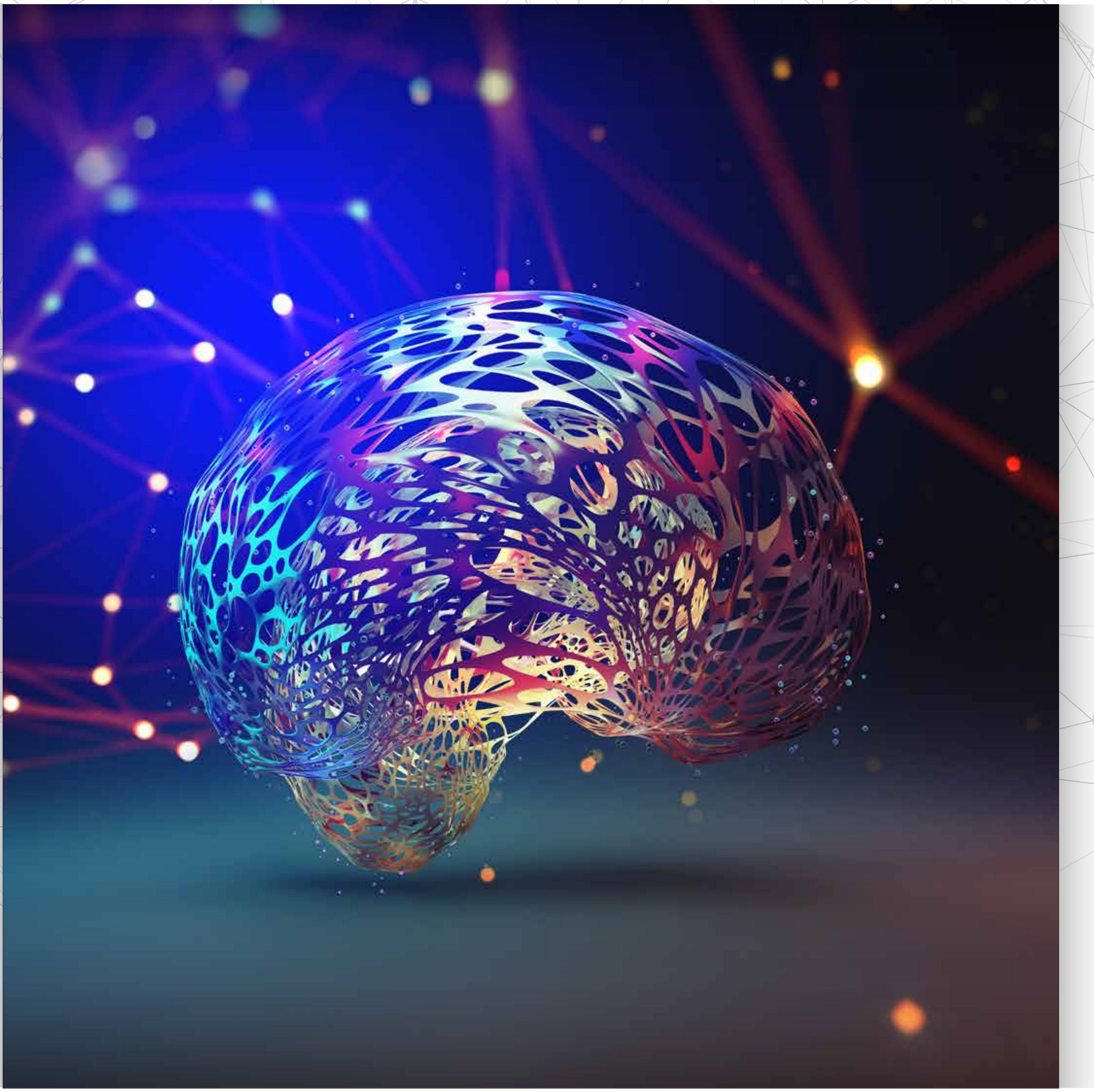
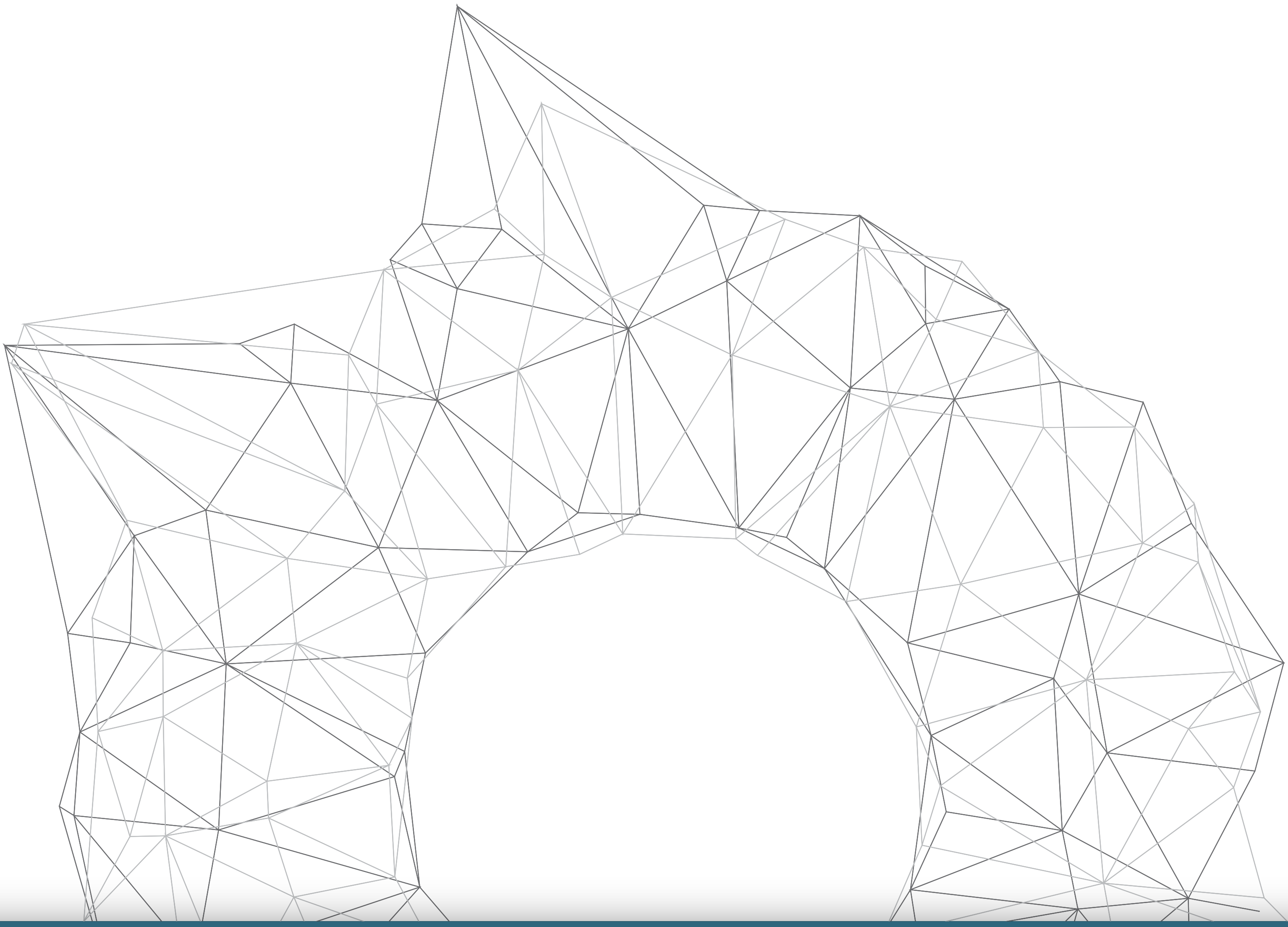


GenAI-powered evaluation function at UNFPA

Strategy for leveraging the benefits of responsible and ethical generative artificial intelligence while minimizing risks

April 2024

UNFPA
IEO
Independent Evaluation Office



Disclaimer: The foundational GenAI models Bard, ChatGPT, and Claude were used while designing, drafting, and revising this strategy document to develop an outline, generate ideas for content, and refine sub-section writing. These AI tools were intentionally used to test and demonstrate their feasibility in supporting the development of this document. The AI tools utilized adhere to UNFPA's AI usage clause upholding ethical and responsible use, transparency, validation of results and compliance with relevant internal regulations. Specifically, no internal information was used on these platforms, and ChatGPT and Claude were used for innovation purposes only, following the March 2023 ITSO guidance.

