



Guide to the responsible use of artificial intelligence in the public sector

Governance & risk management

Israel National Digital Agency



May 2026

Version 1.0

** This is an informal translation of the Hebrew document published by the Israel National Digital Agency in May 2026. Minor modifications have been made for improved clarity and readability. In case of a discrepancy, the Hebrew version prevails.

Table of contents

Executive summary	3
1 Preface	5
1.1 Background	5
1.2 Target audience	6
1.3 Underlying principles	6
1.4 Main components of this guide	7
2 Principles for the responsible use of AI in organizations	8
3 AI Governance, key roles and responsibilities	10
3.1 Senior management	10
3.2 Chief AI Officer	11
3.3 AI Governance Forum	11
3.4 Business Process Manager	11
3.5 End users	12
3.6 Israel National Digital Agency	12
4 Risk management processes for AI systems	13
Annex A: Chief AI Officer – roles and responsibilities	15
Annex B: Roles of the Business Process Manager	17
Annex C: Risk management method for AI systems	18
Annex D: End user guidelines	34
Annex E: Sample policy on the responsible use of AI	37
Annex F: Glossary	43

Executive summary

We are at the dawn of the age of artificial intelligence, with AI applications becoming increasingly ubiquitous and shaping our present and future. Like other revolutionary technologies, AI can be used as a tool to bring about both positive and negative changes. For public service organizations, the challenge is to establish the appropriate frameworks and processes so that AI technologies can be used to improve public services, and to benefit society at large.

Government Decision 3574 on accelerating the government's digitization, data and AI capabilities through the Nimbus public cloud (December 4, 2025), adopted the goals and principles of Israel's data and artificial intelligence strategy. These included the development of appropriate frameworks for promoting the use of AI. The decision highlights "the importance the Israeli government attaches to the continued strengthening of data and development abilities and the integration of AI capabilities, aiming to streamline and enhance the activities of government entities by bolstering responsible use and creating cross-government infrastructure".

Against this backdrop, the Israel National Digital Agency has developed this guide on the responsible use of AI in the public sector, working in collaboration with the Ministry of Justice's Office of Legal Counsel and Legislative Affairs and with the Ministry of Innovation, Science and Technology. It also consulted with a broad range of government entities, private sector experts, research institutes and academia. This guide charts a process for developing an agile, practical and balanced policy for organizations, and offers best practice recommendations for the responsible use of AI in the public sector, in consideration of the associated benefits, challenges and risks.

The guide focuses on three pillars:

1. Principles for the responsible use of AI, based on the OECD principles for responsible stewardship of trustworthy AI;
2. Structured organizational concepts and processes to implement those principles ("AI governance");
3. Risk management process for the integration of a given [AI system](#), based on a risk level classification.

The guide features the following elements:

- ✦ **A traffic light method for risk management:** A risk management process for AI systems that sets forth a differential analysis and classification based on the system's risk level. The higher the risk, the more comprehensive the authorizations and control measures need to be. Accordingly, the guide proposes different authorization tracks for AI systems based on their risk levels. The green track, allowing for rapid authorization, is used for systems with low risk levels. The yellow track, in which authorization is subject to comprehensive risk management, is used for systems with a medium-level risk. The red track, in which authorization is subject to meticulous risk management and oversight by the organization's management, is used for high-risk systems.

- ✦ **Support for experimentation:** Promoting the establishment of a blue track, which allows government entities to engage in limited and supervised experimentation, so that they can learn how the tool works in practice.
- ✦ Comprehensive and uniform holistic **risk and benefit assessment**.
- ✦ **Mapping of technological tools and other means** to address known AI risks (biases, hallucinations, privacy, etc.).
- ✦ **AI Governance Forum:** The guide recommends establishing an AI Governance Forum in certain organizations, and describes the potential roles of such a forum.
- ✦ **An organizational policy template** for the responsible use of artificial intelligence, based on the recommendations in this guide is attached to this document, which organizations can adopt. The template specifies recommended fields of activity, such as transparency and public awareness measures, the method for handling [AI incidents](#), an AI literacy program, and instructions for the organization's employees.
- ✦ **A glossary of common terms** in the field of artificial intelligence.

This is the first version of the guide. It incorporates changes and comments received after the draft version for public comment was published, in June 2025. Israel National Digital Agency invites all stakeholders to continue sending comments and suggestions at the following email address: ResponsibleAI@digital.gov.il.

1 Preface



1.1 Background

In recent years, the use of AI technologies in the public and private sectors has expanded, in Israel and abroad. OpenAI's public launch of ChatGPT, in late 2022, first enabled broad public access to **generative artificial intelligence (GenAI)**. Since then, the technology has been developing at an accelerated pace. This includes the development of AI agents that can perform a range of complex operations, based on detailed instructions.

AI technologies hold great promise for growing the economy and driving sustainable development, improved productivity, greater social well-being, and a higher standard of living. Investment in these technologies and the promotion of innovation are expected to allow Israel to remain at the vanguard of technological developments. In the public sector, AI can enhance public services and make them more accurate, create uniform governance, and make government smarter and more efficient. However, alongside its potential, AI has certain characteristics and limitations, such as the randomness of outputs, biases and hallucinations. Moreover, users tend to be uncritically reliant on the outputs of such systems. These characteristics and limitations bring about various types of risks and challenges – operational, social, legal, ethical, security-related, etc. – which are particularly relevant when AI is used in the public sector. These risks and challenges are further amplified by the increasing ability to integrate different systems and to use end-to-end AI agents to manage complex processes in an organization. To fully harness the enormous potential of these technologies, such challenges must be addressed wisely and methodically.

This guide offers recommendations on work methods for public entities that wish to integrate AI systems into their activities, with an emphasis on organizational, technological and business aspects. It provides a structured process for assessing, managing and minimizing the unique risks of AI systems. This aims, inter alia, to reduce the uncertainty inherent in the application of this technology. This guide does not detract from other applicable protocols or guidelines.

This guide is the product of a collaboration between the Israel National Digital Agency, the Ministry of Justice's Office of Legal Counsel and Legislative Affairs, and the AI Policy and Regulation Center at the Ministry of Innovation, Science and Technology. Furthermore, a number of public sector entities were consulted, including the Israel National Cyber Directorate, the Privacy Protection Authority, the Israel Regulatory Authority, the Accountant General, the Government Procurement Administration, a number of chief data officers (CDOs), researchers, and academics, private companies, and subject-matter experts and counterparts working in leading countries in this field.

1.2 Target audience

This guide is designed for public sector entities that wish to integrate AI into the activities and processes they manage. It includes components of **governance**, i.e., the **roles and responsibilities** of the chief actors in the process of integrating AI into public sector entities, and it sets out guidelines for the risk management processes they should conduct.

This guide also contains guidelines for **end users**, intended for employees of public sector organizations that use AI tools.

1.3 Underlying principles

This guide's general approach is to support and promote the use of AI in the public sector through a structured procurement and development process that weighs benefits against risks. It aims to serve as a meaningful tool to support decisions on how AI is to be integrated into the organization. **The guide is based on a number of principles:**

- ✦ **Risk management:** Using AI systems responsibly does not require complete risk avoidance. The guide advocates for a differentiated approach. Generally, higher risks call for more extensive risk management measures and more comprehensive control processes. Three authorization tracks are proposed for advancing an AI project, using traffic light colors (green, yellow and red).
- ✦ **Experimentation:** The risks involved in the use of a given system are not necessarily fully clear at the outset. At times, experimentation and information gathering are necessary. Therefore, in addition to the three tracks mentioned above, the guide proposes a blue track that allows an organization to conduct limited and controlled experimentation with a particular system.
- ✦ **Generality and genericness:** This guide serves as a general recommendation that can be customized to the organization's area of activity and operational context.
- ✦ **Process guidance:** The guide sets forth the overall process that the organization should adopt but does not prescribe a specific outcome threshold.
- ✦ **Dynamism:** The guide will be updated periodically, to reflect the needs of public entities, technological developments, changes to Israeli and international normative frameworks and best practices.
- ✦ **Alignment with government policy principles:** The guide is aligned with Government Decision 3574 of December 4, 2025 on ["Accelerating the government's digitization, data and AI capabilities"](#). It is consistent with ["AI Journey"](#), a document on the technical and business aspects of integrating an AI system in a public organization. The guide also conforms with other guidelines and policy documents, including the Israel National Digital Agency's ["Directives on ICT Risks"](#),¹ ["Principles of Policy, Regulation and Ethics in Artificial Intelligence"](#), prepared by the Ministry of Innovation, Science and Technology and the Ministry of Justice's Office of Legal Counsel and Legislative Affairs,² the ["National Plan for Artificial Intelligence: the Situation in 2025"](#), and the Israel Regulatory Authority's ["Guide on Regulation Risk Management"](#).

1 Including detailed instructions, primarily guidance for the government's cybersecurity unit ("YAHAV"), on the [Secure use of AI-based chat systems](#).

2 The document recommends, inter alia, adopting an approach of sectoral regulation (as opposed to sweeping horizontal regulations). For an illustration of this approach, see the [Final Report](#) issued by the Interagency task force on AI in the Financial Sector.

- ✦ **Consistency with leading standards and regulatory regimes:** Many international frameworks were taken into consideration, such as: (1) the [OECD's principles on trustworthy AI](#); (2) the Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (the [CAI](#))³; (3) the [HUDERIA](#) methodology, which complements the CAI; (4) similar instructions and guidance in the [United States](#), the [United Kingdom](#), [Australia](#), [Canada](#), and [Singapore](#), and (5) international standards.⁴
- ✦ **Emphasizing the use of the governmental cloud environment:** Israel offers a broad range of AI cloud computing services, available either directly, through cloud services providers in the Nimbus project (Tier 1), or through other suppliers (Tier 5 products),⁵ including some of the world's leading GenAI applications, installed on sites running in secured environments. These services were procured in full compliance with Israeli law, and in accordance with the terms of use determined in the Nimbus tender, themselves based on the requirements of the government ministries. Therefore, preference is given to using AI applications in these environments, and this guide reflects this preference.

1.4 Main components of this guide

This guide contains three short chapters:

1. Principles for the responsible use of AI;
2. A recommended organizational framework for implementing those principles, with defined roles and responsibilities to actors within the organization;
3. A risk management process for integrating a given AI system.

It includes six annexes: a detailed breakdown of the roles and responsibilities of the organization's senior management, those of its Chief AI officer, guidelines for end users, a proposed risk management methodology, a template for an organizational responsible AI policy, and a glossary of terms.

This guide is not a legal document. While it touches on issues that also raise legal questions, the guide does not address these exhaustively. The Ministry of Justice's Office of Legal Counsel and Legislative Affairs is currently drawing up a legal guide (the "**Legal Guide**"), intended to address some key legal issues in connection with public sector AI use, such as transparency, fairness and biases, compliance with administrative law (such as the duty to give reasons), and copyright. The Office of Legal Counsel and Legislative Affairs is available to support the ministries' legal offices with guidance on legal issues that may arise.

Comments, reactions and questions may be sent to the following e-mail address:
ResponsibleAI@digital.gov.il

³ The convention applies mainly to the use of AI in the public sector, and focuses on risks to human rights, democracy and the rule of law. The State of Israel was involved in the drafting negotiations and signed the convention on September 5th, 2024. Israel has not yet ratified the convention, such that it is not formally binding.

⁴ The guide was based on the following documents: **ISO standards**: - Standard [ISO 23894](#) on risk management for AI systems, [ISO 42001](#) on AI system management; [NIST AI 100-1](#) – AI Risk Management Framework Document, written by the National Institute of Standards and Technology (NIST).

⁵ The catalogue of products approved for purchase for Tier 5 is available [here](#).

2

Principles for the responsible use of AI in organizations

Responsible use of AI is an approach that encourages organizations seeking to implement AI to do so in an informed way, while taking into account its expected positive and negative implications and other considerations, throughout the product's lifecycle, from the development stages until cessation of use by the end user.

The OECD described this approach in its recommendations on "[Responsible Stewardship of Trustworthy AI](#)", first published in 2019 and updated in 2024. The recommendations include several non-binding principles of responsible use of AI. They are summarized below:⁶

- 1. AI for growth, sustainable development and general well-being:** Proactive measures to promote AI in pursuit of beneficial outcomes for people and the environment, such as augmenting human capabilities and skills, addressing the exclusion of certain populations, reducing gaps and social inequalities, protecting the environment, and streamlining planning and construction processes.
- 2. Respecting the rule of law, human rights and democratic values, including fairness and privacy:** The development, integration, and use of AI systems should align with democratic values and the rule of law, in a manner that respects human rights (including human dignity, equality, privacy and autonomy). Moreover, it is important to address phenomena impacted by AI, such as disinformation. To conform with these principles, the relevant actors must establish appropriate mechanisms, such as human involvement and risk management, in accordance with applicable law, the context and available technological capabilities.
- 3. Transparency and explainability:** This refers to transparency toward the public with regard to the organization's AI systems. Transparency may include providing relevant information including, for example, information on the system's functions, capabilities and limitations, disclosing when interactions with the AI system occur, providing information about the data sources used and the system's underlying logic, and providing information enabling affected parties to challenge negative outcomes.
- 4. System safety and security:** When developing and using AI systems, it is important to ensure their reliability, safety, and security throughout their lifecycle, so that they operate properly without posing an unreasonable safety or security threat, even in the event of unforeseen situations, misuse or adverse external conditions. Furthermore, mechanisms should be established to address risks caused by the system.
- 5. Accountability:** Organizations are expected to be accountable for the proper functioning of their AI systems and to comply with the above principles, in accordance with their roles and subject to available technological capabilities. To that end, they should develop risk management mechanisms and internal rules of conduct for addressing risks (including bias, violation of privacy or other human rights, violation of workers' rights, IP infringement, safety risks, etc.).

