In collaboration
with AI Commons

WORLD
ECONOMIC
FORUM

The Presidio Recommendations on Responsible Generative AI

JUNE 2023

Introduction

Generative artificial intelligence (AI) has the potential to transform industries and society by boosting innovation and empowering individuals across diverse fields, from arts to scientific research. To ensure a positive future, it is crucial to prioritize responsible design and release practices from the beginning. As generative AI continues to advance at an unprecedented pace, the need for collaboration among stakeholders to ensure that AI serves as a force for good has become increasingly urgent.

On 26-28 April 2023, the summit “Responsible AI Leadership: A Global Summit on Generative AI” took place at the World Economic Forum’s Centre for the Fourth Industrial Revolution based in the Presidio in San Francisco, USA. The event was hosted by the Forum in partnership with AI Commons to guide technical experts and policy-makers on the responsible development and governance of generative AI systems.

The summit emphasized the importance of open innovation and international collaboration as essential enablers for responsible generative AI. The focus was on moving beyond insightful discussions to generate actionable and practical recommendations for various AI stakeholders that could significantly influence the design, construction and deployment of generative AI.

Over 100 AI thought leaders and practitioners participated in the summit, including chief scientific officers, responsible AI and ethics leads, academic leaders, AI entrepreneurs, policy-makers, tech investors and members of civil society. Participants engaged in discussions on numerous aspects of generative AI’s design, development, release and societal impact, and deliberated on key recommendations. These recommendations emerged from interactive panel discussions and working sessions through a bottom-up process, with participants reaching consensus on critical areas related to the governance of generative AI.

This summary presents a set of 30 action-oriented recommendations aimed at guiding generative AI towards meaningful human progress. The recommendations address three key themes that cover the entire life cycle of generative AI: responsible development and release; open innovation and international collaboration; and social progress.

By implementing these recommendations, stakeholders can navigate the complexities of AI development and harness its potential responsibly and ethically. Join us in shaping a more innovative, equitable and prosperous future that leverages the power of generative AI and mitigate its risks to benefit all.

Responsible Development and Release of Generative AI

This section critically assesses the necessity to protect our society from unforeseen outcomes induced by the swiftly developing generative AI systems, and accordingly advocates for responsible strategies concerning their development and deployment. These recommendations are intended for a broad spectrum of stakeholders - ranging from AI developers to policy-makers and users. The objective is to foster accountable and inclusive processes for AI development and deployment, thereby enhancing trust and transparency as generative AI systems continue to proliferate.

01 Establish precise and shared terminology

All stakeholders are called upon to use precise terminology when discussing the design, development, evaluation and measurement of generative AI models' capabilities, limitations and issues. It is the responsibility of experts to define and standardize this language. As soon as a consensus is reached, consistent adoption of this terminology by all stakeholders is essential. This approach will boost clarity and promote effective communication, leading to a shared understanding among different parties. Ultimately, it will facilitate the establishment of strong, standards, guidelines and regulations for a range of generative AI applications.

02 Build public awareness of AI capabilities and their limitations

Public and private stakeholders should prioritize the task of enhancing public understanding. This includes making the terminology related to generative AI models understandable to the general public. Additionally, stakeholders should inform users about the probabilistic (meaning their outputs are not deterministic but based on probability) and stochastic (implying their operation involves a degree of random behavior) nature of generative AI models, while setting accurate expectations for their performance.

03 Focus on human values and preferences

The challenge to align generative AI models with human values and preferences needs to be further acknowledged and addressed. Developers of AI systems should be engaged in discussions about normative values and preferences when designing AI models.

04 Encourage alignment and participation

Public and private sector stakeholders should recognize that AI systems necessitate quality feedback that is diverse and representative of the user base to be truly aligned. Policy-makers should promote the involvement of diverse stakeholders, including non-technical stakeholders, in AI research and development to ensure alignment with human values. AI developers should work to facilitate interactions and feedback from a broad range of participants to create a more inclusive and human-centric development process.

05 Uphold AI accountability with rigorous benchmarknig and use case-specific testing while exploring new metrics and standards

AI developers should commit to the importance of not only holding models accountable against the highest established benchmarks, but also finding new metrics beyond traditional ones and towards other human-centric dimensions. Benchmarking should be complemented by application-specific and task-defined testing to ensure a comprehensive evaluation of generative AI models.