This is a publication by the Independent Evaluation Office of UNFPA.

unfpa.org/evaluation **evaluation.office@unfpa.org** **@unfpa_eval** **@UNFPA_EvaluationOffice**

Foreword

The advent of generative artificial intelligence (GenAI) has ushered in a new era of innovation and transformation, offering immense potential to accelerate the Sustainable Development Goals (SDGs). GenAI's ability to analyse vast amounts of data, lowering barriers for obtaining automated responses, and generate creative solutions may enable new approaches to addressing the pressing development challenges. Recognising the potential of GenAI, the UNFPA Independent Evaluation Office (IEO) has embarked on a journey of experimentation and innovation to maximize the potential benefits of this new technology while minimizing potential risks, to strengthen the evaluation function by enhancing the effectiveness, efficiency, and timeliness of evaluations.

This initiative aligns with the UNFPA strategic plan, evaluation policy, and evaluation strategy emphasizing innovation, digitization, efficiency, and increased utilization of evaluative evidence for decision-making, adaptation, and acceleration. The strategy - informed by a needs assessment - recognizes challenges and risks in using GenAI in evaluation, offers strategic principles, and provides an implementation roadmap for optimizing the evaluation function.

This strategy will serve as a guiding framework for the innovative use of GenAI to power UNFPA evaluation function. It is also a valuable resource for the wider evaluation community to leverage GenAI to optimize evaluation, to accelerate progress towards the SDGs.

Marco Segone

Director, Independent Evaluation Office
UNFPA

Contents

1 Introduction

Evaluation at UNFPA

Evaluation is a critical oversight function at UNFPA as it provides evidence and lessons on UNFPA work on the ground, who benefits and who does not, and gives insight into what needs to be improved and how. Given that the world is halfway through implementing the Sustainable Development Goals (SDGs), and complex mega-trends and development challenges rage on, evidence-informed decisions and actions are even more critical to ensure rights and choices for the people furthest behind.

Therefore UNFPA strategic plan 2022-2025 places a heightened focus on evidence-informed learning, adaptation, accountability, and decision-making to achieve the three transformative results (zero unmet need for family planning, zero preventable maternal deaths, and zero gender-based violence and harmful practices) and fulfilling the promises of the International Conference on Population and Development Programme of Action and the SDGs.

Aligned to this context, the UNFPA evaluation policy 2024, UNFPA evaluation strategy 2022-2025 and the strategy to enhance evaluation use through communications and knowledge management 2022-2025 prioritize enhancing innovation to provide targeted decision-makers with targeted and high-quality evaluative evidence in targeted/real-time for adaptation, learning, accountability, and informed decision-making.

Following the Multi-Year Costed Evaluation Plan 2024-2027, UNFPA will undertake 39 centralized evaluations and exercises, 64 country programme evaluations, and six regional programme evaluations. To meet the increasing demand for evidence in the United Nations system, the Independent Evaluation Office (IEO) also conducts and contributes to several system-wide evidence syntheses. Within UNFPA, there is also a growing demand for timely evaluation evidence and knowledge products to inform decision-making.

With this context, the IEO is exploring innovative ways of leveraging the latest technologies to increase the efficiency and effectiveness of the entire evaluation function, including the conduct, management, communication, and facilitation of the use of evaluations, and for efficient, robust and timely evidence extraction and synthesis. Utilizing new technologies can also enhance the evaluation function's focus on leaving no one behind and reaching the furthest behind, and better integrating social and environmental dimensions in evaluation.

The external context for UNFPA evaluation function

The mega-trends[1] in the external environment and the policy and operational landscape of UNFPA set the stakes for further optimizing the evaluation function at UNFPA. With 2030 fast approaching, current estimates indicate low SDG attainment, with the United Nations Secretary-General stating as of September 2023 that "only 15 per cent of the targets are on track and many are going in reverse."[2] Considering this reality, the United Nations has prioritized digital transformation for SDG acceleration. The United Nations 2.0 policy framework represents this forward-thinking vision of a modern United Nations system that leverages a fusion of data, innovation, digital, foresight, and behavioural science expertise—known as the "quintet of change."

Recent developments in artificial intelligence and machine learning, specifically generative artificial intelligence (GenAI), are being recognized as potential means of accelerating the 2030 Agenda for Sustainable Development.[3] Indeed, the Secretary-General's establishment of an AI Advisory Body to explore the risks, opportunities, and international governance of GenAI that will feed into the 2024 Summit of the Future underscores the significance of this moment. In December 2023, the AI Advisory Body launched its Interim Report: Governing AI for Humanity, calling for a closer alignment between international norms and how AI is developed, rolled out and governed. In March 2024, the United Nations General Assembly adopted a landmark resolution (A/78/L.49) on the promotion of safe, secure and trustworthy AI systems to accelerate progress towards the full realization of the 2030 Agenda for Sustainable Development. The resolution calls for AI systems to be developed and used with full respect for human rights, focusing on privacy, non-discrimination, accountability, transparency and reliability. It also calls for reducing the gaps in access to and use of AI technologies between and within countries.

Background on AI and its use in evaluation

Throughout the development of evaluation, there has been a strong tradition of using digital technology to enable better processes and products. Using technology in evaluation has been especially true within the past two decades with the advancements in computer-assisted data collection, analysis, and reporting.[4] At the frontier of tech-enabled evaluation has been a small cadre of practitioners using emerging technologies such as big data, predictive analytics, machine learning, and traditional artificial intelligence.[5] Many of these cutting-edge technological applications for evaluation have been either limited test cases or high barriers of utilisation for the average evaluator. However, at the end of 2022, a significant inflection point for evaluation—and all knowledge-work sectors—occurred with the widespread exposure and use of GenAI technology, especially large language models (LLMs).[6] As a result, many professions and organizations have been attempting to respond to the disruption GenAI has caused—including in the evaluation sector. Much of this disruption is due

[1] These broadly include climate change, demographic shifts, inequalities and digitalization.

[2] SG/SM21945.

[3] See AI for Good; Security Council debate on AI.

[4] Raftree & Bamberger, 2014; MERL Tech, 2020a.

[5] MERL Tech 2020b; 2020c; Bamberger & York, 2021.

[6] An LLM is a type of artificial neural network, or machine learning model, that has been trained on large amounts of text data and is capable of determining how data relate to each other.

to the significant advancement in AI model performance and significantly lower barriers to use, with little to no technical experience needed by users to obtain automated responses.

Key UNFPA policies and guidelines related to the strategy

The IEO recognizes the potential for GenAI to enhance the evaluation function at UNFPA further, to support accelerated delivery of the UNFPA strategic plan 2022-2025, and has responded to the call of the United Nations Secretary-General to "develop sector-based guidelines to ensure that technology developers and other users have applicable, relatable guidance for the design, implementation, and audit of AI-derived tools in specific settings."[7] To that end, the IEO has developed one of the first known strategies of its kind in the evaluation field—a strategy for leveraging the benefits of responsible and ethical GenAI for the UNFPA evaluation function, while minimizing risks.

There are multiple UNFPA policy frameworks to support the adoption of GenAI in evaluation practices and processes. The UNFPA strategic plan 2022-2025 identifies innovation and digitization as one of six accelerators to achieve the three transformative results. Other policies have operational implications for the digital transformation of the evaluation function, such as the ITSO information security policy, the United Nations System CEB (Chief Executives Board) Digital and Technology Network's Guidance on Generative AI (July 2023), and the UNESCO Recommendation on the ethics of artificial intelligence (2022). Additional relevant ethics-related policies of the UNFPA include the UNFPA gender strategy (2022-2025) and the Guidance note for applying a human rights-based approach to programming in UNFPA (2020). Concerning the evaluation function, relevant policy frameworks inform the adoption of GenAI, such as the UNFPA evaluation policy 2024, evaluation strategy 2022-2025, and the strategy to enhance evaluation use through communications and knowledge management 2022-2025. Together these policies, guidance documents, and evaluation frameworks support the use of GenAI in evaluation, given their focus on innovation, transformative approaches for knowledge management, increased efficiency of evaluation processes and enhanced utilization of evaluative evidence in driving decisions and actions.

Strategy development process

The design of this strategy was a collaborative process. In 2023, the IEO commissioned an initial needs assessment to respond to the demand for more support from IEO staff on what GenAI tools staff can use and how to deploy those tools responsibly to address bottlenecks at various phases in the evaluation lifecycle. The needs assessment identifies the prioritized use cases for GenAI integration into the evaluation function and informs this strategy.

