Strasbourg, 26 August 2026 T-PD(2025)3

CONSULTATIVE COMMITTEE OF THE CONVENTION FOR THE PROTECTION OF INDIVIDUALS WITH REGARD TO AUTOMATIC PROCESSING OF PERSONAL DATA

(CONVENTION 108)

Draft Guidelines on Privacy and Data Protection in the context of LLM-based systems

www.coe.int/dataprotection

CONTENTS

I. INTRODUCTION

These guidelines address the privacy and data protection risks associated to the lifecycle of the LLM-based systems. Large Language Models (LLMs) represent a particularly influential technological development, combining advanced language-processing capabilities with large-scale automation across sectors, thereby amplifying both opportunities for innovation and challenges for the protection of individual's rights and freedoms. These technologies create significant opportunities for individuals, organisations and public authorities, while they may generate serious risks for the effective enjoyment of human rights, the functioning of democracy, and the observance of the rule of law.

Trained on vast amounts of data and embedded in increasingly complex digital infrastructures, LLM-based systems can be used to automate a wide range of tasks, from content generation and summarisation to support decisions and automate interaction. Because their development and operation depend on the large-scale processing of personal data, LLMs raise significant privacy risks and data protection considerations. Their technical characteristics, including opacity, probabilistic outputs and continuous adaptation, challenge traditional data protection safeguards and complicate the effective application of existing data protection standards.

The present Guidelines are interpretative in nature and neither create new rights nor impose obligations additional to those arising under the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 108, "Convention 108") and its Amending Protocol CETS No. 223 (referred in this document as "the Convention" or "Convention 108+"). They adopt a principles-based approach and comprise two complementary elements: interpretative guidance explaining how the principles and obligations of the Convention apply throughout the lifecycle of LLM-based systems, and recommendations supporting their practical implementation.

Guidelines are aimed at providing an overarching, high-level and supplementary guidance on the application of the principles and obligations of the Convention across the lifecycle of LLM-based systems. Where the technical characteristics of LLM-based systems make it difficult to determine how a principle should be applied in practice, the recommendations provide non-exhaustive illustrations of measures through which existing obligations may be fulfilled. Consistent with the technology-neutral character of the Guidelines, those measures do not constitute independent requirements or prescribe any specific technology, architecture or organisational arrangement. Their selection and implementation should be proportionate to the nature, scope and context of the processing and to the risks posed to the rights and fundamental freedoms of individuals. Alternative measures may be adopted where they give equivalent effect to the applicable obligations and provide an equivalent level of protection. The implementation of any one measure does not, in itself, demonstrate compliance with the Convention; compliance must be assessed in the light of the applicable obligations and the effectiveness of the safeguards as a whole.
LLM ecosystems may involve multiple [persons and organisations/actors] whose roles and responsibilities are distributed and are interdependent across different stages of the lifecycle. The recommendations therefore follow a multi-stakeholder approach and are tailored to specific categories of addressees, having regard to their respective legal, technical and operational responsibilities.

This approach is also grounded in the human-centred perspective and commitment to responsible innovation reflected in the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), which ensures that activities within the lifecycle of AI systems are fully consistent with human rights, democracy and the rule of law, while being conducive to technological progress and innovation.

II. SCOPE AND KEY PRINCIPLES

[These Guidelines concern the processing of personal data in the context of the design, development, provision, integration, deployment, use, significant modification and decommissioning of LLMs and LLM-based systems, including AI agents and agentic AI systems, insofar as such processing falls within the scope of Article 3 of Convention 108+. They cover processing at model and system levels and throughout the relevant processing chain, including material intermediate operations and processing carried out through retrieval components, memory, tools, external services, communications between agents and operational records.

They are addressed, as appropriate, to Parties, competent public authorities, data controllers and processors and, to the extent relevant to their functions under applicable law, to other natural or legal persons involved in the relevant stages of the system's lifecycle. Their respective roles and responsibilities should be determined according to their actual functions and decision-making power in relation to the data processing, and not solely by contractual designations or technical descriptions of the system. The distribution of components or functions among several persons or across several jurisdictions should not, in itself, prevent the identification of the controllers and processors responsible for particular processing operations.