06 Employ diverse red teams

Red teaming, a method of critically analysing perspective to identify potential weaknesses, vulnerabilities and areas for improvement, should be integral from model design to application and release. Diversity here implies incorporating members from varied genders, backgrounds, experiences and perspectives for a more comprehensive critique. The public and private sectors should implement frameworks and methodologies to facilitate thorough red teaming.

07 Adopt transparent release strategies

Producers of AI should be held accountable to release AI models responsibly, making them available to the public without compromising safety. Responsible release strategies should be initiated upstream during project ideation and product design to ensure that potential risks are identified and mitigated throughout the development process.

08 Enable user feedback

Users should be empowered with robust controls that allow them to provide real-time feedback on model outputs. Additionally, it is relevant to enable users to have a comprehensive understanding of the limits and responsibilities associated with the generated content.

09 Embed model and system traceability

Developers and policy-makers should align on the importance of creating formal evaluation and auditing structures surrounding traceability throughout the entire AI life cycle, from data provenance to training scenarios and post-implementation.

10 Ensure content traceability

To increase transparency and accountability, companies developing AI-generated content should be responsible for tracing how content is generated and documenting its provenance. This will help users discern the difference between human-generated and AI-generated content.

11 Disclose non-human interaction

In virtual environments, humans should know whether they are interacting with a human or a machine. AI providers should develop mechanisms to support this, for example, via watermarking.

12 Build human-AI trust

To build trust in AI systems, developers and companies should prioritize transparency, consistency, and meeting and managing user expectations. AI developers should be transparent in their processes and decision-making, providing users with an understanding of how they reach their results. By focusing on these aspects, AI developers can create systems that foster trust and facilitate positive human-AI interactions.

13 Implement a step-by-step review process

Policy-makers and businesses should create a step-by-step review process for AI models and products. This should be similar to the detailed checks used in clinical trials or car manufacturing, both before and after a product goes live. There should be an independent auditor or international agency to oversee this to ensure uniform evaluations and continuous monitoring. To help limit potential risks and negative impacts, certification, or licensing system could be used.

14 Develop comprehensive, multi-level measurement frameworks

Policy-makers should emphasize ongoing efforts and incentivize developers and standardization bodies to focus on creating and employing measurement frameworks with an emphasis on socio-technical aspects rather than solely technical performance.

15 Adopt sandbox processes

AI developers, standard-setting bodies and regulators should cooperate on more flexible “sandbox” development environments along with new and associated processes of governance and oversight. Sandboxing could help build trust by demonstrating that AI systems have undergone rigorous testing and evaluation to ensure safety, reliability and compliance.

16 Adapt to the evolving landscape of creativity and intellectual property

With generative AI impacting content creation, it is essential for policy-makers and legislators to re-examine and update copyright laws to enable appropriate attribution, and ethical and legal reuse of existing content.

Open Innovation and International Collaboration

This section focuses on the importance of sharing scientific knowledge and enhancing international collaboration. As frontier research capabilities tend to be concentrated in private sector companies in a select few countries, it is vital that academic researchers remain an integral part of the exploratory process, while countries worldwide participate and influence the governance of generative AI systems. These recommendations are designed for a range of stakeholders, including researchers, AI developers, standard-setting bodies and policy-makers. The overarching goal is to cultivate transparency, accountability and inclusivity in the development, implementation and governance of generative AI.

17 Incentivize public-private research coordination

Public and private stakeholders should actively work to design incentive structures that facilitate greater coordination between academic researchers and the private sector throughout the technology development lifecycle. Possible mechanisms to be considered include joint research programmes, data-sharing protocols and joint IP ownership.

18 Build a common registry of models, tools, benchmarks and best practices

Producers and researchers of generative AI should contribute to a common and open registry of source codes, models, datasets, tools, benchmarks and best practice guidelines, to be shared within the research community, in order to have a platform for academic and private sector collaboration to build future models and systems that are transparent and accountable to the public.

19 Support responsible open innovation and knowledge sharing

Policy-makers and AI providers should contribute to frameworks to democratize AI through responsible sharing of resources, including data, source code, models and research findings; also encourage the sharing certification processes, ensuring transparency and trust among stakeholders. A public-private long-term initiative could be developed to build public-facing platforms that provide open access to compute, data and pre-trained models. This platform could be treated as a digital public good, and usage could be promoted across borders.