Key participants in this exercise include the IEO staff, regional M&E advisors, and country M&E staff/focal points. The IEO sought additional insight from staff in the Policy and Strategy Division (PSD), the Technical Division (TD), the Innovation Unit, the Legal Unit, and the Ethics Office, and in particular, closely collaborated with the Information Technology Solutions Office (ITSO). The IEO will share this strategy and insights from its implementation with the United Nations Evaluation Group (UNEG) and other evaluation offices. Beyond the United Nations system, this strategy will be helpful for the global evaluation community looking to understand how to leverage GenAI in their evaluation practice. This includes, but is not limited to, global networks like the International Organization for Cooperation in Evaluation, EvalPartners, EvalYouth, regional and national evaluation associations, and Global South evaluation networks, among others.

[7] Our Common Agenda Policy Brief 5, 2023, p. 18.

2 Goal and objectives of the GenAI-powered evaluation strategy

The GenAI-powered evaluation strategy supports the existing priorities and values of the UNFPA evaluation function, notably strengthening accountability, evidence-based decision-making, and learning. It is responsive to changing global dynamics and seeks to help the evaluation function be more agile and adaptive amidst changing contexts. Finally, it intends to strengthen an organizational culture of evidence-based decision-making for improved programming and organizational effectiveness.

Strategy goals

Internally, **strengthen the UNFPA evaluation function** by optimizing evaluation processes and products with ethical and responsible use of GenAI tools.

Externally, **co-lead the shaping of responsible and ethical GenAI-powered evaluation through global evaluation advocacy and partnerships**. Leverage collaboration particularly in the Global South, share learning, and advocate for the importance of shaping GenAI-powered evaluation globally with UNFPA staff, partners, UNEG members and the broader evaluation community.

Strategy objectives

1. Pilot/experiment GenAI-powered evaluation at a centralized level first

Test GenAI-powered evaluation tasks to enhance the timeliness, coordination, consistency, and strategic focus of centralized evaluations and, once ready, deploy promising methods to decentralized evaluations

2. Support decentralized GenAI-powered evaluations

Provide GenAI tools and training to decentralized evaluation teams with baseline levels of digital and data literacy ready to integrate GenAI into workflows, enhancing their capacity to conduct rigorous, locally relevant evaluations

3. Ensure GenAI-powered evaluation coherence within the United Nations system

Constantly and systematically share and coordinate among the United Nations entities, fostering coherence and learning about GenAI-powered evaluation across the United Nations system

4. Boost national capacity development for GenAI-powered evaluation

Leverage national stakeholders' evaluation capacities to shape and advance GenAI-powered evaluations at the national level, promoting its local ownership and use, especially in the Global South

Aligned to the M&E framework in the UNFPA evaluation strategy 2022-2025, high-level output and outcome-level measures are available in Annex 1, that will be utilized to track the implementation of this strategy.

3 Challenges and risks in using GenAI in UNFPA evaluations

It is imperative to recognize the challenges in GenAI and determine how to mitigate them to deliver the strategy's objective. The following methodological, ethical, and organizational challenges foreground any responsible integration of GenAI technology in evaluation processes.

Methodological challenges

Matching GenAI solutions to needs: Understanding the universe of GenAI methods, tasks, and solutions and matching it appropriately to the optimization needs of the evaluation function, is a significant and ongoing task. Depending on the role and phase of evaluation, there are diverse and varying optimization needs across the evaluation function and a myriad of GenAI solutions that technology firms develop regularly.

Balancing optimization trade-offs: GenAI can potentially increase the efficiency and quality of evaluation processes and products. However, improving the timeliness of the evaluation function—doing evaluation faster—should not come at the expense of accuracy, propriety, utility, and credibility—doing evaluation better.

Transparency and explainability of GenAI technology: Many GenAI tools contain layers of technology that are not readily transparent nor open for interrogation. The tools used and how evaluators used them in evaluative logic should be explainable, within reason, to support the evaluation quality assessment process.

Triangulation: The integration of GenAI tools risks introducing automation bias into the evaluation process—trusting the outputs of GenAI tool applications merely because they are derived from an automated process, leading to an overreliance on the tool. GenAI-powered evaluation should open evaluation processes to more data sources, methods, and reasoning agents for enhanced triangulation.

Bias reduction: Biases are unknown sources of error that lead to inaccurate and distorted evaluation findings. GenAI should be used to account for and reduce biases in the evaluation process and not be the source of error.

Ethical risks

Emancipatory or extractive use of GenAI in evaluation: The training, development, and maintenance of GenAI technology have historically consisted of extractive practices, from extracting data, to exploiting cheap labour for labelling data,[8] to using vast amounts of natural resources.[9] Some GenAI tech producers are mindful of this history and are working to minimize these harms. If done irresponsibly, evaluation can also be extractive for communities and individuals. GenAI-powered evaluation should be used for empowerment, not exploitation or extraction, and aligned with core ethical principles of human rights and environmental protection.

[8] See investigative reports from Time, WIRED, and Distributed AI Research Institute.

[9] Though researched more than the topic of exploited labour in the GenAI industry, see this article 'Environmental Impact of Large Language Models, for an entry-point to this body of work.

Bias and discrimination: Like the lived experiences of evaluators, GenAI tools have conditioned biases or presumptions about how the world works. Just as human evaluators should declare professional values,[10] the values and assumptions programmed into AI models should be accounted for as much as possible.[11] Integrating GenAI tools should support existing efforts to reduce bias in evaluation and not become the source of amplifying social biases, which can lead to reinforcing inequitable outcomes in evaluation.

Privacy, data protection, copyright: Large corporations have trained GenAI technology on publicly available, often copywritten data, whose creators may not have consented to training AI models. GenAI-powered evaluation should respect copyright and intellectual property rights as much as possible. Further, much of the GenAI technology is operated and maintained on remote servers and data centres, otherwise known as the cloud. This operational reality poses a challenge for adequately protecting the privacy of evaluation participants and sensitive information.

Accountability and responsibility: Using machines for knowledge work like reasoning, analysis, synthesis, and content generation blurs the line of who is ultimately responsible for the consequences of using material produced by GenAI. Clear lines of accountability for GenAI-powered evaluations, including mechanisms for monitoring and meta-evaluating evaluation or evaluation quality assessment, must be established, where humans are ultimately responsible for the outputs of GenAI processes.

Undermining human critical thinking: Related to the challenge of accountability and responsibility for the use of AI-generated outputs, there is a risk that as GenAI continues to augment and automate tasks within the evaluation lifecycle, the human capacity to generate quality content and, more importantly, think and express critical and evaluative thought may diminish among evaluators. As evaluators increasingly integrate machines into evaluation processes, evaluators should strengthen critical and evaluative thinking to interrogate the integrity of GenAI-enabled processes and GenAI-generated outputs.

Organizational challenges

Low digital and GenAI capacity: UNFPA is a large and diverse organization of skilled professionals exposed to various new and emerging technologies. There are varying degrees of digital, data, and GenAI capacity across UNFPA, necessitating customized approaches to digital transformation and evaluation capacity development.

GenAI-powered evaluation capacity building: Although GenAI has lowered barriers to using advanced analysis procedures, such as natural language processing,[12] traditionally only available to those who can code in computer languages, there is still a learning curve. Accordingly, with any learning curve, upskilling takes time and can be discouraging for some professionals who may not immediately see anticipated gains in quality processes or outputs.

Change management: Digital transformation implies changing the status quo, which can bring resistance from those who are used to working styles, known procedures, and established evaluation processes. The process of attempting to integrate GenAI to optimize the evaluation function should address legitimate concerns from UNFPA staff and consultants.

[10] See the Checklist for Evaluation-Specific Standards (CHESS).

[11] For an example of this in practice, see Anthropic's Constitutional AI approach.

