responsible AI Practice, Microsoft

Stella Biderman
Executive Director, EleutherAI

Shane Cahill
Director, Privacy and AI Legislation and Policy Development, Meta Platforms

Suha Can
Chief Information Security Officer, Grammarly

Jennifer Chayes
Dean of the College of Computing, Data Science, and Society, University of California, Berkeley

Kevin Chung
Chief Operating Officer, Writer

Jeff Clune
Associate Professor, Department of Computer Science, Faculty of Science, Vector Institute

Cathy R Cobey
Global Responsible Co-Lead, EY

Umeshwar Dayal
Corporate Chief Scientist, Hitachi

Mona Diab
Director of Language Technologies Institute, Carnegie Mellon University

Mennatallah El-Assady
Professor, ETH Zurich

Gilles Fayad
Adviser, Institute of Electrical and Electronics Engineers (IEEE)

Jocelyn Goldfein
Managing Director, Zetta Venture Partners

Tom Gruber
Founder, Humanistic AI

Lan Guan
Global Data and AI Lead, Senior Managing Director, Accenture

Gillian Hadfield
Professor of Law and Professor of Strategic Management, University of Toronto

Peter Hallinan
Leader, Responsible AI, Amazon Web Services

Or Hiltch
Chief Data and AI Architect, JLL

Babak Hodjat
Chief Technology Officer AI, Cognizant Technology Solutions US

Sara Hooker
Head, Research, Cohere

David Kanter
Founder and Executive Director, MLCommons

Vijay Karunamurthy
Head of Engineering and Vice-President, Engineering, Scale AI

Sean Kask
Chief AI Strategy Officer, SAP

Robert Katz
Vice-President, Responsible AI & Tech, Salesforce

Michael Kearns
Founding Director, Warren Center for Network and Data Sciences, University of Pennsylvania

Steve Kelly
Chief Trust Officer, Institute for Security and Technology

Jin Ku
Chief Technology Officer, Sendbird

Sophie Lebrecht
Chief, Operations and Strategy, Allen Institute for Artificial Intelligence

Aiden Lee
Co-Founder and Chief Technology Officer, Twelve Labs

Stefan Leichenauer
Vice-President, Engineering, SandboxAQ

Tze Yun Leong
Professor of Computer Science; Director, NUS Artificial Intelligence Laboratory

Scott Likens
Global AI and Innovation Technology Lead, PwC

Shane Luke
Vice-President, Product and Engineering, Workday

Richard Mallah
Principal AI Safety Strategist, Future of Life Institute

Pilar Manchón
Senior Director, Engineering, Google

Risto Miikkulainen
Professor of Computer Science, University of Texas at Austin

Lama Nachman
Intel Fellow, Director of Human & AI Systems Research Lab, Intel

Syam Nair
Chief Technology Officer, Zscaler

Mark Nitzberg
Executive Director, UC Berkeley Center for Human-Compatible AI,

Vijoy Pandey
Senior Vice-President, Outshift by Cisco, Cisco Systems

Louis Poirier
Vice-President AI/ML, C3 AI

Victor Riparbelli
Co-Founder and Chief Executive Officer, Synthesia

Jason Ruger
Chief Information Security Officer, Lenovo

Daniela Rus
Director, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT)

Noam Schwartz
Chief Executive Officer and Co-Founder, Activefence

Jun Seita
Team Leader (Principal Investigator), Medical Data Deep Learning Team, RIKEN

Susannah Shattuck
Head, Product, Credo AI

Paul Shaw
Group Security Officer, Dentsu Group

Evan Sparks
Chief Product Officer, AI, Hewlett Packard Enterprise

Catherine Stihler
Chief Executive Officer, Creative Commons

Fabian Theis
Science Director, Helmholtz Association

Li Tieyan
Chief AI Security Scientist, Huawei Technologies

Kush Varshney
Distinguished Research Scientist and Senior Manager, IBM

Lauren Woodman
Chief Executive Officer, DataKind

Yuan Xiaohui
Senior Expert, Tencent Holdings

Grace Yee
Director, Ethical Innovation, AI Ethics, Adobe

Michael Young
Vice-President, Products, Private AI

Leonid Zhukov
Vice-President, Data Science, BCGX; Director of BCG Global AI Institute, Boston Consulting Group

World Economic Forum

John Bradley
Lead, Metaverse Initiative

Karyn Gorman
Communications Lead, Metaverse Initiative

Devendra Jain
Lead, Artificial Intelligence, Quantum Technologies

Jenny Joung
Specialist, Artificial Intelligence and Machine Learning

Daegan Kingery
Early Careers Programme, AI Governance Alliance

Connie Kuang
Lead, Generative AI and Metaverse Value Creation

Hannah Rosenfeld
Specialist, Artificial Intelligence and Machine Learning

Stephanie Teeuwen
Specialist, Data and AI

Karla Yee Amezaga
Lead, Data Policy and AI

Hesham Zafar
Lead, Digital Trust

IBM

Jesús Mantas
Global Managing Director

Christina Montgomery
Chief Privacy & Trust Officer

Production

Laurence Denmark
Creative Director, Studio Miko

Sophie Ebbage
Designer, Studio Miko

Martha Howlett
Editor, Studio Miko

Endnotes

1. IBM AI Ethics Board, *Foundation models: Opportunities, risks and mitigations*, 2023, https://www.ibm.com/downloads/cas/E5KE5KRZ.
2. Solaiman, Irene, "The Gradient of Generative AI Release: Methods and Considerations", *Hugging Face*, 2023, https://arxiv.org/abs/2302.04844.
3. Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic et al., "Deep Reinforcement Learning from Human Preferences", *arxiv*, 17 February 2023, https://arxiv.org/pdf/1706.03741.pdf.
4. Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard et al., "RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback", *Google Research*, 1 December 2023, https://arxiv.org/pdf/2309.00267.pdf.
5. Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayah Kapoor et al, "The Foundation Model Transparency Index", *Stanford Center for Research on Foundation Models and Stanford Institute for Human-Centered Artificial Intelligence*, 2023, https://arxiv.org/pdf/2310.12941.pdf.
6. Smith, Larry, "Shift-left testing", *Association for Computing Machinery Digital Library*, 2001, https://dl.acm.org/doi/10.5555/500399.500404.
7. Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai et al., "Red Teaming Language Models with Language Models", *Association for Computational Linguistics*, 2022, https://aclanthology.org/2022.emnlp-main.225.pdf.
8. Hasani, Ramin, Mathias Lechner, Alexander Amini, Daniela Rus et al., "Liquid Time-constant Networks", *arxiv*, 2020, https://arxiv.org/pdf/2006.04439.pdf.
9. Xiao, Wei, Ramin Hasani, Xiao Li and Daniela Rus, "BarrierNet: A Safety-Guaranteed Layer for Neural Networks", *Massachusetts Institute of Technology*, 2021, https://arxiv.org/pdf/2111.11277.pdf.
10. Willig, Moritz, Matej Zecevic, Devendra Singh Dhami and Kristian Kerting, "Can Foundation Models Talk Causality?", *arxiv*, 2022, https://arxiv.org/pdf/2206.10591.pdf.
11. Roy, Kaushik, Yuxin Zi, Vignesh Narayanan, Manas Gaur and Amit Seth, "Knowledge-Infused Self Attention Transformers", *arxiv*, 2023, https://arxiv.org/pdf/2306.13501.pdf.

COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744
contact@weforum.org
www.weforum.org