Their application depends on the processing actually carried out and the information concerned, rather than on the technical or commercial designation of the model or system. The absence of personal data from the initial input or final output does not, in itself, establish that no processing of personal data takes place within intermediate operations or connected components.

(1) Machine Learning, LLMs and LLM-Based Systems

Machine learning comprises computational techniques through which the parameters of a model are determined or adjusted on the basis of data so that the model can identify patterns or relationships and produce outputs in response to inputs. A machine-learning model is the computational representation resulting from that process. Training refers to the process through which the parameters of the model are determined or adjusted, whereas model inference refers to the application of a trained model to an input in order to generate an output. The generation of a new output during model inference does not, in itself, mean that the model learns from the interaction or that its parameters are modified. In this technical sense, model inference should be distinguished from the inference of personal data referred to elsewhere in these Guidelines.

Where a system is designed to update the underlying model during operation, including through online or continual learning, that updating constitutes a distinct processing operation and should be identified, documented and assessed accordingly. Training, post-training adaptation and model inference may each involve the processing of personal data, although the nature, purposes and risks of that processing may differ.

An LLM is a type of machine-learning model, typically based on a deep neural-network architecture and trained on large quantities of text and, where applicable, other forms of data. It is designed to process and generate language-related outputs, including by predicting sequences of tokens. LLM may provide general-purpose capabilities which can be adapted to a variety of tasks. The model should nevertheless be distinguished from the LLM-based system in which it is incorporated.

An LLM-based system comprises one or more LLMs and the technical components through which their capabilities are made available or used. Depending on its design and purpose, these components may include system instructions, user interfaces, application programming interfaces, retrieval mechanisms, memory, external data sources, tools and services, identity and credential services, monitoring functions and human-control mechanisms. The organisational arrangements governing the use of those components also form part of the context in which the processing should be assessed.

The purpose, context and consequences of processing carried out by an LLM-based system cannot be determined solely by examining the underlying model or its final output. Data protection risks arising from the training data or properties of the model, including memorisation and unintended disclosure, may be reflected or amplified at system level. Additional risks may arise from the system's configuration, data sources, integrations, permissions, recipients and operational environment. Even where the underlying model remains unchanged during operation, personal data may be processed through prompts, retrieved context, outputs, short- or long-term memory, tools, external services and operational records.

LLM-based systems may, but do not necessarily, exhibit agentic characteristics; this depends on their functional configuration and the degree of autonomy enabled at system level. Agentic characteristics arise at system level from the manner in which one or more models are configured and authorised to pursue specified objectives, formulate or select sequences of steps, use tools or other agents and perform actions. Agentic behaviour does not necessarily imply that the underlying model learns, retrains or modifies its parameters during operation. In particular, the use of system instructions, retrieval mechanisms or memory, or the adjustment of actions in response to context, should not, in itself, be treated as training or modification of the underlying model.

These distinctions are relevant to the application of Convention 108+, and hence, to the scope of these Guidelines. The fact that the underlying model is not trained or modified during operation does not, in itself, establish as described above, that no processing of personal data takes place. Personal data may nevertheless be collected, generated, inferred, retrieved, retained, combined, disclosed or otherwise processed through prompts, retrieved context, outputs, memory, tools, external services,

communications between agents and operational records. Conversely, where data obtained during operation are used to update, fine-tune or retrain the model, that use constitutes a further processing operation whose purpose, legitimate basis laid down by law and compatibility with the original purpose should be assessed separately.

The distinction between the model and the system is also relevant to the determination of the respective roles and responsibilities of the natural and legal persons involved in the data processing chain, notably the identification of the applicable processing purposes and legitimate basis, the examination of likely impact referred to in Article 10, paragraph 2, of Convention 108+, the selection of appropriate safeguards and the effective exercise of data subjects' rights under Article 9. The privacy risk impact assessment should therefore address the LLM-based system as deployed and the relevant processing chain, and should not be limited to the underlying model, its training data or the final output. The technical designation of a component, or the distribution of functions among several components or persons, should not obscure or alter the responsibility of data controllers and processors for the processing operations concerned.