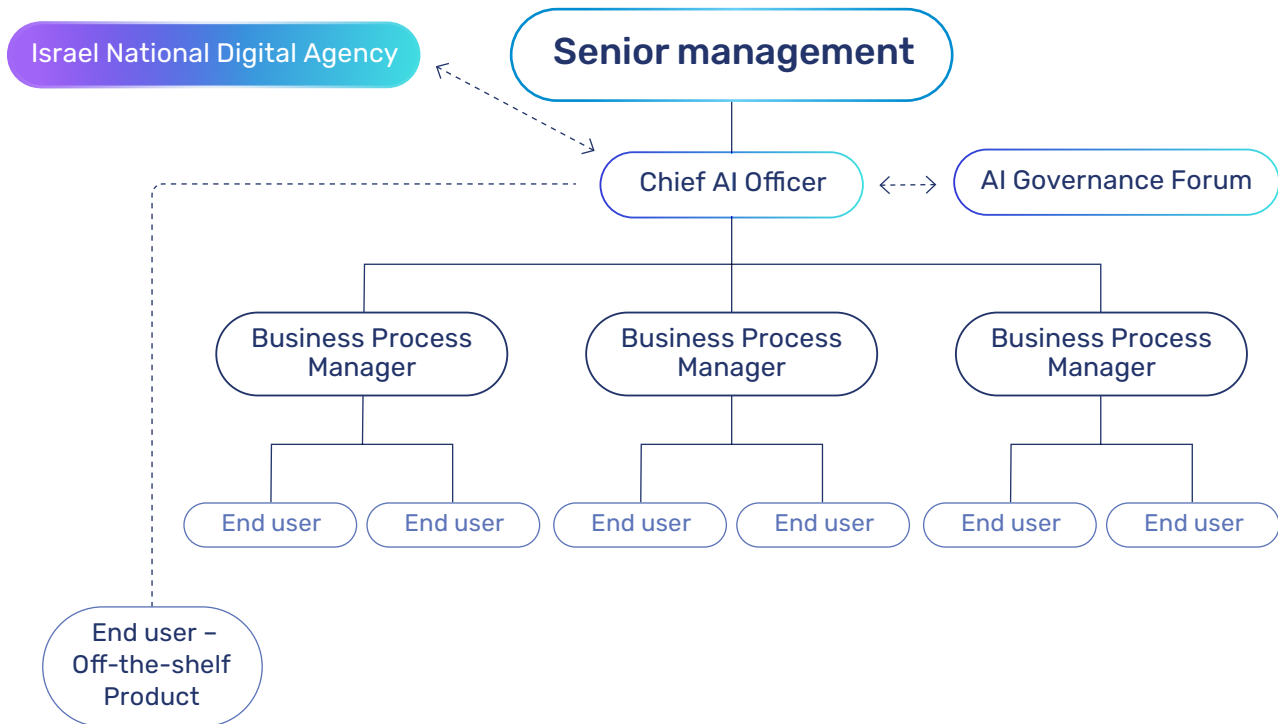
⁶ These descriptions are not a literal translation of the official OECD text.

These principles convey that the responsible use of AI is not merely a technical check-the-box exercise. Rather, it requires, fundamentally, an organizational culture that internalizes and embraces these values. This calls for building organization-wide awareness of AI's implications, promoting AI literacy, and fostering a genuine commitment to stewardship over AI system development and implementation processes. Responsible use of AI requires concerted actions, governance mechanisms, and clearly assigned roles and responsibilities across the organization. Moreover, these principles are not absolute or categorical. The organization should apply them in a manner that is appropriate, pragmatic and adapted to its operational context. For instance, transparency does not imply that an organization must disclose classified tools.

3

AI Governance

Key roles and responsibilities



3.1 Senior management

The organization's senior management initiates and sets the tone for responsible AI processes. To this end, it should provide the requisite operational and budgetary infrastructure. Therefore, it is recommended that management appoint a Chief AI Officer, establish an AI Governance Forum where appropriate, allocate the resources they need for carrying out the activities described in this guide, and report periodically on the organization's activities in this area. Senior management is also expected to develop an organizational AI strategy which prioritizes the areas in which AI should be integrated.

3.2 Chief AI Officer

The Chief AI Officer is responsible for strategic integration of AI systems into the organization's business processes. This includes formulating and implementing protocols for the responsible use of AI, including the organization's AI risk management policy. As a default, it is recommended to assign the chief data officer (CDO) to this role, but another official may be appointed, such as the CTO or the VP of strategy and policy. The role can also be assigned to a committee.⁷ What matters is that the person or entity tasked with this role be knowledgeable about AI, and be given the necessary internal resources (i.e. budget, staff, training) and mandates to perform their duties.

For more information and details on the role of Chief AI Officer, see [Annex A](#).

3.3 AI Governance Forum

In a large organization, or in one where AI systems are highly sensitive or where AI systems are used to handle particularly complex business processes, an AI Governance Forum should be established. This forum advises the Chief AI Officer on cross-organizational issues (e.g. AI integration strategy and policy, the effects of AI use on the organization's employees, and broad socioeconomic repercussions), specific high-risk projects, and AI incident response. The forum's composition can be decided by the organization's senior management. However, a lean forum of 3-4 members is recommended, with the ability to consult additional experts as needed. The AI Governance Forum is intended to support the Chief AI Officer, rather than creating an additional bureaucratic layer. The organization should ensure that the forum indeed operates as intended.

3.4 Business Process Manager

The Business Process Manager is the person responsible for a specific business process, who is considering integrating AI technology. For example, the Business Process Manager can be a division manager responsible for regulating a particular field, providing benefits, or issuing licenses and permits. They can also be responsible for a cross-organizational process, such as accounting, planning and budgeting, legal advice, and IT. When considering the use of an AI tool for a particular business process, the Business Process Manager is responsible for the AI risk management process for that project.

For more information and details on the role of the **Business Process Manager**, see [Annex B](#). For the **risk management process**, see [Annex C](#).

⁷ In an organization where the "Nimbus ministerial team for digital matters, data and AI" has been established, pursuant to Government Decision 3574 ("**Nimbus Team**"), the organization may wish to assign the role of Chief AI Officer to one of the members of the team, or to another official acting in coordination with the team.

3.5 End users

End users are public sector employees who use an AI system to carry out their duties. Whether the AI system in question is provided by the organization or an off-the-shelf tool available to the general public, end users are expected to act responsibly and prudently.

For guidelines for **end users**, see [Annex D](#).

3.6 Israel National Digital Agency

The Israel National Digital Agency (INDA) supports and guides public sector organizations in integrating responsible AI practices, chiefly via an [AI Governance Hub](#). It can help conduct AI risk management in a given project. The INDA also has a Center of Excellence that assists in the implementation of AI projects in the public sector. In addition, the INDA develops and provides AI literacy training courses through its digital literacy academy, [Hadigitalit](#). In the future, the INDA will publish use cases and examples of risk management programs. It intends to connect organizations with one another, as needed, to prevent duplication of efforts and enable knowledge sharing. The INDA can be consulted regarding AI risk management for certain high-risk uses, and can act as a source of knowledge for the Chief AI Officer.

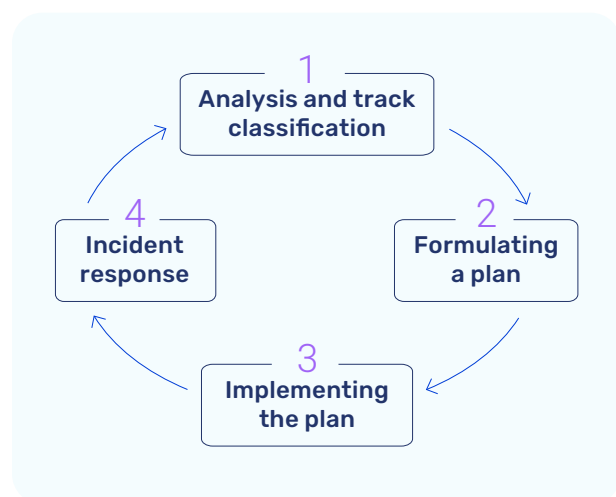
4

Risk management processes for AI systems

In general, the business process begins with setting an operational objective, defining the requisite specifications, planning a supporting information system, and selecting a suitable AI product. This process requires an understanding of the technology's capabilities and limitations, its benefits and risks, and the operational and regulatory environment in which the system is intended to operate. For more on this topic, see the [AI Journey](#) guide. The following is an overview of the risk management process for assessing an AI system.

The process comprises four stages, each with its own substages. The Business Process Manager is responsible for carrying out these stages and documenting each substage for internal audit purposes.

The process is cyclical and occurs throughout the system's entire lifecycle – its development, deployment, production, upgrading, and decommissioning. Where the AI component is being integrated into an existing operational system and business processes, the risk management process proposed here relates to this AI component only.



The four stages are summarized below:

Stage 1: Analysis and classification of the system – This stage involves classifying the system, which in turn prescribes how its development and deployment ought to be governed. A very low-risk system does not require the same level of oversight as a high-risk one. This guide recommends a differentiated approach, based on the traffic light method (details on the different tracks are provided in [Annex C](#)).

- ✦ The **Green Track** is expeditious and simple, and is meant for low-risk systems;
- ✦ The **Yellow Track** is a more comprehensive framework for moderate-risk systems;
- ✦ The **Red Track** has strict conditions, meant for high-risk systems;
- ✦ Additionally, it is recommended to establish a **Blue Track** that allows for limited experimentation with high and moderate-risk systems, to collect data on the system in a safe and controlled setting, before its full deployment.

For the initial risk classification, the Business Process Manager should consult with the relevant officials, based on the anticipated risk level. Together with the organization's Chief AI Officer, they determine which track will be used. At any given moment, the Business Process Manager may recommend changing the classification to a more lenient or stringent one, based on findings and consultations. At any time, the Business Process Manager can also suggest placing the project on the Blue Track.

Note: The classification stage should be coordinated with the development and procurement processes of the system being evaluated. Every tool development or procurement process is expected to include testing, checks, and dialogue with the developers or the supplier, to further refine the characteristics of the system and enhance its performance. During this process, the risk classification is not expected to be a single, clean-cut exercise – instead, it will involve several rounds of consultation. The substages described below are simplified to outline key milestones, but the process itself is dynamic.

Stage 2: Formulating a risk management plan – The goal of this stage is to formulate a risk management plan suited to the system’s risk classification. In the case of Red Track systems, it is recommended to consult with the AI Governance Forum. Note: As in Stage 1, the risk management plan may be formulated in several steps, with an exchange of ideas and positions, rather than in a linear fashion.

Stage 3: Implementing the plan – At this stage, the system is operational, and the approved risk mitigation measures must be implemented in practice. In appropriate cases, the Business Process Manager must ensure that the planned public transparency requirements are implemented as well. The Business Process Manager must also monitor the system to ensure that it is operating as expected, and report problems and AI incidents in accordance with the risk management plan.

Stage 4: Incident Response – This stage refers to the real-time response to issues and AI incidents, in accordance with the risk management plan, including reports to the organization’s senior management. Whatever the system classification, severe AI incidents require an immediate response. Public transparency considerations come into play here as well, and the Business Process Manager should consult with the organization’s AI Governance Forum and legal department. In relevant cases, it may be necessary to consider restricting or suspending the use of the system.

For more details, and an explanation of the risk management method and the various tracks, see [Annex C](#).