20 Enhance international collaboration on AI standards

Standard bodies must foster international collaboration on AI standards, ensuring the participation of all AI stakeholders, including all geographical locations.

21 Establish a global AI governance initiative

To address the challenges and potential risks posed by AI technologies, policy-makers should consider devoting efforts towards creating a global AI governance initiative. This initiative should bring together experts from a wide array of fields. The key focus should be on promoting global understanding of responsible generative AI, ensuring broad inclusion, facilitating access to infrastructure, and fostering collaboration to harmonize response structures at the national level against AI challenges and risks.

Social progress

This section examines the hurdles tied to AI-driven transformations, spanning from workforce transitions to educational shifts, as well as the necessity of championing AI for societal benefit and advocating for equitable AI access in developing nations. The recommendations are intended for a broad array of stakeholders, including educational institutions, community organizations, corporations, individuals, policy-makers and governments. The primary objective is to cultivate a society that is more informed, engaged and resilient in the face of these emerging changes.

22 Prioritize social progress in generative AI development and adoption

All stakeholders must ensure that the technology's societal implications remain front and centre. This involves a focus beyond technical proficiency towards the technology's role in enhancing social progress. Comprehensive support must be provided to communities and workers affected by the shift to an AI-enabled society, encompassing learning initiatives, guidance on surmounting generative AI-specific challenges and assistance in navigating the ethical, social and technical shifts inherent in an AI-influenced environment with an active participation of workers throughout the process.

23 Drive AI literacy across society

Educational bodies and community institutions must take the initiative to increase AI literacy among the general public. A proactive approach is needed to demystify generative AI tools, outline their potential uses and discuss their ethical implications. This will empower individuals to better understand, interact with and contribute to the evolving landscape of AI, fostering a more informed and participative society.

24 Foster holistic thought approaches in AI-driven environments

Foster diverse modes of thinking – critical, computational and responsible – to better equip society for the generative AI era. Encourage these core competencies across sectors and communities to empower individuals to engage critically with AI-generated content, understand the underlying technology and make responsible decisions about its use.

25 Steer generative AI’s transformative impact

Address the transformative influence of generative AI on societal systems. Understand its effect on human interactions, knowledge dissemination and evaluation mechanisms. Proactively adapt to the evolving landscape, supporting roles that may transform due to generative AI, and explore innovative ways to evaluate its impacts within our rapidly evolving digital ecosystem, to harness its potential for driving positive societal transformation.

26 Incentivize innovation for social good

Policy-makers should encourage the development and implementation of generative AI technologies that prioritize social good and address complex and unmet societal needs, such as in healthcare and climate change, to improve the overall quality of life.

27 Address resource and infrastructure disparities

Policy-makers should increase public investment in national and international research infrastructure. That includes work to ensure greater access to computing resources for researchers, especially those from underrepresented regions and institutions. The private sector is encouraged to contribute to the development of datasets and support governments in making more resources available to researchers.

28 Promote generative AI expertise within governments

Governments should invest in fostering AI expertise, ensuring an informed, effective and responsible approach to public policies and regulation of these transformative technologies. By leveraging mechanisms such as targeted incentives, private sector collaborations, and exchange programs, governments can nurture AI talent. This commitment while expanding in-house AI proficiency is crucial in securing a future where these technologies advance societal progress and serve the public interest effectively.

29 Increase equitable access to AI in developing countries

To ensure that the benefits of generative AI technology are accessible to all, public and private stakeholders should focus on establishing initiatives that can provide support and resources at scale, particularly in developing countries where there may be limited access to digital infrastructures. Efforts should focus on providing resources, training, and expertise to make AI more accessible and inclusive, fostering national and international partnerships across sectors to promote diversity and inclusion in the development and deployment of generative AI technology.

30 Preserve cultural heritage

All stakeholders need to contribute to preserve cultural heritage. Public and private sector should invest in creating curated datasets and developing language models for underrepresented languages, leveraging the expertise of local communities and researchers and making them available. This will improve access to AI technologies to help preserve linguistic diversity and cultural heritage.