(2) AI Agents and Agentic AI Systems

For the purposes of these Guidelines, an "LLM-based AI agent" means an LLM-based system that, in pursuit of a specified objective, can receive and interpret information from its environment, formulate or select a sequence of steps, invoke tools, services or other agents and perform actions capable of affecting a physical or virtual environment, with varying degrees of human involvement. An "agentic AI system" may comprise one agent or several interacting agents and may include an orchestrator, short- or long-term memory, retrieval components, tools and application programming interfaces, identity and credential services, communication protocols, monitoring functions and points at which effective human review, authorisation or intervention is required or available. These functional descriptions do not require every system to display the same features or degree of autonomy.

The range and nature of actions that an agent is technically capable of performing, and those that it is authorised to perform, should be distinguished from its degree of autonomy. The agent's action-space concerns the systems, personal data, tools and operations made accessible to it, including whether access is read-only or permits creation, modification, communication, deletion or execution. Its degree of autonomy concerns the extent to which it may select objectives, steps, timing and means without prior human approval. Privacy and data-protection risks may increase with either the breadth and potential consequences of the action-space or the degree of autonomy. A broad or consequential action-space may give rise to significant risks even where a formal human oversight or approval step is present.

The terms "agent" and "agency" are used in these Guidelines to describe functional characteristics. They do not imply consciousness, intention, legal personality or an independent legal role. The apparent autonomy of an agent should neither displace nor dilute or alter roles and responsibilities of data controllers and processors. Describing a system as an agent, or increasing its technical autonomy, should not obscure or alter obligations of the natural or legal persons involved in the processing.

Agentic AI is a means of processing and does not, in itself, constitute a specified purpose or a legitimate basis laid down by law for the processing of personal data.

(3) Personal Data and Applicability of Convention 108+

From a data protection perspective, personal data may be processed at every stage of the lifecycle of an LLM or LLM-based system. The nature and severity of the resulting risks depend, *inter alia*, on the application, design and development, and deployment contexts[1]. Pursuant to Article 3 of Convention 108+, the Convention applies to processing within each lifecycle stage where that processing falls within its scope, with a view to securing every individual's [right to the] protection of personal data and thereby contributing to respect for human rights and fundamental freedoms, in particular the right to privacy.[2]

In determining whether information constitutes personal data, regard should be made to the information processed and to the means reasonably likely to be used to identify an individual, including through combination with other information. This assessment should be carried out throughout the lifecycle and should cover training and post-training data, prompts and outputs, inferred data, retrieved context, memory, tool parameters, operational records and data disclosed or made available to other agents or services. Technical designations such as "synthetic", "de-identified" or "internal" do not, in themselves, determine whether information is personal data. Considerations related to special categories of personal data and the examination of likely risk impacts are addressed in sections IV(3) and IV(7) of these Guidelines, respectively.

III. PRIVACY AND DATA PROTECTION RISKS ACROSS THE LIFECYCLE

The nature, likelihood and severity of privacy and data protection risks depend on the purposes, context and scale of the processing, the categories of personal data and data subjects concerned, the operational environment and the safeguards applied. Particular attention should be paid to children and to individuals belonging to vulnerable groups. These factors should be considered across the following lifecycle stages:

- ***Model creation:*** This stage includes the collection and preparation of data for training and the development of the model. Training datasets may contain personal data collected without an appropriate legitimate basis or beyond what is necessary for the specified purpose. Relevant risks include insufficient

[1] Covers all activities from the design of an AI system to its retirement, regardless the actors involved. Explanatory Report, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), para 15.

[2] This approach is consistent with the holistic assessment and contextualisation of various risk factors afforded by the Council of Europe risk and impact assessment framework of AI systems from the point of view of human rights, democracy and the rule of law. Council of Europe risk and impact assessment framework of AI systems from the point of view of human rights, democracy and the rule of law (HUDERIA), p. 14.

transparency, poor data quality, memorisation and unintended disclosure, bias and adversarial interference.