Annex A

Chief AI Officer Roles and responsibilities

Recommended actions	Details
1 - Policy development	
1.1 Responsible use	Determining a general policy for the responsible use of AI within the organization, with the support of the AI Governance Forum where necessary. It is recommended to base the internal policy, in full or in part, on the Responsible AI principles, adapted to the organization's operational context (Chapter 2 of the guide), and on the risk management process described in this guide (Annex C). It is important to align the policy with the organization's cyber and privacy policies, the Legal Guide, and the Regulatory Risk Management Guide issued by the Israel Regulatory Authority. Annex E serves as basis for an organization's policy. It can be customized to the organization's needs.
1.2 Data governance for AI	Developing rules for managing the organization's data used in AI models and systems, in coordination with the CDO (if he/she is not the Chief AI Officer). This ensures that the data used as the basis for AI applications is valid, accurate, and up-to-date. It enables the organization to anticipate issues and risks, including: (1) rights and interests of data subjects, particularly privacy rights, and (2) the need to train models on diverse, representative, and up-to-date data, to reduce the chances of algorithmic biases, incorrect outputs and harmful outputs. The Chief AI Officer should ensure the organization's data is protected to maintain continuity, and put measures in place to preserve institutional data, knowledge, and know-how when a Business Process Manager leaves the organization or changes roles.
1.3 Off-the-shelf AI tools and AI systems policy	Establishing a policy regarding authorized products and applications, in coordination with the organization's cybersecurity officer, DPO, chief accountant and other relevant officers. Among other things, this guide recommends regulating the permissions for internal and external users, determining whether to ban or otherwise restrict certain AI-based tools, under certain conditions, and communicating these policies within the organization. The use of government cloud-based (Nimbus) services, rather than external procurement, is recommended, so that this technology can be used more securely. Note: Procuring off-the-shelf products that incorporate AI services is subject to the relevant TAKAM directives.⁸
1.4 AI incident response policy	The policy should address, among other things, the following topics: establishing an AI Incident Response Team, discontinuing the use of an AI system, adjustments to algorithms, changing or updating the model or the data set used to train the model, reporting to the relevant individuals in the organization (e.g., management, DPO) and outside of the organization (e.g. Israel National Cyber Directorate, Privacy Protection Authority, Cyber Defense Unit in the Israel National Digital Agency).

⁸ [TAKAM directive 7.10.7](#) – Contracting for the purchase of AI services. For AI services offered by AWS and Google, see [TAKAM directive 16.12.2: "Provision of public cloud services to government ministries by AWS and Google"](#). With regard to AI services offered by Salesforce in a central tender (i.e. through the Government Procurement Administration), see [TAKAM directive 16.2.4: "Provision of Customer Relationship Management \(CRM\) services on the cloud"](#). For AI services offered through Nimbus by third-party vendors, see [TAKAM directive 16.12.3: "Procurement of third-party services through the AWS and Google public cloud."](#)

Recommended actions	Details
2 - Policy implementation	
2.1 Establishing mechanisms	Establishing the mechanisms developed pursuant to Section 1 above, and specifically: the AI Governance Forum, the AI Incident Response Team, communicating the procedures and templates for relevant authorizations, and sharing relevant information on the organization's website.
2.2 Spearheading AI literacy processes	In conjunction with the organization's HR department, establishing a training program that will include continuing professional development courses and internal and external training on AI and responsible use (including through the Israel National Digital Agency's Digital Academy).
3 - Operationalizing the policy throughout the organization	
3.1 AI risk management directives	The Chief AI Officer should participate in the development of the risk management plan for an AI system being considered, commensurate with its expected AI risk level, by providing guidance to the Business Process Manager on the responsible use of AI. The Chief AI Officer is also responsible for approving risk management plans for high-risk systems. See Annex C .
3.2 Organizational AI products database and risk map	Creating a list of AI systems being used at the organization, as well as those pending approval, along with an up-to-date status report on the mapping and scoring of risks. The list should indicate the Business Process Manager in charge of risk management for the entire system, and should include details regarding the monitoring conducted by the Business Process Manager and other relevant officials, such as actions performed to identify data leaks. The status report should indicate potential risks created or amplified by interactions between AI systems at the organization.
3.3 Transparency	Publishing general information on AI systems being used by the organization, pursuant to the organization's responsible use policy, risk management plans of each AI system (for details, see Annex C), and the Legal Guide. This is subject to applicable restrictions on classified or protected information, such as confidential methods and tools. The Israel National Digital Agency operates a government dashboard that publishes the government's AI use cases. ⁹ It is recommended to send the information on the system to the Israel National Digital Agency so that it can be publicized on the dashboard.
3.4 AI Incident response	Pursuant to the procedures established under Section 1.4 .
3.5 AI incident reviews and lessons learned	Proactively and comprehensively investigating AI incidents, including reviewing decisions tied to the specification and operation of the system and the method that was employed to manage the risks. Lessons learned: Based on the findings of the investigation, and in consultation with the AI Governance Forum and senior management, the organization should (1) decide whether to modify the way in which the system operates or the way risks are managed, (2) for severe incidents, determine whether to discontinue usage of the system, either temporarily or permanently, and (3) in cases where substantial or serious damage was caused, determine whether and how to notify the public and those affected by the incident.
3.6 Internal reporting	An annual report, with documentation, should be submitted to the Director General and to the AI Governance Forum (as relevant) regarding the AI-enabled systems being used by the organization, their benefits and risks, the relevant control and risk management processes that were adopted, and any AI incidents that occurred. Recommendations should be formulated on improving organizational processes for responsible AI use.

⁹ AI systems will also be published on the [AI Watch](#) system, established by the Israel National Digital Agency, which is meant to serve as a general database and create situational awareness on the use of AI at the various organizations for government actors and residents. This website will be updated.

Business Process Manager Roles and responsibilities

Recommended actions	Details
1. Specifying how AI components will operate in the project	During the appropriate stages of the process, and as needed: • Involving the relevant actors within the organization, such as the DPO, CDO, CISO, Head of Digital and Information Technologies, Cloud Lead, as well as users and customers. • Based on the nature of the system and possible legal aspects, the legal department should be consulted in order to identify the legal requirements with which the system will need to comply. This should be done as early as possible, to anticipate potential legal concerns. • Consulting with colleagues at other public agencies facing an identical or similar business need, to learn from their experience with comparable risks and benefits. • Documenting the findings and conclusions from the consultations. For more on developing project's specifications, see AI Journey.
2. AI Risk Management	Implementing the risk management process described in Annex C (risk classification, risk management plan, audits and reporting to the Chief AI Officer). Documenting all findings comprehensively and with precision, clearly mapping the risks and benefits of the system and formulating a risk mitigation plan.
3. Directives for end users	The relevant directives on the responsible use of AI, including third-party vendor instructions, if any, should be made available to end users. If these directives cannot be provided, the information contained in the End User Guidelines (Annex D) should be provided as baseline guidance, and complemented with additional guidance tailored to the product's specification and to conclusions drawn from the risk management process (Annex C).

Risk management method for AI systems

This annex is designed to provide a uniform basis for managing AI risks with respect to a specific system being considered for use. It does not address regulatory or policy risk management. If such questions arise, the [Regulation Risk Management Guide](#), published by the Israel Regulatory Authority, should be consulted.

The method proposed here is based on the traffic light approach, with three main tracks, based on the risk level of the evaluated system (green, yellow and red), as well as the experimentation track (blue).

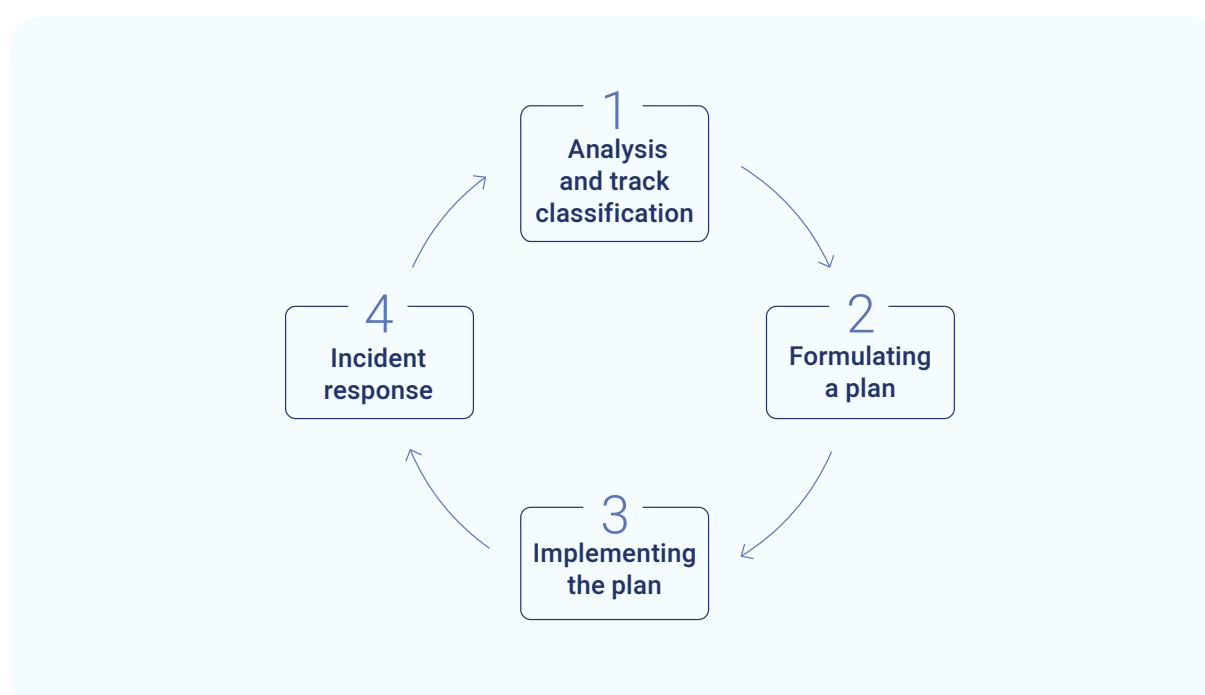
The risk management method is comprised of four parts:

- [1. Stages of the system risk management process](#)
- [2. Classification of benefits and risks](#)
- [3. Description of approval tracks](#)
- [4. Formulating a risk management plan](#)

The organization's Chief AI Officer may, at his/her discretion, expand or modify the following model (though it is recommended that all of its basic components be retained in some form).

1. Stages of the system risk management process

The process of identifying and managing risks is iterative. It begins as early as the system specification stage. It is routinely conducted throughout the lifetime of the system, as described in the following diagram:



The following table describes the actions that each stage involves. Decisions made at each stage must be documented.

Stage description	Action		
Analyzing and classifying the system	The goal: To holistically evaluate the desired AI system’s benefits and risks and decide which risk management track is to be taken (green/yellow/red). When experimentation is desirable before reaching a decision on which risk management track is to be used, the blue track can be taken.(see details below)		
Initial classification	✦ Identifying the business need, examining alternatives, and determining the relative advantage that the AI system is expected to deliver. ✦ Performing an initial mapping of the potential risks and benefits arising from the system being assessed.		
	<table><tr><td>Assessing the benefits, including an assessment of: ✦ The impact of potential benefits ✦ The likelihood of materialization </td><td>Assessing the risks, including an assessment of: ✦ The impact of the risks ✦ The likelihood of materialization </td></tr></table>	Assessing the benefits, including an assessment of: ✦ The impact of potential benefits ✦ The likelihood of materialization	Assessing the risks, including an assessment of: ✦ The impact of the risks ✦ The likelihood of materialization
	Assessing the benefits, including an assessment of: ✦ The impact of potential benefits ✦ The likelihood of materialization	Assessing the risks, including an assessment of: ✦ The impact of the risks ✦ The likelihood of materialization	
See the table in Part 2 of this annex.			
Consultations	Consultations should be conducted based on the anticipated classification. ✦ Green & Yellow Tracks: Consult with the Chief AI Officer. Other relevant stakeholders may be consulted as well, as necessary (see the following list). ✦ The Red track: Consult with the Chief AI Officer and other relevant individuals both inside and outside of the organization, as necessary. The following are the relevant actors within and outside of the organization that may be consulted at each stage (this is a non-binding, yet non-exhaustive list): ✦ Privacy issues: DPO (the organization’s data protection officer) ✦ Cybersecurity and impacts on the network, interfaces, communications and information: the organization’s cybersecurity officer; ✦ Use of the cloud: the organization’s cloud management lead; ✦ Implementing technical components to mitigate the risks: the organization’s Digital Technologies and Information director. ✦ Data quality and accessibility: CDO (the organization’s data officer) ✦ Unit employees within the organization; ✦ The organization’s partners, including its suppliers; ✦ External business entities expected to use the system; ✦ Those receiving the service within the organization, including regulatees; ✦ The general public. Additionally, in the various tracks, and based on the nature of the system and emerging legal aspects, the legal department should be consulted in order to map the legal requirements with which the system will need to comply.		
Classification of benefits and risks	✦ Comprehensive evaluation of expected benefits: very low/low/moderate/high/very high benefit ✦ Comprehensive evaluation of expected risks: very low/low/moderate/high/very high risk ✦ Considering whether to move to the Blue Track, based on its parameters. As a general rule, if the risks are not fully understood, or if it is difficult to measure them with sufficient accuracy before the system is deployed, it is recommended to move the system to the Blue Track. See more details here. ✦ Based on the assessment, the system’s track is preliminarily determined. ✦ Note: In general, if the system presents one “very high” risk, it will be classified as a red-track system. See the table in Part 3 of this annex, which describes the tracks.		
Track validation	After consulting with the relevant entities and performing a more thorough risk and benefit assessment, the track used to evaluate the system can be approved: ✦ Green/Yellow Track: the Business Process Manager approves the track independently, while consulting with the Chief AI Officer and subject to his/her professional guidance ✦ Red Track: Approval by the Chief AI Officer, after notifying the Director General and AI Governance Forum (if established). If it is unclear whether the system warrants classification under the red track, the Chief AI Officer may consult the AI Governance Forum. If a system is classified as a Red Track system, the organization should consider moving it to the Blue Track, to enable a controlled and gradual initial deployment.		