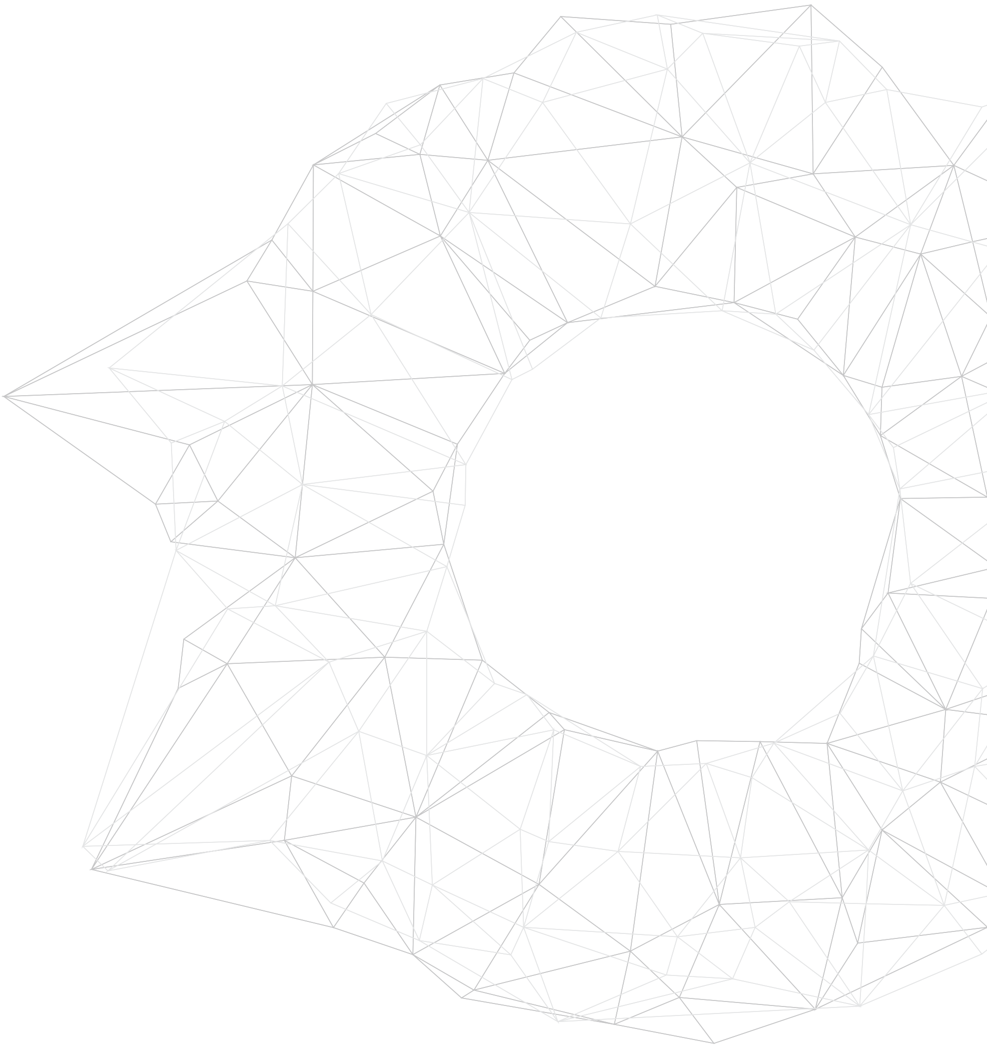
[12] For example, non-specialists can use natural language (without the need for prior computer programming languages) to perform tasks and procedures for enabling computers to process, interpret, and generate human languages.

Interagency coordination: Promoting compatibility and avoiding undue duplication[13] between United Nations agencies' evaluation units that will experiment and develop strategies and policies for responsibly using GenAI in evaluation.

Infrastructure and resources: Investing in innovation and digitization requires supporting a level of coherence with the existing suite of tools for evaluation and understanding how GenAI tools can complement, not overcrowd. The appeal of new and promising technology may entice some to pursue innovation for its own sake, not driven by users' needs.

Keeping pace with rapid GenAI advancement and evolving user needs: The GenAI space is dynamic, and the IEO anticipates obsolescence. New systematic biases may present themselves in unanticipated ways, and evaluators must make plans for staying informed and relevant, with support from ITSO and AI experts and by learning from experience of other United Nations agencies.

Incorporate principles and guidance to address evolving and unknown ethical concerns: As a large organization, UNFPA's attempts to entirely prevent or monitor users from misusing GenAI may be too challenging. The strategy for responsible and ethical use of GenAI technology in evaluation can provide guidance. However, the dynamic and fast-changing landscape means not all ethical dilemmas can be considered and prescribed against in advance. Providing strategic outlines and effective principles, as well as safe and vetted tools for responsible GenAI use is vital.



[13] While experimentation and variation in GenAI-powered evaluation practice and policy is expected in evaluation units across United Nations agencies, there should be some level of coordination.

4 Potential harms in GenAI-powered evaluation

A growing body of AI ethics research has documented how irresponsible and unethical use of AI harms and places marginalized groups at risk. While the potential gains for successful completion, increased efficiency, and higher quality of work tasks with GenAI support are compelling,[14] there is potential for misuse and unintended consequences. The following possible risks and known harms set the stakes for why using principles for GenAI-powered evaluation is critical. While the United Nations policy landscape is adapting to the disruption caused by GenAI, principles-focused evaluation (see page 13) can examine the extent to which principles for responsible and ethical GenAI-powered evaluation are being adhered to within the UNFPA evaluation function.

Potential harm to individuals

Data privacy breaches can put individuals at risk. Unmitigated bias from GenAI analyses may misrepresent individual needs or behaviour changes. Gaps in those willing to utilise GenAI in evaluation and those who are not, may exacerbate the digital divide.

Potential harm to groups

Algorithmic bias can cause or reinforce unfair discrimination or misrepresentation of sub-groups' needs during evaluation processes. These distortions can have an impact on social programme design and effectiveness.

Potential harm to communities

Affected communities may lose public trust in institutions, including UNFPA, due to perceived non-transparency or unfairness of GenAI-powered evaluations. This lack of confidence may lead to inequitable access to social programmes due to unprincipled GenAI-powered evaluation practices.

Potential harm to institutions

Reputational damage from misuse or misinterpretation of GenAI-powered evaluations is possible. Rapid scaling of flawed GenAI-powered evaluation systems (example institutions include the United Nations system and its agencies, and implementing partners) may lead to economic fallout and or partnership fallout from poor decisions.

Potential harm to ecosystems

The rapid adoption of GenAI systems with little to no environmental standards may lead to unintended ecological impacts from GenAI-driven decisions, unsustainable resource use, and carbon footprint of GenAI tool use.

[14] Dell'Acqua, et al., 2023.

5 Risk mitigation for responsible and ethical GenAI use in evaluation

Given the risks of integrating GenAI technology into knowledge work, Table 1 presents a risk assessment matrix for GenAI-powered evaluation. The table offers key questions and considerations for mitigating known harms and potential risks to assist the adoption of GenAI technology in evaluation following a responsible and ethical approach.[15]

Table 1: Risk assessment matrix

Potential risks	Key questions	Considerations for mitigating risks
Individuals: Breaches of data privacy	• How is personal/sensitive data being protected in GenAI tools? • What data security measures are in place for cloud based GenAI solutions?	• Align GenAI tools with UNFPA IT guidelines and policies • Ensure strong encryption, access controls and compliance audits • Consult with the legal team
Groups: Biased GenAI analysis or decisions	• How are the known biases in training data being addressed? • Is the GenAI model fair and unbiased regarding gender, race, and other social identities? • Are GenAI outputs accurate or convincing hallucinations?	• Evaluators exercise human judgment of the integrity of GenAI outputs, and include it as part of the evaluation quality assurance and assessment process • UNFPA evaluation managers and external evaluation teams sign the disclosure (Annex 3) on the ethical and responsible use of GenAI, assuming responsibility for the validity and quality of evaluation products • Oversight from AI ethics experts • External meta-evaluation of GenAI-powered evaluations

[15] For ideas on adapting more detailed procedural assessments from the national level to the organizational level, see UNESCO's Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence.

Potential risks	Key questions	Considerations for mitigating risks
Communities: Lack of transparency	• Are processes for collecting, cleaning, and labelling data documented? • What specific GenAI and Natural Language Processing (NLP) tasks, methods, and GenAI tools did workers use? • Can inputs, outputs, and evaluative conclusions from GenAI be explained?	• Declare the use of GenAI for transparency in evaluation materials • Maintain detailed documentation for how workers use each GenAI-powered tool • Enable human-in-the-loop learning at key stages of evaluation
Institutions: Over-reliance on GenAI	• What are the human checks on essential decisions in GenAI-powered evaluation? • Is the use of GenAI undermining evaluators' critical thinking? • How will the IEO handle new issues and challenges with the technology?	• Human judgment is paramount for evaluation quality assurance and assessment • UNFPA evaluation managers and external evaluation teams sign the disclosure (Annex 3) on the ethical and responsible use of GenAI, assuming responsibility for the validity and quality of evaluation products • Define processes for escalating unclear cases to evaluation function leadership
Reputational damage	• How are participants engaged in GenAI-powered evaluation use? • Are harms from using GenAI in evaluation monitored and addressed quickly?	• Ongoing communication with evaluation partners • Establish processes to identify and resolve issues
Ecosystems: Environmental impacts	• What is the carbon footprint and resource use of tools? • Are sustainable computing practices used? • Does using a GenAI tool conflict with guidance from UNFPA on environmental standards in evaluation?	• Identify models used in GenAI tools and seek their model card[16] documentation • Assess and monitor energy use and materials • Select GenAI tools with progressive energy policies

[16] AI model cards are concise documents that provide essential information about machine learning models. They serve as a transparent and accessible way to communicate key details about a model, including its intended use, performance metrics, limitations, and potential biases. See this guidebook on model cards.