Authors

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

Benjamin Larsen
Lead, Artificial Intelligence and Machine Learning, World Economic Forum

Hubert Halopé
Lead, Artificial Intelligence and Machine Learning, World Economic Forum

Lucia Velasco
Lead, Artificial Intelligence and Machine Learning, World Economic Forum

Summit Co-Chairs

Amir Banifatemi
Director, AI Commons

Pascale Fung
Chair Professor, Hong Kong University of Science & Technology

Francesca Rossi
IBM Fellow and IBM AI Ethics Global Leader; AAAI President

Joaquin Quiñonero-Candela
Technical Fellow for Artificial Intelligence, LinkedIn

Summit Steering Committee

Esteban Arcaute
Head of Responsible AI, Meta Platforms

Yoshua Bengio
Head of the Montreal Institute for Learning Algorithms, University of Montreal

Mona Diab
Lead Responsible AI Research Scientist, Meta Platforms

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

Hiroaki Kitano
Senior Executive Vice-President and Chief Technology Officer; Chief Executive Officer, Sony Research, Sony Group Corporation

Yann LeCun
Vice-President and Chief AI Scientist, Meta Platforms

Pilar Manchón
Senior Director of Engineering, Google

Peter Norvig
Director of Research, Google

Contributors

Blaise Aguera
Vice-President and Fellow, Google Research, Google

Xavier Amatriain
Vice-President of Engineering – Product AI Strategy, LinkedIn Corporation

Stephen Augustus
Head of Open Source, Cisco Systems

Ricardo Baeza-Yates
Director of Research, Institute for Experiential AI, Northeastern University

Anthony Bak
Head of AI and Machine Learning, Palantir Technologies

Houman Behzadi
President and Chief Product Officer, C3 AI

Kimmy Bettinger
Expert and Knowledge Communities Lead, World Economic Forum

Seth Bergeson
Manager, AI and Emerging Technology, PwC

Jamie Berryhill
Artificial Intelligence Policy Analyst, Organisation for Economic Co-operation and Development (OECD)

Marc Boxser
Vice-President, Policy and Communications, Chegg Inc.

Kirk Bresniker
Fellow and Chief Architect, Hewlett Packard Labs, Hewlett Packard Enterprise

Joanna Bryson
Professor of Ethics and Technology, Hertie School

Sebastian Buckup
Head of Network and Partnerships, Deputy Head. Centre for the Fourth Industrial Revolution; Member of the Executive Committee, World Economic Forum

Jill Burstein
Principal Assessment Scientist, Duolingo

Cansu Canca
AI Ethics Lead at Institute for Experiential AI, Northeastern University

Diane Chang
Director of Data Science, Intuit Inc.

Joshua Cohen
Member of the Faculty, Apple University, Apple

David Cox
Director of Exploratory AI Research, IBM Corporation

Natasha Crampton
Chief Responsible AI Officer, Microsoft Corporation

Joris Cyizere
Head ad interim, Centre for the Fourth Industrial Revolution, Rwanda

Umeshwar Dayal
Corporate Chief Scientist, Senior Vice-President and Senior Fellow, Hitachi America

Anil Dewan
Senior Advisor, US Department of Homeland Security

Daniel Dobrygowski
Head, Governance and Trust, World Economic Forum

Anne Marie Engtoft Larsen
Tech Ambassador, Ministry of Foreign Affairs of Denmark

Mojdeh Eskandari
Founder and President, Enovant Foundation

Aldo Faisal
Professor of Artificial Intelligence and Neuroscience, Imperial College London

Gilles Fayad
Advisor, Institute of Electrical and Electronics Engineers

Rebecca Finlay
Chief Executive Officer, Partnership on AI

Kay Firth-Butterfield
Executive Director, Centre for Trustworthy Technology

Gwenda Fong
Deputy Secretary (Development and Regulation, Ministry of Communications and Information of Singapore

Edward Fu
Head of Government Affairs, Duolingo

Krishna Gade
Chief Executive Officer and Co-Founder, Fiddler Labs

Eugenio Garcia
Deputy Consul-General, San Francisco, Ministry of Foreign Affairs of Brazil

Tiffany Georgievski
AI Attorney, Sony Research

Matthew Graviss
Chief Data Officer, US Department of State

Tom Gruber
Co-Founder/Chief Technology Officer, Siri and Humanistic.ai

Peter Hallinan
Leader, Responsible AI, Amazon Web Services

Ruimin He
Chief Technology Advisor, Ministry of Communications and Information of Singapore

Brittan Heller
Fellow, Digital Forensics Research Lab, The Atlantic Council

Cyrus Hodes
Co-Founder of AIGC Chain and Stability AI, Harvard Kennedy School of Government

Babak Hodjat
Chief Technology Officer AI, Cognizant Technology Solutions US Corp.