Stage Description	Action
Stage 2: Formulating a risk management plan	Goal: Assessing the measures that may be employed to reduce risks.
Consultation	Consultation with the entities listed in Stage 1: ✦ Green Track: as necessary ✦ Yellow Track: the Chief AI Officer and organizational actors such as data security officer and the DPO. ✦ Red track: The Chief AI Officer, relevant actors within and outside the organization, and the AI Governance Forum (if established)
Formulating a risk management plan	Formulating a detailed risk management plan that integrates the risk mitigation measures specified in Part 4 of this annex, as necessary in the circumstances. Assessment of the costs of the risk management plan includes: ✦ An assessment of the various alternatives; ✦ An approximation of the cost of risk management measures needed to reduce the presented risk. ✦ If the system is classified as a Blue Track system, the project's operating parameters should be addressed (duration, metrics and criteria for progression). Note: "Costs" are not merely financial. Consideration should be given to broader repercussions, including efficiency and public trust. With regard to arising legal issues, the legal department should be involved at this stage to ensure that the appropriate measures are taken to comply with legal requirements, as set out in the Legal Guide (e.g., the duty to give reasons, transparency, handling biases, etc.).
Approving the plan	Green & Yellow Tracks: The Business Process Manager approves the plan independently, but in consultation with the Chief AI Officer and in accordance with his/her professional guidance. Red Track: The plan must be approved by the Chief AI Officer and the Director General or someone acting on his/her behalf. As mentioned above, in the event of any legal issues, the legal department should be involved as well.
Stage 3: Implementing the plan	Goal: Deploying the system and implementing the risk management plan.
Implementation	Implementing risk management measures, pursuant to the approved plan.
Recommended audits	Collecting data from the system itself; Implementing the risk reduction measures determined in Stage 2; Taking into account information from affected stakeholders, based on information received from public inquiries, media channels, dialogue with relevant civil society organizations, etc.; Data audits: Spot checks or periodic sampling, as a function of developments, needs and risks.
Audit frequency	Based on system classification, unless otherwise stated in the risk management plan: ✦ Green: Annual ✦ Yellow: At least twice a year ✦ Red: At least once every quarter
Public transparency	Based on the system's classification and the affected population, and in accordance with the risk management plan: ✦ Continuous transparency regarding the organization's use of AI systems, the types of data the system uses, the way the system works, in a general and accessible manner (including capabilities and limitations). ✦ Transparency and enhancing awareness of cases in which a direct or substantial interaction exists between citizens and the AI system. Additional details on transparency will be provided in the Legal Guide.

Stage Description	Action
Stage 4: AI Incident Response	Goal: To enable optimal handling of AI incidents and failures.
Periodic reporting – frequency	Based on system classification, unless otherwise stated in the risk management plan: ✦ Green: Annual ✦ Yellow: At least once a year ✦ Red: At least four times a year
Reporting AI incidents	Immediately reporting to the Chief AI Officer when an AI incident occurs. Note: The report ought to include recommendations on the required adjustments.
Public transparency	Transparency towards the general public (or the affected population) when an AI incident occurs, in accordance with the Chief AI Officer's guidance.
Adjustments and lessons learned	Making adjustments in accordance with the Chief AI Officer's guidance. Adjustments may include, inter alia, changes to the product's specifications or the method used to gather or tag data. Documenting as part of the change management process

Blue Track – Risk management process

All stages of the risk management process should reflect the experimental nature of the Blue Track, while routinely consulting with the Chief AI Officer. Naturally, the Blue Track provides some flexibility regarding the system's planned operation. Still, the process should, at a minimum, address the following components, in consideration of the extent of the risks:

- ✦ Consulting with the relevant entities within and outside of the organization, including the legal department;
- ✦ Formulating a risk management plan;
- ✦ Setting the parameters and metrics that will be evaluated in the project;
- ✦ Testing frequency;
- ✦ Frequency of reporting to the Chief AI Officer;
- ✦ Setting timetables, taking into account the duration of the project's expected pilot phase;
- ✦ Defining severe AI incidents and how they are to be handled, when very high risks are posed.

Changes may be made dynamically during the above stages (e.g., deciding to extend the period of time that was originally defined).

Note: If a system poses high risks, the Business Process Manager and/or Chief AI Officer are also expected to notify the organization's management at various junctures, and, if needed, receive its approval.

2. Classification of benefits and risks

The starting point of the process is assessing the AI system's various benefits. The use of AI may lead to valuable results, such as an improvement in public service efficiency (through the automation of various processes, leading to shorter wait times or file processing times); better decisions; AI systems operating around-the-clock; prevention of human error; proactivity when designing and delivering public services; and adapting public services to address the unique needs of residents and businesses.¹⁰ Generally, when AI systems are used responsibly to advance the organization's goals, the public gains from better public goods and services. Consequently, this may enhance the organization's reputation and public trust in the organization.

However, AI systems are based on complex internal processes, advanced technological systems, and human interactions that also create various risks. Inaccurate, outdated or biased data could harm the quality of the system and create errors or distortions in the results. They can also make it harder to follow the system's logic.¹¹ Furthermore, the use of AI systems exposes the organization to unique data security risks that could harm privacy and data integrity, and lead to unintended disclosure of [protected information](#). This damage could also be caused by attacks on the tools themselves – hostile actors may try to exploit vulnerabilities in AI systems and tamper with results or try to extract information from the data management systems. Ultimately, if these risks materialize, there is risk of harm to the public, thus undermining public trust in the organization.

How can an AI system's risks and benefits be assessed? The following is a non-exhaustive map of the main types of risks and benefits that should be considered during a risk management process. This mapping corresponds with the OECD's Principles on the responsible use of artificial intelligence. Additional risks and benefits relevant to the organization's activities, or the tool's expected use cases, can also be examined.

Benefits

Category – based on the OECD principle	Sample benefits	Comments/examples
Social, environmental and economic benefits OECD principle: AI for growth, sustainable development and general well-being	✦ Reducing social gaps ✦ Protecting the environment ✦ Improving public health ✦ Economic growth ✦ Improving a work environment ✦ Making decision-making processes better and more accurate	✦ A tool to improve the accuracy of climate forecasting ✦ Streamlining traffic management (the contribution to the environment) ✦ A tool for analyzing financial requests or issuing licenses ✦ Reducing workloads
Advancing individual well-being OECD Principle: respecting the rule of law, human rights and democratic values, including fairness and privacy	✦ Adapting and streamlining citizen services ✦ Reducing wait times ✦ Reducing human error	Examples: ✦ Tools to facilitate citizens' access to their rights ✦ A tool to help with the anonymization (redaction) of files containing personal information, before they are sent to a third party ✦ A tool for analyzing government or court decisions If legal issues are present, including the assessment of systems that affect human rights, the organization's legal department must be consulted separately. See also the Legal Guide.

¹⁰ One such example is [a call for proposals issued by the Israel National Digital Agency and the Ministry of Innovation, Science and Technology](#) that supports public-sector AI use cases with high potential value. Examples from other countries can be found on the [OECD catalogue](#).

¹¹ Due to the "black box" (opaque) nature of AI systems.

Category – based on the OECD principle	Sample benefits	Comments/examples
Benefits to public transparency OECD principle: transparency and explainability	✦ Summarizing and analyzing government decisions and other public texts for auditing purposes ✦ Improving information sharing processes with the public	Examples: ✦ Public engagement tool for policy and regulation development ✦ Government chatbot to enhance citizen's access to information ✦ Translation of government websites into different languages Transparency and explainability have legal implications - the Legal Department must be consulted
Public security The OECD Principle: safety and security	✦ Strengthening the cybersecurity of critical systems ✦ Enhancing security in public spaces (highways, public transportation, policing)	Examples: ✦ Proactively protecting critical systems ✦ Monitoring the cyberspace of a governmental network ✦ Forecasting natural disasters ✦ Rapidly and effectively responding to emergencies.
Governance OECD Principle: accountability	✦ Tracking internal processes and decision-making processes within the organization ✦ Implementing quality controls on intra-organizational processes	Examples: ✦ A tool to help summarize internal meetings ✦ A permission management dashboard ✦ A tool to help analyze public comments

Risks

Category – based on the OECD principle	Sample risks	Comments/examples
Social, environmental and economic risks OECD principle: AI for growth, sustainable development and general well-being	✦ Widening of social gaps ✦ Harm to the environment ✦ Harm to public health ✦ Economic damage	Harm to the environment: Errors in a system designed to help issue air emission permits and environmental conditions for business permits. Harm to public health: Errors in a system that helps with the process of issuing regulatory approval for drugs Economic damage: Errors in an AI model for forecasting and overseeing the securities market might undermine financial stability in the country and in various companies and establishments.
Risks tied to protecting human rights and democratic values OECD Principle: respecting the rule of law, human rights and democratic values, including fairness and privacy	✦ Possible discrimination and bias ✦ Possible violation of privacy ✦ Possible violation of human dignity or other fundamental rights	AI systems may raise legal challenges as well. For example, biases and discrimination emanating from these systems raise concerns over the ability to safeguard the principle of equality. When integrating AI systems into a public agency's activities, legal compliance should be ensured, including the protection of individual rights, the protection of the right to privacy, and compliance with administrative law obligations, such as a public body's obligation to act within the scope of its authority, the duty to give reasons, transparency requirements, and reliance on sound factual foundations. To do so, the legal department should be consulted. The Legal Guide will address these topics.

Category – based on the OECD principle	Sample risks	Comments/examples
Risks to public transparency OECD principle: transparency and explainability	✦ Results lacking a clear or consistent rationale ✦ The inability to understand how the system had reached a specific conclusion ✦ Use of the system outside its intended purpose	The black box phenomenon makes it difficult or impossible to understand an AI system's output. Moreover, blindly relying on an AI system's output may pose business risks. In addition, transparency and explainability have legal ramifications; thus, compliance with those precepts needs to be assessed.
Risks to public safety OECD Principle: safety and security	✦ Operational risk ✦ Damage to critical systems, including cyberattacks ✦ Loss of life or property damage ✦ Social manipulation and disruption of public order	Operational risk: A risk arising from a system malfunction, or system misuse. This risk may arise from inaccurate, outdated or biased data, <u>model drift</u>, etc. Damage to critical systems: A risk arising from reliance and dependence on AI for activities such as managing electrical grids, water grids, or traffic control. Loss of life: A risk of physical or emotional harm to human beings, for example, if a traffic accident occurs, or in the case of a misdiagnosis. Cyber risks: The organization's CISO should be consulted in order to determine the scope of the risk and the ways to contend with it.
Governance risks OECD principle: accountability	✦ Overreliance on AI systems, could lead to an erosion of internal accountability.	Overreliance on AI systems can lead to a lack of accountability. Literacy and audit mechanisms help address this risk.

This guide focuses on identifying and analyzing risks that directly arise from **AI systems** as such, and not on risks arising from other factors, such as cybersecurity vulnerabilities in the organization's IT systems, or flaws in the underlying policy which the AI tool implements.

Legal issues: AI systems occasionally give rise to legal questions on issues such as applicable administrative law (the public organization's obligation to act within its authority, the duty to give reasons), privacy, biases and discrimination, and the protection of intellectual property.

Therefore, based on the nature of the system and potential legal questions, the legal department should be consulted in order to map applicable legal requirements.

Note that the legal assessment is separate and independent. Nonetheless, legal requirements and issues should be flagged early on and duly considered when developing the risk management plan as well as when the system is deployed.

The following measures must be taken to evaluate the AI system’s benefits and risks:

- ✦ Mapping the expected benefits with precision and specificity, including an estimation of the potential ROI;
- ✦ Rating the benefits in terms of the relative advantage they provide (from “very low” to “very high”);
- ✦ Mapping and rating the various risks – including their potential of materializing and the anticipated scope of damage (from “very low” to “very high”);
- ✦ Holistically weighing the benefits and risks. As mentioned earlier, when risks and benefits involve the legal issues specified in the table, such as those tied to **rights**, these issues cannot be weighed and analyzed solely in terms of their cost and benefit. Therefore, the legal issues should be reviewed with the legal department.

Integrating the AI system into the process might reduce existing risks, amplify them, or create new risks. Therefore, the incremental risk should be properly assessed and measured against the baseline state, by isolating the AI-specific risk driver, and comparing it to the expected benefits. For example, a reduction of the system's error rate from 15% by human to 1% by AI indicates that there remains a risk, but in comparison to the baseline, the AI system represents a major improvement. Each alternative should be evaluated with respect to the other alternatives and the corresponding changes to risk levels, including the option of integrating non-AI tools (such as business rule engines).