6 Principles underlying GenAI use in evaluation

Fairness, Accountability, Transparency, and Ethics (FATE) frameworks for AI

The challenges and risks listed in the previous sections are not specific to integrating GenAI in evaluation. For years, private corporations, organizations, multilateral agencies, and research institutions have been developing a growing body of guidance frameworks for Fairness, Accountability, Transparency, and Ethics (FATE) consisting of values, principles, rules, prescriptions, and policy recommendations for the ethical and responsible application of AI from their experiences.[17] The following section presents an overview of a few recent FATE frameworks in the United Nations system and their guiding values and principles.

Sample of United Nations FATE frameworks on AI[18]

AI Advisory Body's Interim Report: Governing AI for Humanity (December 2023)

- Inclusivity in access to and use of AI, focusing on the Global South
- Broader accountability framework in the public's interest, going beyond the do no harm principle
- Centrality of data governance
- Universal, networked and multistakeholder AI governance
- Anchored in the United Nations Charter, International Human Rights Law, and SDGs

UNESCO Guidance for Generative AI in Education and Research (September 2023)

- Promote inclusion, equity, linguistic and cultural diversity
- Promote plural opinions and plural expressions of ideas
- Monitor and validate GenAI systems for education
- Develop AI competencies, including GenAI-related skills for learners
- Build capacity for teachers and researchers to make proper use of GenAI
- Protect human agency
- Test locally relevant application models and build a cumulative evidence-base
- Review long-term implications in an intersectoral and interdisciplinary manner

[17] See the United Nations Resource Guide on Artificial Intelligence Strategies, particularly Annex III.

[18] In addition, see the UNFPA Guidance on safe and ethical use of technology to address gender-based violence and harmful practices (under update). The principles in this guidance apply to GenAI as well.

Digital & Technology Network (DTN) Guidance on the Use of Generative AI Tools in the United Nations System (July 2023)

- Be mindful of using raw GenAI content for delegated decision-making or unsupervised provision of answers
- Ensure the access rights applied to training data are preserved in the deployed model.
- Maintain transparency, accountability, and explainability
- Assess the risks, costs, and limitations of the specific GenAI model
- Leverage technical expertise or obtain it at the right level
- Follow established review and approval processes for institutional initiatives
- In building and training new models, data used to train the model should be quality assured, representative and unbiased
- Encourage diversity and inclusion
- Be mindful of the impact of AI and GenAI

Chief Executives Board for Coordination (CEB) Principles for the ethical use of artificial intelligence in the United Nations system (October 2022)

- Do no harm
- Defined purpose, necessity, and proportionality
- Safety and security
- Fairness and non-discrimination
- Sustainability
- Right to privacy, data protection, and data governance
- Human autonomy and oversight
- Transparency and explainability
- Responsibility and accountability
- Inclusion and participation

UNESCO Recommendation on the Ethics of Artificial Intelligence (January 2022)

- Proportionality and do no harm
- Safety and security
- Right to privacy and data protection
- Multi-stakeholder and adaptive governance and collaboration
- Responsibility and accountability
- Transparency and explainability
- Human oversight and determination
- Sustainability
- Awareness and literacy
- Fairness and non-discrimination

Effectiveness and ethical principles for human rights-based GenAI-powered evaluation

The frameworks above present some values and principles for the responsible and ethical use of GenAI in the United Nations system. Specifically, where framework authors express guidance as principles (or prescriptions), they are often effectiveness principles or ethical principles. Effectiveness principles focus on what works or how to use GenAI effectively. Ethical principles focus on what is right and how to uphold ethics in using GenAI. This strategy mainly guides the use of GenAI in evaluation practice in the form of effectiveness and ethical principles.

Why effectiveness and ethical principles?

Using principles is a reasonable approach for navigating the uncertainties of evaluation in the age of GenAI. The development and practical application of GenAI is a dynamic landscape, and, due to their nature, principles can mitigate the obsolescence of overly prescriptive rule-based guidance. The extent to which GenAI responsibly optimizes evaluation processes and products depends on the extent to which effectiveness and ethical principles are applied.

Principles-focused evaluation underpins UNFPA’s GenAI-powered evaluation strategy

The following principles of GenAI-powered evaluation in section 5 were informed by the principles-focused evaluation (PFE) approach explicitly developed for principles-driven organizations to understand how and how well they navigate innovation in complex contexts.[19] The IEO used the following GUIDE criteria in crafting GenAI-powered evaluation principles:

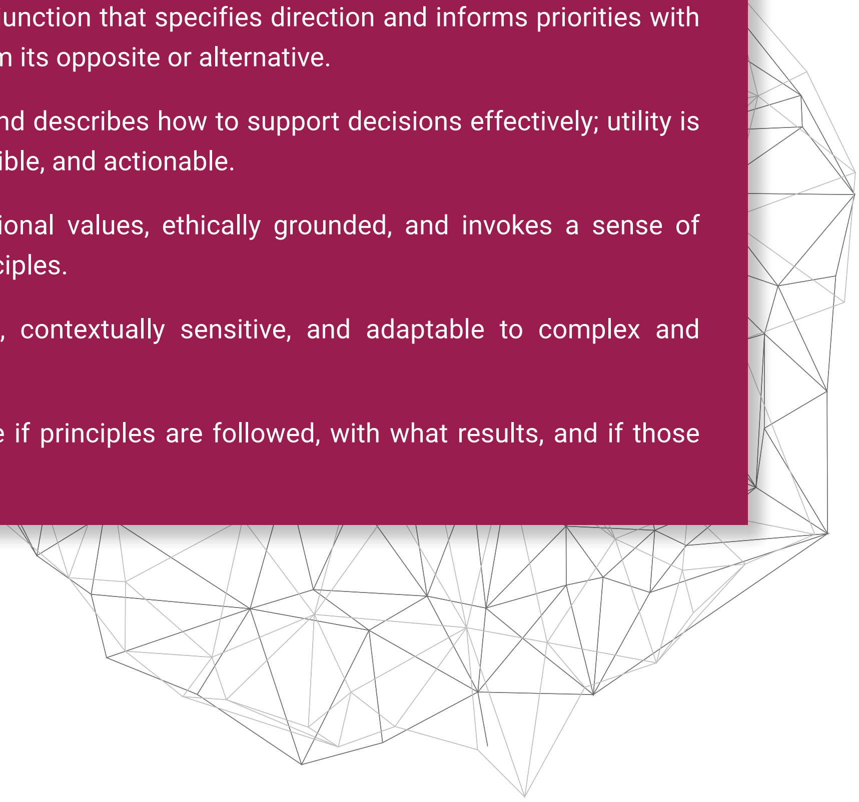
(G) meaningful Guidance: a prescriptive injunction that specifies direction and informs priorities with action-oriented wording and distinctive from its opposite or alternative.

(U) Useful: points toward desired results and describes how to support decisions effectively; utility is high when principles are interpretable, feasible, and actionable.

(I) Inspiring: derived from core organizational values, ethically grounded, and invokes a sense of purpose among those intended to use principles.

(D) Developmentally adaptable: enduring, contextually sensitive, and adaptable to complex and dynamic situations.

(E) Evaluable: able to document and judge if principles are followed, with what results, and if those results lead to desired outcomes.



[19] See Principles-focused Evaluation—The Guide by Michael Quinn Patton.

7 Key principles for GenAI-powered evaluation at UNFPA

The following section presents principles for GenAI-powered evaluation informed by the UNFPA evaluation strategy 2022-2025, notably strategic principles. As a note, the IEO intends these principles to evolve and sharpen over time by implementing the GenAI-powered evaluation strategy.

1. Pursue a demand-driven approach for GenAI-powered evaluation: Ensure GenAI-powered evaluation focuses on the demand of meeting needs and not integrating specific solutions for the sake of using GenAI. If a non-GenAI solution meets an evaluation need better than a GenAI solution, pursue the non-GenAI solution.

2. Promote diversification and innovation in evaluation: Regularly seek and integrate new GenAI-powered tools in evaluation processes. Consider how GenAI-powered evaluation can lower technical barriers to data collection and analysis.

3. Uphold quality and credibility in evaluations: Ensure rigorous and transparent GenAI-assisted data analysis, contributing to credible findings and recommendations. Apply the Evaluation Quality Assurance and Assessment (EQAA) framework for GenAI-powered evaluations and adapt the EQAA to account for unique quality considerations for GenAI-powered evaluations.20

4. Maximize the use and utility of evaluations: Utilize GenAI to enhance the accessibility, timeliness, and relevance of evaluation outputs, facilitating their use in decision-making and learning. Providing targeted information to targeted audiences for targeted decision-making is the aim—not the adoption of this new technology for its own sake.