Jerremy Holland
Director of AI Research, Apple

Matissa Hollister
Assistant Professor of Organizational Behaviour, McGill University

Sara Hooker
Head, Cohere for AI

Eric Horvitz
Chief Scientific Officer, Microsoft

Xinghai Hu
Head of TikTok Data US, Bytedance

Anil Kamath
Fellow and Vice-President AI/ML, Adobe Systems

Vijay Karunamurthy
Vice-President of Engineering, Scale AI

Anja Kaspersen
Member, Council on Extended Intelligence and Industry Activity on Life Science, Institute of Electrical and Electronics Engineers

Jeffrey Ladish
Head of AI Insights, Center for Humane Technology

Yolanda Lannquist
Director of AI Governance, The Future Society

Federico Lecumberry
Associated Professor, Universidad de la República

Chase Lochmiller
Co-Founder and Chief Executive Officer,
Crusoe Energy Systems

Leland Lockhart
Director, Artificial Intelligence & Machine Learning, Vista Equity Partners

David Luan
Chief Executive Officer, Adept AI

Emily McReynolds
Senior Fellow, Center for Responsible AI,
New York University

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

Steven Mills
Partner and Chief Artificial Intelligence Ethics Officer,
Boston Consulting Group

Joshua New
Technology Policy Executive, IBM Corporation

Loren Newman
Government Affairs Lead, World Economic Forum

Vaibhav Pahwa
Product Manager, Platform Fairness
and Responsible AI, TikTok

Gleb Papyshev
PhD Candidate in Science and Technology Policy, The Hong Kong University of Science and Technology

Vijay Parthasarathy
Head of Artificial Intelligence and Machine Learning,
Zoom Video Communications

Jonnie Penn
Assistant Teaching Professor of AI Ethics
and Society, University of Cambridge

Nazneen Rajani
Research Lead, Hugging Face

Martin Rauchbauer
Co-Director and Founder, Tech Diplomacy Network

Stuart Russell
Professor of Computer Science,
University of California, Berkeley

Sultan Saidov
Co-Founder and President, Beamery Inc.

Nayat Sanchez-Pi
Chief Executive Officer, INRIA Chile

Supheakmungkol Sarin
Head of Data and Artificial Intelligence Ecosystems,
World Economic Forum

Silvio Savarese
Executive Vice-President, Chief Scientist, Salesforce

Yoav Schlesinger
Architect, Ethical AI Practice, Salesforce

Craig Shank
Advisor, Responsible Artificial Intelligence Institute

Joanna Shields
Chief Executive Officer, BenevolentAI

Karen Silverman
Founder and Chief Executive Officer,
The Cantellus Group

Sarvjeet Singh
Principal Engineer/Engineering Director,
Google Research, Google

Navrina Singh
Founder and Chief Executive Officer, Credo AI

Uyi Stewart
Chief Data and Technology Officer, data.org

JoAnn Stonier
Chief Data Officer, Mastercard

Murali Subbarao
Vice-President, AI Solution Success, ServiceNow

Candace Sue
Vice-President for Academic Relations, Chegg

Arun Sundararajan
Harold Price Professor of Entrepreneurship and Technology, Stern School of Business, New York University

Josephine Teo
Minister of Communications and Information
of Singapore

Kellee Tsai
Dean of Humanities and Social Science,The Hong Kong University of Science and Technology

Thomas Wolf
Co-Founder, Hugging Face

Andrea Wong
Head of Platform Fairness, Bytedance

Lauren Woodman
Chief Executive Officer, DataKind

Daniel Wroblewski
Managing Director, Head of Investment Science,
Canada Pension Plan Investment Board

Alice Xiang
Global Head of AI Ethics, Sony Research,
Sony Group Corporation

Kevin Yancey
Staff AI Research Engineer, Duolingo

Masaru Yarime
Associate Professor, The Hong Kong University
of Science and Technology

Grace Yee
Director of Ethical Innovation, AI Ethics, Adobe

Polina Zvyagina
Privacy and Data Policy Manager, AI/ML Products,
Responsible AI, Meta Platforms

WORLD
ECONOMIC
FORUM

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