After reviewing the parameters illustrated above, **a separate numerical severity and probability score should be given to each risk** (1 being the lowest score, and 5 being the highest). The overall risk score for each risk is the product of the risk’s severity score (1-5) and its probability score (1-5). Note that these estimates are qualitative, not quantitative. The recommended scoring method is designed to provide a general overview and prompt the Business Process Manager to think about the risks systematically.

When the risk identification and assessment process is complete, a “heat map” should be drawn, to prioritize risk mitigation resources. For example, suppose that three risks were identified in a system under evaluation: a data security risk, a financial risk, and an operational risk. The severity of the operational risk is determined to be low (level 1), but it has a high probability of occurrence (level 4). Conversely, the severity of the data security risk is determined to be high, but its probability of occurrence is low. The financial risk is deemed moderate (level 3), but its probability of occurrence is high (level 4).

Visualizing the various risk levels helps the Business Process Manager address the various risks adequately. Moreover, the Chief AI Officer can use heat map tables received from all Business Process Managers in the organization to create an organizational heat map, from which practical insights can be drawn.

Sample heat map



3. Description of approval tracks

Risk management is a dynamic, multi-stage and differentiated process. The higher the risk, the more comprehensive the authorizations and audit measures need to be. The following are four recommended tracks for approving systems and performing risk management: green, yellow and red tracks, using the familiar traffic-light method, and a blue track, which allows for controlled experimentation.

The system classification may vary according to new findings, changing circumstances of use cases, technological developments, inputs from the operation of the system and changes in the system's impact level. To reclassify a system, the risk management plan must be reviewed and adjusted to the required audits and reports, which may lead to either stricter or more lenient requirements.

The following is a breakdown of the approval tracks, based on the system's risk level.

A table appears at the end of this annex, which displays the approvals required in each track.

1. Green Track

The Green Track is meant for low-risk systems. It enables AI that poses no significant challenges to be authorized through a simpler, faster process. Notably, even when a system is classified as "green", a risk management plan must be prepared (though the plan may be fairly rudimentary, given its risk level). Continuous audits, AI incident response and reporting must still occur, and transparency must be constantly maintained in the appropriate cases. However, the plan is expected to be less extensive and require relatively fewer resources from the organization.

1.1 When is this track appropriate?

When each of the system risks is classified as low-risk (e.g., a score of 1-4 out of 25 in each of the risk categories listed in [Part 2](#) of this annex), even if the expected benefit is considered low. In general, systems can be considered for the Green Track in the following cases (these are indicative, and not necessarily cumulative or sufficient conditions):

- ✦ The system does not violate the principles of democracy and the rule of law, and it complies with legal requirements;
- ✦ The system does not have any direct effect on rights, freedoms or eligibility for services;
- ✦ The system does not carry out complex multi-use actions – it operates in a single, circumscribed area.
- ✦ The system does not use personal information or other protected information (due to factors like confidentiality) (see the definition of [Protected information](#));
- ✦ If it has a generative component, the system's [Temperature](#) is very low;
- ✦ The system can be subjected to robust human oversight, and this is how it is intended to be used.
- ✦ The system's output is simple and explainable.

Examples (solely for illustrative purposes - a risk assessment should be conducted in all cases):

- | | |
|---|---|
| <ul style="list-style-type: none">✦ Office equipment inventory management✦ Dashboard for internal projects✦ Presentation generator tool✦ Summarizing minutes of meetings✦ Providing standardized answers to common questions (FAQ)✦ Tools that help analyze public documents (such as NotebookLM)✦ Spellcheckers for public announcements | <ul style="list-style-type: none">✦ Trend analysis and data analysis tools, and tools for recognizing patterns and detecting anomalies, in non-sensitive contexts (e.g., cases that are not expected to affect individual rights)✦ Optical character recognition (OCR) for document scanning, in non-sensitive contexts✦ Translating simple and non-sensitive announcements✦ Generating internal draft reports, in non-sensitive contexts✦ Translating public documents into Hebrew |
|---|---|

1.2 Authorization and oversight

Fast-tracked process: The Business Process Manager carries out the planned activities independently. The Business Process Manager must consult with the Chief AI Officer, and should also consult with any relevant entity, as necessary, to save time and avoid future surprises. If the Business Process Manager believes that the risks are very low, he/she can continue independently. A Green Track system also requires a risk management plan (prepared in consultation with the Chief AI Officer), as well as reporting and audits, though these audits need only be implemented once a year.

2. Yellow Track

The Yellow Track is relevant to systems whose overall risk level is determined to be moderate. In general, it will include risks that may be deemed material for the organization, its partners and its customers. Importantly, classifying a risk as “material” does not automatically preclude the use of the system. Various useful and beneficial systems can be classified as entailing a “moderate” risk, while being worthy of use and even providing significant value to the organization or to the public. However, organizational resources must be invested to address this type of risk.

2.1 When is this track appropriate?

When the system is not classified as a Green Track system and none of its risk scores exceeds 19 – that is, at least one of its risk scores ranges from 5 to 19 (out of 25). In short, its risks taken as a whole are “moderate”. Possible examples: (1) systems that analyze public inquiries (e.g., email messages, forms) and classify cases as urgent, or perform the initial routing of requests to the various departments; (2) anomaly detection algorithms that indicate potential errors in a report or a fraud attempt.

2.2 Authorization and oversight

After preliminary classification by the Business Process Manager, he/she must consult with the Chief AI Officer and other relevant actors in the organization. The Business Process Manager then drafts a risk management plan, in consultation with the Chief AI Officer. Frequent audits and reports are required.

3. Red Track

The Red Track is relevant to systems whose risk level is determined to be high. In general, it is suitable for cases in which the anticipated benefit is **very high and no less risky alternatives have been found** that provide an identical or similar benefit. The Red Track allows the system to be used under strict conditions, subject to additional authorizations from the organization’s senior management.

3.1 When is this track appropriate?

When at least one of the evaluated risks is classified as high (with a score of 20-25)

3.2 Authorization and oversight

Given the risks involved, the authorization process includes comprehensive controls and authorizations from senior management at key stages. Comprehensive consultations are required at the outset, once the system has been preliminarily classified as high-risk. Validation of this track requires the Chief AI Officer’s authorization, and the organization’s Director General must be notified. The risk management plan also requires the approval of the Chief AI Officer and the organization’s Director General, or someone acting on his/her behalf. If the organization has an AI Governance Forum, it should be involved as well. If the project is ultimately approved, it is advisable to route it to the Blue Track, to allow for controlled and gradual deployment, unless there is an emergency situation or if there are other exceptional justifications.

Exceptional risks and use cases that should be declined

When using the Red Track for high-risk systems, cases may arise in which the AI tool poses an exceptional risk of harm or danger, or one that could substantially and materially violate basic rights.

In such cases, consideration should be given to whether the project should proceed at all. Whenever a system could pose an exceptional risk, or whenever rights may be violated in a substantial or material manner, **the organization's legal department should be consulted** to assess whether there is a substantial legal hindrance or obstacle that precludes the use of the system. If the project receives legal approval, **the organization's senior domain-specific policymakers must be consulted**.

4. Blue Track

The Blue Track is an AI risk management framework designed for controlled experimentation, in cases where, due to insufficient data, it is not yet possible to classify the system's risk or to develop an accurate risk management plan. Under this framework, AI tools can be evaluated before they are conclusively classified in a controlled setting, within a predetermined time period. The Blue Track could be implemented in several ways. For example:

- ✦ Training a model based on [Protected information](#), or entering an input that contains protected information, in a sterile environment, to develop the product (the PoC stage);
- ✦ Partial or full deployment in the production environment, using specific product functions gradually and in a limited way.

4.1 When is this track appropriate?

A system or project may be placed on the Blue Track when there is uncertainty regarding the scope of the risks, when the system's abilities need to be assessed in a sterile environment, or when developers need to learn how it operates in practice and understand the risks it poses. This allows the organization to optimally reduce expected risks. Examples of tests that can be performed during the evaluation period: performance testing based on various parameters, verification of output accuracy in different scenarios, risk evaluation regarding sensitive topics (such as fairness), and assessing the system's interaction with end users and the affected population.

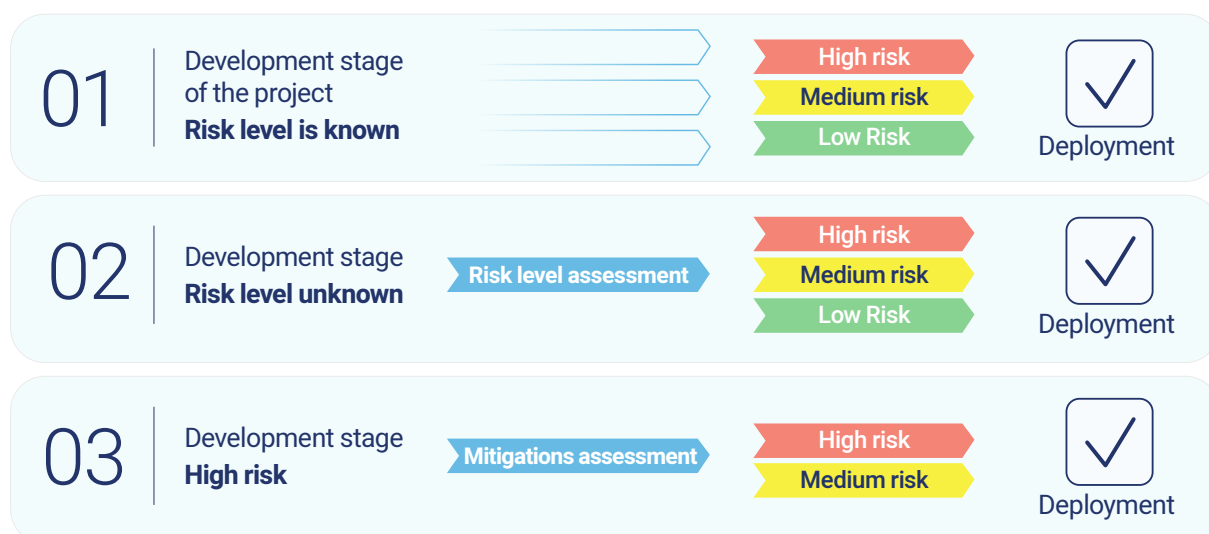
4.2 Authorization and oversight – stages

Since the Blue Track is designed to assess systems under conditions of uncertainty, in a variety of scenarios, the risk management process is more flexible. This is meant to allow the Business Process Manager to create a tailored risk management plan, together with the Chief AI Officer. However, Blue Track projects should be time-limited and subject to robust audit mechanisms. Once this period is over, the Business Process Manager must submit a comprehensive report to the Chief AI Officer. The Business Process Manager can then decide, after consulting with the Chief AI Officer, whether the project should be fully or partially deployed in one of the other tracks (green/yellow/red), based on the findings, identified risks, possible mitigation measures, and recommendations made by other actors.

This track supplants the internal authorization processes proposed in this guide (Yellow/Red Track, as applicable). However, in any event, the project requires assistance from the organization's legal department to ensure compliance with applicable legal obligations, including those pertaining to its authority to train models using [Protected information](#).

At any stage, the Israel National Digital Agency may be consulted. It can be contacted via e-mail at ResponsibleAI@digital.gov.il.

Track flow based on risk classification



5. Project authorization stages by track

The following is a summary table of the stages of the risk management process for each track. Note that the stages are provided merely as recommendations; the organization may modify them based on its needs and processes.

In addition to the relevant actors involved in the risk management process, as specified in the following table, the legal department must be involved as early as possible, whenever a legal question arises.

Stage Description	Green Track	Yellow Track	Red Track
Stage 1: Analyzing and classifying the system			
Consultation	Chief AI Officer and additional AI actors, as necessary	Chief AI Officer and additional AI actors, as necessary	Chief AI Officer and additional AI actors, within and outside of the organization
Track validation	Business Process Manager, in consultation with the Chief AI Officer	Business Process Manager, in consultation with the Chief AI Officer	Chief AI Officer & notification of the organization's Director General [as well as the AI Governance Forum, if established]
Stage 2: Formulating a risk management plan			
Consultation	As necessary	Chief AI Officer and inter-organizational actors such as the data security officer and the DPO	Chief AI Officer and inter-organizational actors such as the data security officer and the DPO, and entities outside of the organization [and the AI Governance Forum, if established]
Approving the plan	Business Process Manager, in consultation with the Chief AI Officer	Business Process Manager, in consultation with the Chief AI Officer	Chief AI Officer & the Director General, or someone acting on his/her behalf
Stage 3: Implementing the plan			
Audit frequency	Once a year	At least twice a year	At least once every quarter
Stage 4: Incident Response			
Routine reporting – frequency	Once a year	At least once a year	At least once every quarter
Blue Track: Milestones for these components will be determined in conjunction with the Chief AI Officer.			