5. Adhere to a human rights-based approach in evaluation: Embed the United Nations principles for ethical AI use into the evaluation field, as well as uphold the UNEG ethical principles for AI in United Nations evaluations. The later focuses on ethical principles such as transparency, fairness, participation, privacy, accuracy, a human-centered approach, and human rights. For example, aligning with the human rights approach and the principle of 'leaving no one behind' ensures that GenAI usage avoids exclusion or disadvantage to any group.

6. Foster evaluation capacity development: Cultivate GenAI capacity among evaluation stakeholders especially in the Global South, empowering them to effectively use, interpret, and critique GenAI-powered evaluations and not be left behind in realizing technological advancements.

[20] See Change from the Outside: towards credible third-party audits of AI systems in UNESCO's Missing Links in AI Governance.

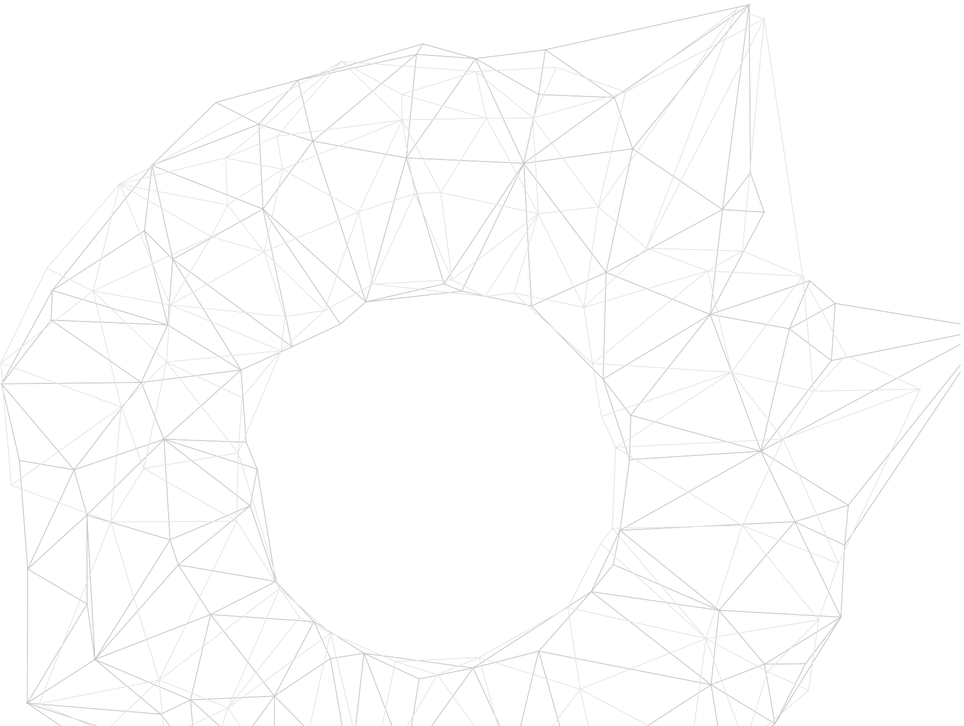
8 GenAI use cases for UNFPA evaluation

As mentioned earlier, the IEO conducted an internal needs assessment of the evaluation function in parallel with and to inform the development of this strategy. The needs assessment used the following criteria to prioritize user needs by evaluation phase: number of individuals affected by need, size of need/magnitude of discrepancy, degree of anticipated optimization, and data source agreement. Needs across each of the phases of the evaluation function were identified along with potential solutions, including information management, evidence synthesis, data analysis, report writing, dissemination, and quality assurance. For a list of prioritised user needs by evaluation phase, see Annex 2.

Selection of GenAI technologies and tools

Given the vast array of turn-key solutions, the IEO recommends the following criteria for assessing custom and market-based solutions. These include the availability of resources; legal requirements; procurement requirements; IT requirements; key features of solution; alignment with UNFPA IT architecture; robust information security and data privacy; start-up, continuation, and maintenance costs; anticipated optimization of an evaluation process or workflow; anticipated risks and harms across use cases; training requirements; other unique factors. These criteria will guide the selection, piloting, and scaling of GenAI solutions to meet UNFPA evaluation function needs.

While a host of third-party market-based solutions could be leveraged for a suite of GenAI solutions for different evaluation needs, locating GenAI solutions within the Google ecosystem is viewed as an ideal scenario that fully or partially meets most criteria for solution selection. Given the institutional relationship, the IEO should work with Google enterprise representatives and ITSO staff to test the feasibility of developing no- to low-code solutions to meet most use cases emerging from the needs assessment. In addition, IEO will continue to explore new use cases as more GenAI tools become available.



9 Role and responsibilities for advancing GenAI-powered evaluation

The successful implementation of a GenAI-powered evaluation strategy requires intentional coordination among and collaboration between various evaluation function key players and related business units, including the IEO Director, AI specialists internal to IEO, external evaluation advisors, evaluation specialists, evaluation capacity and communications staff, regional M&E advisors, country M&E staff/focal points, ITSO, PSD, TD, UNEG, and potential external experts in meta-evaluation, AI ethics, and digital transformation.

Differentiated capacities for different target groups

Building GenAI expertise and skill sets should be tailored to the roles and responsibilities that comprise the evaluation function. Training and upskilling of staff will include general GenAI literacy for all IEO staff and targeted upskilling for managers, analysts, quality assurance leads, and external contractors based on their specific evaluation roles. The IEO will seek partnerships and collaborations among United Nations agencies working on similar digital transformation initiatives.

10 Strategy implementation roadmap

1. A phased approach to GenAI adoption and deployment: Develop and field test principles for deciding when, how, and to what extent the IEO can integrate GenAI with evaluation based on needs, capacities, and ethical considerations. An initial principle for this decision is to pursue GenAI-powered evaluation optimization when there is high user demand, medium to high optimization potential, and non-GenAI solutions may not adequately address the evaluation function needs.

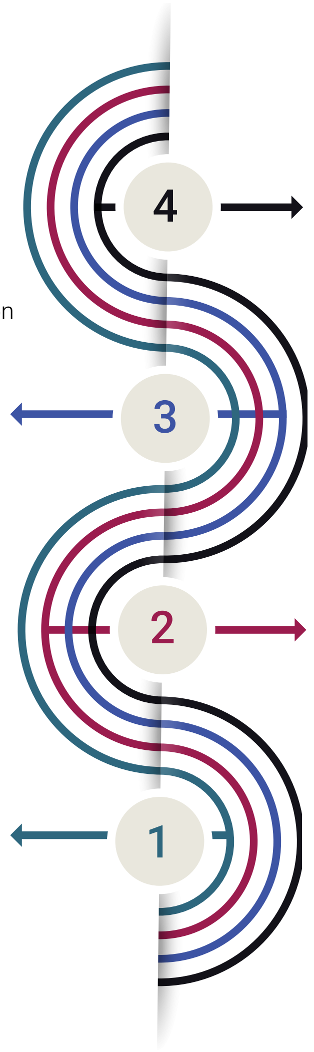
2. Custom GenAI solutions for evaluation: Those who integrate GenAI into evaluation processes and products should recognize there is no silver-bullet or one-size-fits-all solution to optimizing the evaluation function. The needs assessment revealed 30 unique user needs that the IEO could optimize through GenAI solutions. Developing custom solutions to prioritize user needs can complement adapting available GenAI solutions from institutional partners and third parties. In partnership with ITSO, the IEO is exploring custom applications for evaluation evidence synthesis and evaluation quality assurance and assessment.

3. Change management and communication with stakeholders: Integration of GenAI in evaluation processes will be most successful when done collaboratively, responsive to the concerns and needs of primary intended users, including evaluation advisors, evaluation managers, evaluation specialists, evaluation communications staff, Regional M&E advisors, and country M&E staff/focal points. Regular communication to keep staff updated on processes and experimentation with GenAI-powered evaluation will help realign the evaluation function as UNFPA evolves from traditional to tech-enabled approaches. A key change management aspect is planning for and facilitating the upskilling and GenAI-powered evaluation capacity development. Conduct a baseline assessment of evaluation function optimization efforts to monitor the rate of change and communicate results.

4. An iterative and adaptive approach to digital transformation: A dynamic GenAI-powered evaluation strategy requires constant follow-up and review, assessing the feasibility of new tools and upgrading of tools. The IEO will track changing user needs, optimization potential, and the changing landscape of third-party solutions. Given the GenAI ecosystem is evolving rapidly, the strategy implementation roadmap ahead will need to be agile, and this agility is vital to a successful GenAI strategy.