4. Establishing a risk management plan

In general, there are four risk management methods: avoidance, non-involvement, risk transfer and mitigation. When a public entity assesses the use of an AI system, these methods may be relevant in various contexts. However, they are not necessarily suitable for every situation. The following is a brief demonstration of how these methods can be used in different contexts.¹²

Risk management method	Sample implementation
Risk avoidance	When risks are high and the expected benefit does not justify taking them. This may be particularly relevant for systems classified as “red”, with exceptionally high risks.
Non-involvement	When the risks are very low and are not sufficiently material for the organization or for the individuals or third parties associated with the organization. This method may be more appropriate for systems classified as “green”. Nonetheless, a risk management process should be conducted, and risk reduction measures should be considered, if appropriate.
Transferring and sharing the risk	Risk transfer or sharing the risk with a third party (through an indemnification mechanism, for example) could be suitable for low-level financial or operational risks (Yellow or Green Track risks), but are not appropriate for all risks. For instance, different types of operational risks can be transferred by imposing certain service requirements on the third-party service provider. Indemnification requirements can also be imposed, in case the risk materializes. However, under applicable law, a public authority retains material responsibility for its actions with respect to individuals and the public, and may not simply transfer it over to a third party, except in a limited way on the financial-operational level.
Risk mitigation	When taking certain material risks is justifiable under the circumstances, mitigation measures are available, routine monitoring may be performed, and the system can be operated safely and responsibly. See further details below on the types of possible actions.

The measures described below are non-specific, and may be applied to a variety of risk types presented by the use of AI systems. When choosing the appropriate measures, different measures may suit different goals. For example, human involvement may be suitable for checking grant requests that were denied, but it may not be suitable for a queue management system designed to work quickly and autonomously.

Analyzing the methods used to optimally reduce the risk is another important topic: sometimes, relatively simple actions can reduce risks more efficiently than complex procedures. For example, in an AI system that flags errors in documents submitted to a government department, a “standard” method to reduce AI risk is to add a human in the loop. However, this type of measure could entail significantly higher costs. Another method is retraining the model or adding a [RAG](#) component based on updated business data. Moreover, if the system is an agent that connects to a database, various protocols – such as [MCP](#) or [A2A](#), depending on the situation – can be used to enter data into the system, thereby calibrating for accuracy and implementing automated controls (e.g., referring a specific output to human oversight). Furthermore, in many cases, these technologies include full log capabilities. This can assist with explainability, and eventually, with transparency as well. Nonetheless, connecting an AI model to databases, both for training and inference, can lead to complex dilemmas regarding privacy protection and data security that require consulting with relevant actors within the organization and obtaining their authorization.

Risk mitigation efforts generally fall under one of the following main categories: technical and architectural safeguards, adding components to the AI product, and organizational/business measures. The following are select examples of measures that can be implemented in each category:

¹² See also page 18 of the Israel Regulatory Authority’s [Regulation Risk Management Guide](#).

A. Technical and architectural safeguards

To reduce risk, **technical and architectural safeguards** should be built into the system's design and architecture from the outset. The following techniques can mitigate various risks in different ways. The organization should familiarize itself with their advantages and disadvantages, and consider how they can be integrated optimally, in consultation with the organization's IT Division and the Chief AI Officer.

Risk type	Technical and architectural safeguards
Violation of privacy or the exposure of other Protected information	Privacy enhancement technologies (PET) ¹³ Federated learning Data anonymization and synthetic data Homomorphic encryption Differential privacy
Inaccuracies and hallucinations	Tool use Temperature adjustment MCP (in the case of AI agents) CAG/RAG Monitoring tools for concept drift/model drift risks MLOps (machine learning operations)
Biases and inequality	A debiasing tool (reweighting, adversarial debiasing) Fairness testing Bias evaluation frameworks
Data security and adversarial attacks	Guardrails Model Armor On-prem model A secure infrastructure and platform, such as Vertex AI, for isolating data and network security protection Purchasing a system that provides protection to organizations, or privacy protection to users, while complying with privacy and data security standards
Lack of transparency and explainability	xAI technologies The use of models that integrate reasoning techniques Model cards/datasheets
Losing control of an agent	MCP A2A

Notes:

1. **FinOps and DevOps issues** demand a unique approach that integrates performance measurement and risk assessment for the system, during both the development stage and the production stage. This guide does not directly address these issues, but they overlap with certain aspects of the responsible use of AI. The choice or modification of a certain architecture, attribute or method of operation of an AI system can impact the risks being identified and managed as part of the risk management plan. The Business Process Manager should engage in constant dialogue with the technology experts to customize and update the plan in accordance with the proposed changes to the system.
2. **Using a pre-defined AI agent for a designated task:** AI agents are not, in and of themselves, risk mitigation measures, but deploying an AI agent for a well-defined, designated task (such as conducting an online search of a reliable database, measuring outputs against objective parameters, or implementing fail-safe measures) may reduce the risks arising from errors and biases in the AI model or increase system accuracy. Furthermore, by employing AI agents, organizations can identify and automatically escalate cases where human intervention is necessary.
3. **Experimentation with a technological sandbox:** With the GovAI Studio app (once it is fully deployed) or the Vertex AI Platform, the organization can run models and applications and perform initial model corrections.

¹³ For additional information, see the [Guide on implementing privacy-enhancing technologies in AI \(PETs AI\)](#), published by the Privacy Protection Authority.

B. Product components

Components that enhance transparency and end-users' ability to oversee and learn can be integrated in the design of AI-based products:

1. **Disclosure:** The system communicates to users that it is AI-based and discloses its limitations. This improves human oversight and helps frame how the system should be used. Disclosure and transparency have legal dimensions, which will be addressed in the Legal Guide.
2. **User feedback/reporting:** A feedback component that enables users to report problems, miscomprehension or harm, can enhance the organization's oversight capabilities for the product. In appropriate cases, it can allow users to challenge the output.

C. Organizational and business measures

Risks can also be mitigated through organizational processes that enable human discretion, professional oversight, an integrated culture of responsible use, and continuous learning throughout the system's lifecycle. For example:

1. **Human involvement:** Ensuring adequate human oversight throughout the AI system's lifecycle, or at critical junctures in the business process, helps achieve continuous operational control. Such oversight should apply to the development, deployment and use of the system, and should cover specific decisions made by the system. This involvement could be ex ante, real-time or ex post, depending on the risk level and usage context. Full human involvement may not be necessary in all circumstances, but rather in specific cases, such as in a system that screens submitted forms and verifies whether predetermined conditions are met. Legal aspects of this issue will be covered by the Legal Guide.
2. **Consultation with experts:** By involving experts in the relevant fields, the organization can better identify risks and build appropriate mitigation mechanisms. In addition to AI and software engineering experts, experts in other fields (e.g. law, privacy (DPO), equality and inclusion, finance and cybersecurity) should be consulted, as needed, as early as in the product design and development stages.
3. **Periodic sampling:** periodic sampling of system outputs, and spot checks in sensitive cases, can help identify risks throughout the lifetime of the system. [Model drift](#) – i.e. the tendency of a model to deteriorate or develop biases over time – should be taken into consideration. Model drift underscores the need for periodic sampling that can reevaluate pre-established KPIs. Problematic outputs require special attention such as the use of CAG to fine-tune the results, or retraining the model.
4. **AI cybersecurity specialists** to evaluate the system's safety, reliability and resilience, and help identify weaknesses, robustness failures, vulnerabilities, biases, and possible risks. The organization's CISO should be consulted.
5. **Strengthening data governance:** The reliability of an AI system is closely and directly dependent on the data that the models run on. If the risks posed by the system stem from flaws in the original data or the way it is being managed, the following steps can be taken:
 - ✦ Ensuring that the right data is used for the right model – keeping data up-to-date, integral, representative and reliable.
 - ✦ Management of access, authorizations, quality assurance and data integrity.
 - ✦ Taking organization-wide measures to identify and reduce data biases.

Data subjects' rights and interests should be duly considered in the process.

6. **Increasing awareness:** An organization can take several measures, such as conducting training and workshops, or publishing policy papers and internal case studies for users and decisionmakers. The organization can also take measures designed to encourage the reporting of AI risks that materialize.

7. Transferring a project to the Blue Track: As noted, the Blue Track enables experimentation for a limited time, under clear restrictions and with rigorous controls, such as:

- ✦ Determining at the outset the work assumptions, tests and metrics; the measurements should address each of the responsible use principles;
- ✦ Limiting the project to 3-12 months;
- ✦ Limiting the data used, user population, and use cases, based on risk level;
- ✦ Embedding pause or rollback options in the system, without adversely affecting services;
- ✦ Disclosing to users that the system is an experimental system.

Each tool comes with its own costs and risk mitigation capabilities. The trade-offs should be weighed together when choosing which measures to use.

Illustrative example:

A government ministry handling submissions from citizens is considering the use of an AI-based system that determines whether the applicant submitted all of the required documents, or if a document is missing. Prior to implementation of the system, 5% of the submissions were incorrectly flagged as incomplete (false negatives), and 4% of the submissions were incorrectly confirmed as complete (false positives). The proposed AI system is expected to reduce the rate of false negatives from 5% to 3%, but also to increase the rate of false positives from 4% to 6%. In such a case, various mitigation measures can be planned, such as reducing the aggressivity of the system, to reduce false negatives, even if this means increasing the number of false positives. Alternatively, given the time saved and the expected drop in false negatives, the organization could decide to add a human reviewer to check all negative outputs, bringing the false-negative rate close to zero.

End user guidelines

This annex is designed for end users of AI-based systems – i.e. any public servant using AI applications, including GenAI. It contains general recommendations on the responsible use of AI to enhance productivity. Some of the recommendations are applicable to the use of off-the-shelf products available to the general public, while others are relevant to the use of both off-the-shelf products available to the organization and custom applications designed specifically for the organization.

In general, AI systems made available to the staff of an organization will be accompanied by instructions and guidance from the organization's Chief AI Officer and from the relevant Business Process Manager. These will provide a general and complementary framework for the applicable directives and guidelines, as necessary.

This guide does not analyze legal issues or laws applicable to public servants. Rather, it provides a set of recommended principles for the use of AI systems.

Characteristics and limitations of AI systems

End users should be aware of the various limitations of the AI systems they intend to use. The following is a description of the main limitations.

AI models are designed differently and are trained on different databases. Each AI model is unique, with its own characteristics, strengths, weaknesses, and biases. The following are a number of common limitations that users should be familiar with:

- ✦ **“Hallucinations”:** limitations related to reliability and precision: GenAI systems might provide unreliable and inaccurate information, sometimes without the required disclaimers. For example, the results of LLMs occasionally contain citations from inaccurate or fabricated sources.
- ✦ **Harmful content and discrimination:** Algorithms may use discriminatory data based on certain characteristics (such as nationality, gender, etc.) or other data that correlates with discriminatory data. Because AI systems are typically deployed to handle a large number of operations, their errors, biases, or other harmful outcomes may be replicated on a far greater scale than if decisions were made individually by a human.
- ✦ **Dangers associated with protected information:**
 - ✦ **Use of copyrighted information or trade secrets:** AI systems might be collecting IP-protected information, such as copyrighted materials and trade secrets from users, or sharing such information with them. This could result in IP infringement or the unauthorized use of such materials, exposing the state to legal action.
 - ✦ **Use of personal information and the generation of cumulative data:** In many cases, especially when AI systems are provided free of charge, they store large quantities of data entered by users (via prompts, files, etc.), including personal data pertaining to data subjects, as well as the conclusions reached by merging the input data.

- ✦ **Other aspects of the use of protected government information:** Various other types of restrictions and legal protections apply to other government-held information, including information that could prejudice the country's foreign relations. There are other types of information, the disclosure of which could harm other state or third-party interests (such as information generated in the course of policy development discussions, transcripts of conversations between state officials, etc.).¹⁴ Posting such information on AI systems and off-the-shelf tools that have not been approved by the organization, creates certain risks, including legal risks. Disclosure of such information to unauthorized individuals can adversely impact the state, the public and other actors.

Some applications, including those made available free of charge, allow users to delete the information. However, this does not sufficiently preclude the possibility of misuse or improper storage of the information by the system provider. Moreover, many free tools are trained on the information provided by the user.

- ✦ **Outcome explainability problem:** Advanced AI models are like "black boxes", in the sense that the reasoning they use to reach a conclusion cannot be fully explained.
- ✦ **Anthropomorphism:** In some models, especially LLMs (large language models), the system communicates with users using natural and human-like language. Consequently, users may mistakenly attribute empathy, morality, authority or expertise to the system – characteristics it does not actually have. In practice, the system operates through probabilistic text processing.
- ✦ **Human dependency:** By using an AI tool, users may develop a dependence on the tool. This could undermine their ability for creative, autonomous and critical thought. It could also degrade their ability to write text, plan, analyze, design, and more.
- ✦ **Cyber risks:** AI systems are vulnerable to prompt injection attacks from hostile actors, which could affect the system output. Moreover, hostile actors can potentially gain access to the information provided by users.

General guidelines for end users

1. Before using an AI tool, the user must verify that the use of the tool (or the requested use case) has not been banned by the organization's CISO due to reliability or data security concerns.
2. The user should also check whether the organization has issued any relevant guidelines. Such guidelines may have been issued by the legal department or the Chief AI Officer, for example.
3. The user should check whether the AI tool is **third-party system** (such as a private Gemini or Claude account) or one that was authorized by the organization to process its data (such as the organization's cloud, or its information systems).
4. If the tool is **authorized by the organization**, it must only be used in accordance with the instructions for use given to end users.
5. If it operates within a **third-party environment** (e.g., an online chatbot using a private account), the following guidelines must be followed:
 - 5.1 No protected information, such as personal information, information that may not be disclosed under the Freedom of Information Law,¹⁵ IP-protected information or trade secrets, or classified information, may be entered into external systems.

¹⁴ See the limitations and exceptions to the Freedom of Information Law of 1998 for more on this, particularly those listed in Sections 9(a) and 9(b) of the law.

¹⁵ See Section 9 of the Freedom of Information Law of 1998.

- 5.2 Prompts that convey an intention to take action against a specific identifiable individual, or to take a particular sensitive governmental action, must not be entered.
- 5.3 The user should review the AI system's policy and terms of use to better understand the tool's limitations and data processing practices. When in doubt, employees should consult with their supervisors.
- 5.4 Automatic interfaces between the third-party AI tool and the local environment may not be created independently without consulting first and obtaining authorization from the relevant officials within the organization, such as the CISO, the DPO, the actor responsible for AI risk management.
6. AI output should be used only to support decision-making, not as the sole basis for a final decision. However, if the AI tool in question has been integrated for use within the organization in accordance with the Guide to the responsible use of AI in the public sector, and it has obtained all necessary clearances – including legal approval – it may be utilized beyond mere decision support to allow for a degree of automation, provided that such use is explicitly authorized by the governing policy.
7. If a given output indicates that personal, sensitive or classified information has been disclosed (whether from an off-the shelf product or a system approved by the organization), this should be accurately documented and reported to the relevant Business Process Manager or to the Chief AI Officer.
8. **Writing accurate prompts:** It is recommended to write complete and accurate prompts, with explicitly instructions, to ensure accurate and relevant results.
9. **Content should not be assumed to be correct if it has not been checked:** The outputs should be cross-checked against reliable information sources, such as scientific publications or well-known websites. For example, if the user normally deals with property valuation and asks a question on taxation, the user should consult with colleagues who are taxation experts to verify the reliability of received outputs.
10. **Transparency and documentation:** If an AI system was used substantially to produce a certain deliverable, this should be stated explicitly, as appropriate, in the final deliverable. If an AI system (such as a customized system integrated into the organization's activities) is used to make a decision, the use of the system should be documented. Because transparency raises certain legal issues, employees should follow the organization's transparency guidelines, which will be developed, inter alia, on the basis of the Legal Guide.
11. **Note whether IP-protected information was used.** When using generative AI, the output may contain copyrighted material. Discretion should be exercised when using such output. In case of doubt, the tool's information sources should be located, if possible, and the organization should assess whether references must be included. In any case, the organization's legal department should be consulted.

Annex E

Sample policy on the responsible use of AI



This is a template to help organizations draft an internal policy based on the principles set out in this guide, including its risk management method.

Chief AI Officers formulating their organization's policy may use the following text and adjust it as they see fit, based on the characteristics and needs of their organization. Text in square brackets indicates areas that the chief AI officer must complete.

Table of Contents

1. Preface
2. Basic principles
3. AI Governance: key roles and areas of responsibility
4. AI project approval process
5. Transparency and public awareness with respect to the use of the AI system
6. Audits
7. AI incident response
8. Training
9. Employee guidelines

1. Preface

This policy sets forth the organization's principles, processes, and responsibilities to advance the informed, responsible and ethical use of AI systems. It is based on the Israel National Digital Agency's Guide to the responsible use of artificial intelligence in the public sector. (the "Guide"). The aims of the policy are as follows:

- ✦ Supporting the use of AI systems to improve the services the organization provides and advance the organization's goals;
- ✦ Promoting an organizational culture that champions the responsible use of such systems for both internal and external needs;
- ✦ Advancing AI literacy among the organization's employees;
- ✦ Integrating AI systems in a way that benefits employees and enhances their work environment.

The policy applies to all of the AI systems being used, developed, or assessed by the organization.

It applies to all stages of the AI system's lifecycle, from the product requirements analysis to the design, development/procurement, deployment, commissioning, and decommissioning of the system. Note, however, that the policy does not detract from other rules and regulations applicable to the various stages.

This policy is complementary to the relevant TAKAM provisions and to the Guide.¹⁶ [references to other organizational policies should be added].

In general, when assessing the deployment of an AI tool, tools available under the Nimbus Agreement (Tier 1 and Tier 5) will be given preference.

2. Basic principles

The organization will take steps to implement the principles for the responsible use of AI adopted in the Guide.

These principles will be implemented through proactive steps taken at each stage of the lifecycle of the AI systems used by the organization, and will be integrated into the organization's decision-making processes. The following table lists the principles, alongside the main measures that the organization will use to implement them.

Principle	Implementation
Growth, sustainable development and well-being	Before developing or deploying any AI system, a comprehensive assessment of the potential benefits will be performed, while evaluating the system's contribution to general well-being and its ability to improve public services and reduce the exclusion of vulnerable populations and harm to the environment.
Respecting the rule of law, human rights and democratic values, including fairness and privacy	Each AI system will undergo a rigorous ethical and legal assessment during the design and development stages, led by the legal office and the DPO.
Transparency and explainability	The organization will publish relevant information about the main AI systems it uses on its website and on AI Watch . If the system interacts with citizens, a clear statement to that effect will be displayed, identifying the AI system and indicating how citizens may contact officials in the organization, where appropriate.
System safety and security	AI systems will be developed and deployed in accordance with strict data security and cybersecurity standards. Tools available on the governmental cloud (Nimbus) will be given preference. Ongoing testing, monitoring, and malfunction response mechanisms will be deployed. High and medium-risk AI systems will undergo comprehensive testing, and appropriate risk mitigation measures will be applied.
Accountability	Roles and responsibilities for each stage of the AI system's life cycle will be clearly defined and documented. Routine reporting and control mechanisms will be established, and the principles of the policy will be enforced. Each AI incident will be thoroughly investigated. The organization will learn from incidents and take corrective actions as needed.

¹⁶ Additional guides and instructions will be published (such as the YAHAV policy and the legal guide on AI use in the public sector), to be published by the Office of Legal Counsel and Legislative Affairs. Once these guides are published, the policy will need to be updated accordingly.

3. AI Governance: key roles and areas of responsibility

The following are the main roles for the responsible use of AI in the organization:

Role name, as per the guide	Relevant individual in the organization [please complete]
Chief AI officer	[Designate whether this is the CDO, CTO or another official]
Business Process Manager	[Designate which individuals will be authorized to advance AI projects in the organization. Examples: heads of divisions, team directors, VP Finance, VP HR, etc.]
End users	The employees in the Business Process Manager's team (including external advisers, as appropriate)
AI Governance Forum	A forum comprised of the following roles: [specify the roles]
Legal department	[Designate the legal advisors in the organization who will support AI processes]

[Other roles may be added, as necessary]

These officers must collaborate to ensure that this policy is successfully implemented, and to ensure good governance of AI systems in the organization.

4. AI project approval process

The scope of the approval process depends on the system classification, which is determined based on the Guide (the green, yellow, red and blue tracks). [Here, the relevant sections from [Annex C](#) of the Guide (risk assessment) can be copied and adapted to the needs of the organization].

Organization-specific guidelines relating to the risk management process described in the Guide:

A. Consultations:

- ✦ If the system is particularly sensitive [describe the organization's criteria for such a designation separately], individuals outside of the organization may be consulted, subject to the approval of the Chief AI Officer and to confidentiality requirements.
- ✦ If the system directly and substantively impacts citizens, the public should be involved early on, and feedback should be collected from relevant stakeholders, in coordination with the Chief AI Officer and the legal department.

B. Formulating a risk management plan:

- ✦ The business process manager will draft a risk management plan, based on [Annex C](#) of this guide.
- ✦ The Chief AI Officer will support the Business Process Manager, providing information and tools throughout the system's lifecycle.
- ✦ The Chief AI Officer will contact the Israel National Digital Agency if other tools or knowledge are needed to help create the plan.
- ✦ Others within the organization can be consulted at any time, to better understand the risks and develop approaches for handling those risks.

C. Auditing: As per the risk management plan.

D. Reporting: The plan should include a timetable for the submission of periodic reports, and reports to the Chief AI Officer in the event of an AI incident.

E. The tests performed and conclusions reached at each stage of the project must be documented.

5. Transparency and public awareness with respect to the use of the AI system

The following instructions pertain to transparency in the organization's interactions with the public and with other stakeholders within and outside of the organization, when the AI system is used. These instructions may supplement or expand upon the legal instructions, which will be developed in consultation with the legal department.

The organization will publish a register of approved AI systems that may have a direct impact on citizens. The register will specify:

- ✦ The purpose and use of the system;
- ✦ Its main operating principles;
- ✦ Data that the organization has used to validate the model or provide context;
- ✦ Types of data the system will use when it is commissioned;
- ✦ The main risks posed by the system;
- ✦ Mitigation measures that were put in place.

Moreover, each AI system used by the organization (including "green track" systems) will be publicized on the [AI Watch](#) website. If a main system or platform is comprised of two or more agents, it may be sufficient to publish general information about the system as a whole, while detailing the roles of the main agents.

The organization will establish a public feedback channel using an online form and a dedicated email account that will receive the feedback, questions and public inquiries on the organization's AI use cases. The Chief AI Officer is responsible for managing these channels and responding to inquiries.

For systems that decide on the rights of businesses or individuals, or that play a significant role in supporting such decisions, the organization will clearly explain – in plain language – how anyone adversely affected can obtain details about the decision or challenge it.

6. Audits

[The table appearing in Annex C, Part 1 of the Guide can serve as a foundation that can be expanded upon]

7. AI incident response

(This section relates to the recommendation in Annex A of the Guide regarding the establishment of an AI incident response team)

An **AI Incident Response Team (AIRT)** comprised of [the organization's AI Governance Forum can be mentioned here, as well as the Business Process Manager, Chief AI Officer, a technology officer, a representative of the legal department, and a senior member of the organization's management]. The AIRT's role is to investigate and respond to AI incidents.

The following steps will be taken to handle AI incidents:

- A. If an end user, or the system itself, detects a potential AI incident, either through routine monitoring or as a result of an external report, the end user (or the system) must report this immediately to the Chief AI Officer. The report must state the date, a description of the event, and the assessed impact.

- B. As soon as the report is received, the Chief AI Officer must convene the AIRT. The AIRT will then collect the technical data (i.e. logs, model versions, I/O data) and analyze the scale of the damage and the reasons for the failure (e.g., a technical malfunction, misuse, a security breach, bias, violation of privacy, or unauthorized use). The AIRT will determine whether the incident is indicative of a broader issue in the system or an isolated incident. The AIRT will document its findings.
- C. Response: Based on the findings and the scope of damage, the AIRT will decide which actions must be taken. Those measures may include the following:
- ✦ Disconnecting or temporarily suspending the system or certain system components;
 - ✦ Defining workarounds, to ensure functional continuity;
 - ✦ Handling the incident, either locally or broadly (changing the result, compensation);
 - ✦ Deciding how and to what extent the incident should be reported to the public, if relevant;
 - ✦ Recommissioning the system by taking measures such as carrying out system repairs, changing the way data is collected or tagged, or retraining the model;
 - ✦ Conducting operational and functional testing before the system is made available to users.
- D. Lessons learned and continuous improvement: If necessary, the AIRT will recommend changes to the development, testing and control processes and to training procedures, as well as steps to boost the awareness of employees and the relevant teams to prevent further AI incidents.

8. AI training program

AI training aims to achieve the following:

- ✦ Increase awareness and understanding of artificial intelligence, including its advantages and potential risks, among all of the organization's employees;
- ✦ Instill the responsible use of AI as part of the organizational culture;
- ✦ Help employees understand how AI tools can contribute to their work;
- ✦ Encourage innovation and entrepreneurship in the organization on how new tools can be integrated;
- ✦ Provide teams that implement AI with practical tools for formulating and operationalizing an AI risk management plan.

The training program will include courses on the following topics, which are listed in the table. It will be updated at least annually, and on an ongoing basis to reflect technological and regulatory developments. The precise contents and logistical aspects will be determined in consultation with the organization's HR department.