5. Strategy update and sustainability: Developing practices and policies that ensure the ethical and responsible use of GenAI ensures the long-term sustainability of GenAI-powered evaluation efforts. The IEO will seek to keep track of external guidance developments across relevant disciplines and fields, coordinate and learn from UNEG agencies, take stock of piloted solutions, and scale at the appropriate pace. The IEO will attempt to "future-proof" the evaluation function by anticipating potential disruptive technologies and planning for unexpected obstacles within emerging technologies that could impact or alter existing GenAI strategies and implementations.

Milestones and timelines for specific GenAI initiatives: Actionable steps for immediate implementation



Feasibility and needs assessment (Q2-Q3 2023)

a. Identify evaluation process user needs

b. Map existing GenAI solutions to prioritized use cases

Piloting phase (Q4 2023-Q2 2024)

a. Establish appropriate measures to observe changes in evaluation efficiency and effectiveness

b. Custom build high-priority GenAI solution for evaluation evidence synthesis and evaluation quality assurance and assessment.

c. Pilot GenAI-powered evaluation methods for secondary data analysis and synthesis for a centralized evaluation

d. Pilot GenAI-powered method for data extraction and analysis for a meta-synthesis exercise

GenAI capacity building (Q1-Q4 2024)

a. Develop and deliver training modules on GenAI applications in evaluation, focusing on specific evaluation task optimizations, GenAI-powered evaluation principles, and responsible and ethical GenAI-powered evaluation.

b. Offer on-demand consultation and troubleshooting support for GenAI applications in evaluations.

c. Facilitate networking and knowledge exchange among GenAI and evaluation professionals

Scaling and sustainability (Q1-Q4 2025)

a. Take stock of key lessons from the use of GenAI in a centralized evaluation

b. Conduct internal quality assurance on centralized GenAI-powered evaluation

c. Phased and prioritized roll-out of GenAI solutions in decentralized evaluations

d. Support continued development of GenAI tools and their respective clauses and disclaimers to ensure ethical and responsible use

Annex 1: M&E framework for GenAI-powered evaluation strategy

The IEO will use the following monitoring and evaluation framework, aligned with the Evaluation Strategy 2022-2025, to implement the GenAI-powered Evaluation Strategy and ensure the IEO is progressing through this strategy's piloting and scaling phases.

Outcomes	Outputs	Indicators/Targets
Area 1: Effective centralized evaluation systems are implemented		
Centralized UNFPA evaluations have optimized their processes and products	Centralized UNFPA evaluations have adopted GenAI-powered evaluation methods	**Output Indicator:** Percentage of centralized evaluations that have responsibly and ethically integrated GenAI solutions into evaluation processes **Target:** 50% **Outcome Indicator:** Percent of centralized GenAI-powered evaluations demonstrating optimized processes (efficiency) and products (quality) **Target:** 40%
Area 2: Effective decentralized evaluation systems are implemented for greater accountability, improved programming, and a stronger culture of results		
Decentralized UNFPA evaluations have optimized their processes and products	Decentralized UNFPA evaluations have adopted GenAI-powered evaluation methods	**Output Indicator:** Percentage of decentralized evaluations that have responsibly and ethically integrated GenAI solutions into evaluation processes **Target:** 25% **Outcome Indicator:** Percent of decentralized GenAI-powered evaluations demonstrating optimized processes (efficiency) and products (quality) **Target:** 20%

Outcomes	Outputs	Indicators/Targets
Area 3: Evaluation coherence within the United Nations system is promoted		
Joint and system-wide evaluations have optimized their processes and products	Joint and system-wide evaluations have adopted GenAI-powered evaluation methods	**Output Indicator:** Percentage of joint and system-wide evaluations that have responsibly and ethically integrated GenAI solutions into evaluation processes **Target:** 10% **Outcome Indicator:** Percent of joint and system-wide GenAI-powered evaluations demonstrating optimized processes (efficiency) and products (quality) **Target:** 5%
Area 4: National evaluation capacities for monitoring and evaluation systems are strengthened		
National-level evaluation function is optimized through GenAI-powered evaluation	Strategic multi-stakeholder partnerships on responsible and ethical use of GenAI have been established	**Output Indicator:** Number of international events in which UNFPA IEO led discussions on GenAI-powered evaluation discussions **Target:** 2 **Outcome Indicator:** At least one strategic multi-stakeholder partnership focusing on the Global South, is actively addressing responsible and ethical use of GenAI **Target:** Yes

Annex 2: List of prioritised user needs by evaluation phases

Preparatory phase

Contact Relationship Management/Information Management System (recruitment): identifying and recruiting talented evaluation consultants/contractors

Generate draft ToR: generate initial ToR based on key inputs.

Design phase

Evidence synthesis (evaluation team): evidence repositories, general synthesis support, portfolio synthesis, cross-cutting issue and recommendation synthesis

Evidence synthesis (search & retrieval): search & retrieval of conclusions, recommendations, intervention types, citations, enterprise resource planning, and other evidence discovery

Evidence synthesis (summarization): summarization of documents for descriptions of the object of evaluation, background, and context for ToR, senior meeting briefs, and other reports.

Field phase

Data analysis (qualitative): qualitative analysis of unstructured text data, including categorization, pattern recognition, coding, and sentiment analysis.

Data analysis (transcription): automatic voice recognition for transcription generation

Data analysis (quantitative): quantitative analysis of extensive secondary dataset analyses to identify patterns, correlations, and trends.

Reporting phase

Report writing (initial): support the initial generation of report designs, drafts, background and context, questions, findings and recommendations, edits, visualizations, and CPD development.

Report writing (readability): improve accessibility and readability of writing with non-technical language.

Report writing (feedback process): improve burdensome comment feedback process.

Dissemination and facilitation of use phase

Dissemination (translation): inclusive translation of reports and comms assets into Braille, closed captions, non-United Nations languages, sign language, indigenous languages

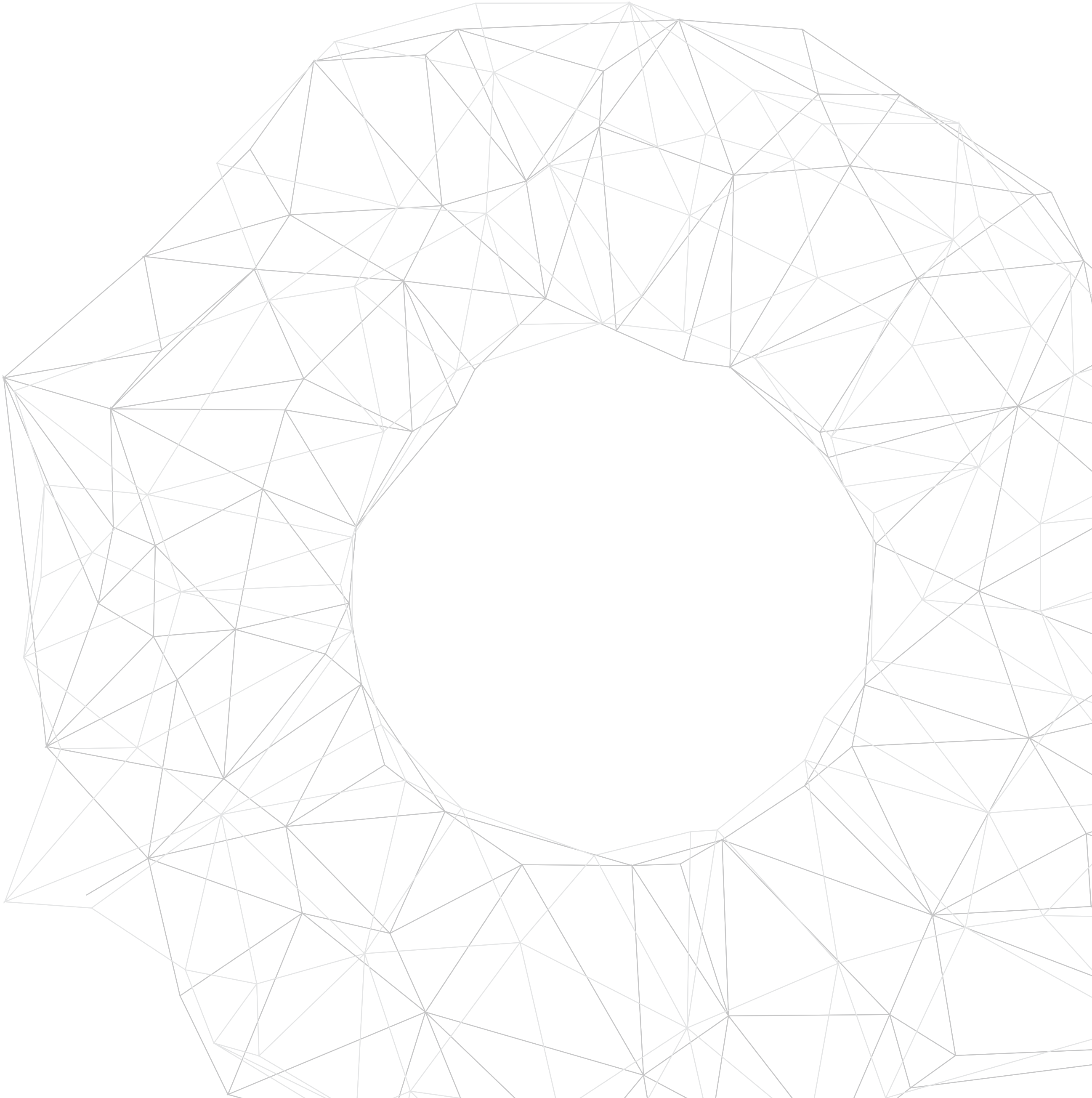
Dissemination (audience segmentation): audience identification and targeting.