Subject	Content	The Target Audience
Introduction to artificial intelligence and responsible use	What is Artificial Intelligence, and how does it work? Types of AI (generative AI, predictive AI, etc.); Potential advantages and applications of AI in the public sector; Main risks (e.g., hallucinations, algorithmic biases, violation of privacy, data security); Responsible AI principles, and end user guidelines	All employees
The organization's responsible AI policy	In-depth review of the organization's AI policies, principles, and roles and responsibilities Track classification processes (green, yellow, red and blue tracks), risk and benefit analysis, technologies used for mitigating risks and detecting, monitoring, and reporting AI incidents An overview of legal aspects, in conjunction with the legal department, or the Office of Legal Counsel and Legislative Affairs, in the Ministry of Justice	The organization's Business Process Managers
Artificial intelligence as a work support tool – tools and practical use cases	Hands-on experimentation with AI tools approved for use in the organization, development tools, the tools' limitations, and details on risk reduction technologies	The organization's Business Process Manager and the technology teams

9. Employee guidelines

When using an AI system, employees must follow the supplier's terms of use, as provided to them by the business process manager. Additional guidelines:

For systems implemented in the organization through an internal procurement process: Follow the instructions for use provided by the [individual tasked with providing such instructions – the Chief AI officer or the Business Process Manager]

For systems appearing on the Tier 5 of the Nimbus system: follow the Israel National **Digital Agency's [End User Guidelines](#)**.

For systems made available to organizations by the Israel National Digital Agency (through the [Governmental Data Highway](#)): follow the instructions in the "product policy" page, on that system's website.

Generic third-party systems (such as a private account on Gemini or Claude): follow the Israel National Digital Agency's End User Guidelines.

Automatic interfaces between the third-party AI tool and the local environment may not be created independently without the authorization of the [organization's CISO]

These guidelines are subject to the data security rules issued by the [organization's CISO]

Annex F

Glossary



The following is a glossary of common terms in the field of artificial intelligence that appear in this Guide.¹⁷

Term	Definition
Generative artificial intelligence (GenAI)	AI systems capable of creating new content, such as text, images, videos and audio files, based on examples or data they were trained on. These models can generate creative outputs, without needing to be specifically trained to perform each task. There are different types of applications that generate new content, including: ✦ Visual content: Generating images, videos, or graphics. ✦ Audio content: Generating synthesized speech, soundtracks or narrations. ✦ Textual content: Writing creative, professional or technical texts. These tools rely on extensive databases and machine learning to create content that mimics human-generated content. Nano Banana, Sora (for visual content), Resemble AI (for audio content), Claude or ChatGPT (for textual content).
Explainable AI (xAI)	The development of methods that allow humans to understand, interpret, and trust the results AI systems produce. Explainable AI aims to make the model's "black box" more transparent by explaining how a given AI system reached a certain decision.
Fairness Testing	A quality control process designed to assess whether the model behaves consistently and without discrimination across different population groups. The process involves the use of testing techniques (such as combinatorial testing) to detect incidents in which the model generates different results for similar inputs that differ only in an attribute that may be discriminatory.
Prompt	A prompt is an instruction given to a generative artificial intelligence system. The prompt is written in free text and is used to communicate with the system.
Homomorphic encryption	An encryption technique allowing computations on encrypted data without decrypting it. AI models can be trained on encrypted data, ensuring that the data remains protected throughout the training process. This way, both the trained model and its outputs can remain encrypted, accessible only to authorized users. For example, an insurance company can transmit encrypted medical data to a cloud service provider for the purpose of training a predictive model, without exposing the underlying information. The provider performs computations on the encrypted data, and the final result can be decrypted only by the company. (Source: The Privacy Protection Authority's Guide to Privacy Enhancing Technologies (PETs))
Anonymization (redaction)	Data anonymization is the removal of features or alteration of values in order to reduce or prevent identification of the data subject or the ability to recover personal information from the training input or from the trained model. Applying anonymization to training data hinders the ability to link specific data entries to individuals, thereby reducing the risk of re-identification from the trained model. One common approach to anonymization involves rounding values or grouping them into ranges, for example, representing age as a range (e.g. 40–49) instead of an exact value (e.g. 47), or using income range in place of precise monthly income figures. This reduces the risk of re-identification but also decreases the granularity of the data. (Source: The Privacy Protection Authority's Guide to Privacy Enhancing Technologies (PETs))

¹⁷ See also the Glossary of IT terms

Term	Definition
Privacy enhancing technologies (PET)	PETs are technologies that allow data to be processed, analyzed or shared, while protecting the privacy of the people whose data was collected. Using these technologies and employing privacy by design will minimize violations of privacy.
The RAG or CAG technique (Retrieval-Augmented Generation/Cache-Augmented Generation)	These are techniques that provide an AI model with a specific dataset that it draw from when it processes the question being asked, thus generating more accurate, up to date and detailed responses to complex questions. When a generative model retrieves information from reliable sources, risks stemming from “hallucinations” and system errors tend to decrease, and the output tends to be more explainable. Moreover, by feeding up-to-date and reliable data sources administered either by the organization or by third parties, these techniques reduce the risk of unreliable or erroneous outputs.
Temperature	Temperature is a parameter in generative models that sets the output’s randomness and creativeness level. A low temperature (0.1-0.3) will result in consistent, safe, and cautious answers. A high temperature (0.7-1.0) will result in diverse answers that are also less predictable, and may occasionally be incorrect or problematic. In certain LLMs, temperature control can be performed by end users through prompts written in natural language, and in certain models, a technical officer will need to set the temperature by calling the model’s API.
“Personal Assistant” applications using GenAI	Programs or applications designed to help users with their day-to-day lives or business tasks. Users can use GenAI to plan meetings, suggest ideas, provide customized insights, and even create customized content. Examples: Siri, Google Assistant, Microsoft Copilot.
Off-the-shelf AI tools	AI-based tools, software or platforms that can be used “out-of-the-box”. These are general-purpose AI tools that can be used for different uses, and in some cases, these tools can be customized.
Monitoring tools	Tools and processes used to detect changes in a model’s performance over time, especially model drift.
A debiasing tool (reweighting, adversarial debiasing)	Techniques for reducing discriminatory biases in algorithms. Reweighting is a method used to assign different weights to different data items, before a model is trained, to compensate for the underrepresentation of certain groups. Adversarial debiasing is a method used to train the model to generate accurate predictions, and prevent an “adversarial” component from inferring protected attributes from the predictions (such as gender or ethnicity).
Federated learning	Federated learning allows collaborative training of AI models using participants’ data (local models), after which the patterns identified in the local models are aggregated into a single, accurate, global model, without the participants having to share the underlying raw data used to train the model. Federated learning can be achieved using several methods: 1. Centralized Federated Learning: A central server generates an algorithm or a model, and sends it to distributed data sources. The model is updated according to the local data sources and sent back to the central server, which creates a unified weighted model. 2. Decentralized Federated Learning: In this model, there is no central server involved. The parties communicate directly with each other and update the model based on their local data. (Source: The Privacy Protection Authority’s Guide to Privacy Enhancing Technologies (PETs))
On-prem model	The deployment and operationalization of AI models on an organization’s own infrastructure, rather than third-party cloud providers. This approach allows the organization to retain direct control of data security and makes it customizable, and it is generally used in organizations with highly sensitive information.
Reasoning model	Models designed to carry out logical thinking processes, such as chain-of-thought (CoT) and self-reflection, to solve complex problems more accurately, by mimicking human thought patterns: it breaks down complex problems into stages, experiments with different strategies, and self-corrects during processing.
Protected information	In this document, protected information is a general term referring to information to which various protections apply, including personal information, as defined in the Privacy Protection Law, information subject to confidentiality obligations (pursuant to any applicable law, judicial ruling, or contract), information protected for reasons of state security, information that contains intellectual property or a trade secret, privileged information (e.g., due to a government or public interest), etc.

Term	Definition
Synthetic data	Synthetic data refers to artificially generated data designed to mimic “real” data, based on the statistical properties of genuine datasets, but is not supposed to preserve the original details of any person or event. In other words, it is fictional yet structurally accurate data intended to serve as a substitute for sensitive data, such as during the development, testing, or training of AI systems. Synthetic data allows artificial intelligence models to be trained while reducing the risk of exposure to personally identifiable information. (Source: The Privacy Protection Authority's Guide to Privacy Enhancing Technologies (PETs))
Conversational search engines	Systems that combine traditional search engines with natural language processing (NLP) technologies, allowing users to have natural and intuitive conversations with the system to search for data. Instead of sufficing with rule-based results, these search engines respond with dialogue, explaining the information and suggesting advanced contexts based on their understanding of complex queries. Examples: ChatGPT, Google AI Search, Perplexity
Bias evaluation frameworks	Structured methodologies (like NIST AI Risk Management Framework) for detecting, measuring, and reducing potential biases in AI systems throughout their life cycles. These frameworks define reliability criteria, map potential social impacts, and recommend courses of action.
AI system	An artificial intelligence system is a machine-based system that uses the inputs it receives to infer ways to generate outputs, such as forecasts, content, recommendations, or decisions that could affect physical or virtual environments, for explicit or implicit purposes. Different AI systems vary in terms of their autonomy and their adaptability after deployment. ¹⁸
Guardrails	Defining the technical rules and control mechanisms used to ensure that the system operates in line with the policy that was set. Guardrails can, for example, prevent the system from addressing prohibited or sensitive topics, or answering problematic questions. They can be used to issue alerts, set the tone of a conversation with a user, and define acceptable response patterns.
AI agent	An autonomous or semi-autonomous AI system used to carry out tasks or achieve defined goals while interacting with its environment. In general, this term relates to a tool that fuses an LLM with an application that can enable it to conduct various actions. An AI agent system also contains a routing agent that directs tasks to the various agents, based on the tools they have.
Concept drift/model drift	Drift occurs when the relationship between inputs and their correct outputs varies between laboratory settings and the real world. Drift implies that the model's accuracy has been degraded, making it vital to retrain or make adjustments to the model.
Differential privacy	This technique limits the impact of any single data point on the learning process, preventing the identification or reconstruction of personal information even if an adversary has access to the model's outputs or external knowledge. By applying differential privacy, the model substantially reduces its dependency on any single identifiable example, making it extremely difficult to reconstruct individual information. The main advantage of this category lies in its relative simplicity and the ability to implement the techniques without requiring extensive computational resources or collaboration across multiple entities. However, data transformation may adversely affect model quality. When noise is added or generalization is applied, the data loses precision or granularity, which can, in turn, impact the performance of AI models. For example, a model trained using age ranges may be less accurate than one trained on precise age data. (Source: The Privacy Protection Authority's Guide to Privacy Enhancing Technologies (PETs))
Black box	The term “black box” is often used to describe the way AI systems work, in the sense that AI outputs do not necessarily lend themselves to being explained. In other words, it is not possible to fully understand how or why a particular decision was made. This is because of the complexity of the AI system's underlying algorithms and the quantity of data that those decisions are based on. (Source: Final Report of the Interministerial Team for Assessing the Regulation of AI in the Financial Sector.)
Transcribing and summarizing meetings	Tools that use ASR (automatic speech recognition) and NLP (natural language processing) to accurately transcribe conversations or meetings and create lucid and concise summaries. They contain advanced features, like speaker recognition, speaker diarization, noise filtering, key insight extraction, and action item extraction. Examples: Notion AI, Microsoft Teams transcription, Otter.ai

Term	Definition
MLOps (machine learning operations)	MLOps is a methodology that combines the development of machine learning models with system operations to create an automatic and continuous process of developing, deploying, monitoring and administering the models in the production environment, similarly to what DevOps does in the software domain.
AI incident	An event, circumstance or set of events in which the development, use or malfunction of one or more AI systems directly or indirectly causes damage, such as harm to a person's health, disruptions to the management and operation of critical infrastructure, or a violation of legal obligations. AI incidents can arise from events such as changes to an AI model, attacks against the model itself or its underlying data, or unexpected results from a system.
A2A - Agent to agent communication	A technological trend allowing autonomous AI systems ("agents") to communicate with each other, transfer information, or carry out orchestrated actions. Human control of this process is preserved through the use of comprehensive security protocols and well-defined interfaces that predetermine the way agents can "detect" each other, and the information they are allowed to exchange.
Tool use	An AI model's ability to access third-party resources and communicate with other systems (through an API, for example), to perform actions or retrieve data, instead of relying solely on internal information or on its own abilities. Using additional tools may enhance system accuracy.
Model Armor	One of the Google Cloud's designated security services for LLMs, which serves as a real-time "layer of protection" between the user and the AI model. The tool helps detect and even block attempts to bypass the model's restrictions, prevent personal information from leaking, and filter out inappropriate content.
MCP	An open-source standard that enables developers to create secure bidirectional integration between AI assistants (like ChatGPT) and business systems and databases (like Google Drive). This standard eliminates the need to create a separate integration for each model, and creates a common protocol for sharing contexts.

מערך הדיגיטל הלאומי
نظام الديجيتال الوطني
Israel National Digital Agency