Dissemination (slide deck generation): generate a presentation from a text document.

Evaluation quality assurance and assessment

Quality assurance (process): speed up the process, make it more iterative and efficient, improve procedures, and improve adherence to existing evaluation policy.

Quality assurance (product): evaluation designs, communication assets, the strength of evaluative reasoning in reports, degree of triangulation.



Annex 3: AI usage clauses

For IEO consultants

Ethical use of Artificial Intelligence (AI) in evaluation work

AI technologies cannot be used in the framework of this contract unless a prior written agreement is obtained from IEO. Upon this prior agreement, the consultant is obligated to disclose the utilization of AI tools in evaluation and commits to upholding ethical standards and accuracy in the application of AI tools.

1. **Prior approval for utilization of AI tools:** The use of AI tools must be explicitly agreed upon and approved in writing by the IEO Director.
2. **Declaration of the utilization of AI tools:** If the use of AI tools in evaluation is agreed upon with IEO, the consultant must be transparent and declare the use of AI tools in evaluation work and other work-related tasks, specifying the nature of AI usage. The AI tools utilized in work-related tasks must include only those tools that are vetted by IEO.
3. **Verification of accuracy:** The consultant commits to diligently checking the accuracy of AI-generated results and assumes full responsibility for its reliability and validity.
4. **Ethical and responsible use:** The consultant is obligated to uphold ethical principles in the use of AI in work-related tasks, as well as relevant regulations that govern the use of AI in the United Nations system. This includes the Digital & Technology Network Guidance on the Use of Generative AI Tools in the United Nations System, Principles for the Ethical Use of Artificial Intelligence in the United Nations System, and UNFPA Information Security Policy. The consultant commits to employing AI tools that adhere to principles of non-discrimination, fairness, transparency, and accountability. The consultant will adopt an approach that aligns with the principle of 'leaving no one behind', ensuring that AI tool usage avoids exclusion or disadvantage to any group.

For companies

Ethical use of Artificial Intelligence (AI) in UNFPA evaluation work

1. **Ethical principles:** *[Add company name]* (hereafter known as 'the company') agrees to uphold and adhere to the Principles for the Ethical Use of Artificial Intelligence in the United Nations System, in all aspects of AI utilization in UNFPA evaluation work and analysis.
2. **Verification of results:** The company commits to implementing verification processes to validate the results generated by AI analyses and assumes responsibility for the reliability and validity of the analyses.
3. **Monitoring of analyses:** Following a review by UNFPA, the company agrees to rectify any biases, errors, or shortcomings in the AI analyses promptly.
4. **Transparency and explainability:** Along with the AI analyses, the company shall provide a clear description of the methodology employed for the analyses.

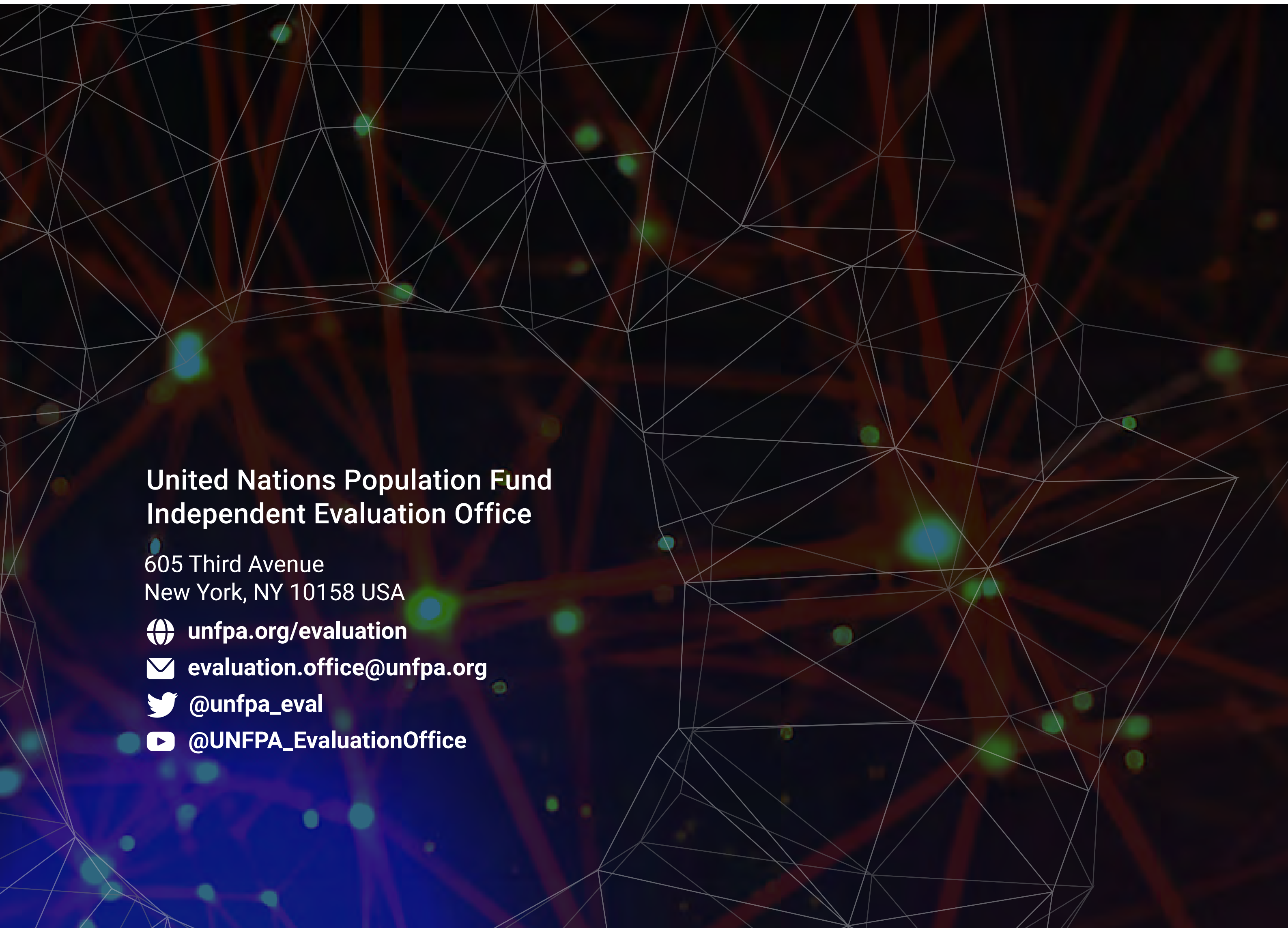
By signing below, the company acknowledges their understanding and acceptance of the obligations outlined in this 'Ethical use of AI in UNFPA evaluation work' clause.

Annex 4: Template disclaimer on the use of AI in evaluation reports and materials

This report incorporates the use of Artificial Intelligence (AI) technologies to enhance and support *[mention for which purpose the AI tool was used, e.g., for information synthesis, data analysis, generating insights, etc.]*. The AI tools utilized in this report adhere to UNFPA's AI Usage Clause, ensuring ethical and responsible use, transparency, validation of results, and compliance with relevant internal regulations. For details on the specific AI methodologies and tools used and details regarding the validation of AI-generated results, refer to the annexure *[add annex #]* on ethical AI utilization.

Driving evidence-based actions

Delivering rights and choices for all



United Nations Population Fund
Independent Evaluation Office

605 Third Avenue
New York, NY 10158 USA

unfpa.org/evaluation
evaluation.office@unfpa.org
@unfpa_eval
@UNFPA_EvaluationOffice