- ***Post-training adaptation:*** This stage includes system instructions and alignment, fine-tuning and adaptation to specified tasks or operational purposes using targeted datasets and task-specific inputs. While identifiability and quality of data remain as one of the primary concerns (due to the amplification of potential risk of data inaccuracies and bias), repurposing data processing beyond the original purpose pursued, insufficient transparency and safeguards represent additional challenges and do not reflect the intended context of use.
- ***System integration:*** This stage concerns the integration of an LLM into an application, service or compound system, including through orchestration, retrieval components and external tools or services. It may increase risks of excessive processing, cross-context profiling and inference, re-identification of data subject through data combination and inferring additional identifiers, unauthorised access. Transborder flows of personal data due to cloud-hosted LLMs (API-based processing) require appropriate safeguards and adequate level of protection.
- ***Operational deployment and end-user interaction***: This stage covers live operation, user interactions, autonomous workflows, memory functions, monitoring and post-deployment governance. Risks include the generation or disclosure of personal data, retention and secondary use of interaction data, decisions or actions based on inaccurate information, barriers to the exercise of data subjects' rights, security incidents and insufficient accountability.
- ***Significant modification and redeployment:*** Changes to the model, system instructions, orchestration, tools, data sources, memory, permissions, degree of autonomy, agent composition, recipients, geographic routing or deployment context may alter the purposes, means or risks of processing. Such changes should be identified before implementation and should trigger reassessment of data protection and privacy-related risks where they may affect the basis or conclusions of an earlier examination.
- ***Decommissioning and retirement:*** The end of operation should include the termination of access, revocation of credentials and permissions, and the secure return, deletion or lawful preservation of personal data, memory and operational records. Measures should prevent continued processing by connected tools, services or agents after the system has been retired.

These characteristics may amplify risks of excessive collection, cross-context linkage and inferential profiling; persistent or inaccurate memory; unauthorised disclosure through tools; identity or credential misuse; propagation of errors between agents; and actions capable of producing significant or irreversible effects when based on probabilistic outputs. Access to accounts, files, communications, sensors or external services may further increase informational asymmetry and reduce the practical ability of data subjects to foresee, understand and exercise control over processing.[3;4]

Beyond this lifecycle approach, LLM-based systems may have broader implications for human rights, democracy and the rule of law that call for appropriate scrutiny and,

---

[3] Information Commissioner's Office, ICO Tech Futures: Agentic AI — Data protection and privacy risks, 2026

[4] *Beduschi A.,* Data protection in the era of agentic artificial intelligence, Computer Law & Security Review, vol. 61, 2026, 106342, DOI: 10.1016/j.clsr.2026.106342.

where applicable, broader impact assessment.[5] Privacy and data-protection risks may also be mapped at model and system levels.[6] The integration of LLM-based systems with diverse and rapidly evolving technologies may enable persistent, adaptive, multimodal and autonomous interactions, continuous monitoring, profiling and advanced inference, thereby increasing informational asymmetries and testing the adequacy of existing safeguards. Such risks should be addressed within the broader risk- and impact-management framework for human rights, democracy and the rule of law.[7]

/

LLM is a machine learning model trained on a massive amount of data and designed to process and generate natural language-like outputs by predicting sequences of tokens. Given its technical characteristics, it serves as a foundational architecture for building LLM-based systems, which creates a relatively hierarchical structure between the model and the system. An LLM provides general-purpose language abilities for autonomous systems pursuing complex objectives and sequences of actions. While not every LLM-based system is agentic, the latter is built on one or multiple AI models, driven towards specific objectives and equipped with autonomous features allowing for planning, decision-making, memorising (including through external resource tools) and dynamic functioning by adjusting its actions[8]. Given this hierarchical correlation, privacy-related risks emerging at the model level are reflected and amplified in systems.

From data protection perspective, personal data is processed across activities within the lifecycle[9] of LLM models and LLM-based systems, giving a rise to privacy- and data protection related risks in consideration of its application context, design and development context and deployment context[10]. With an objective of securing individual's right to protection of their personal data and contributing to respect for their human rights and fundamental freedoms, in particular to the right to privacy, and pursuant to Article 3 of the Convention 108+, the applicability of thereof extends to each lifecycle phase of AI technology.]

IV GENERAL GUIDANCE ON APPLICATION OF FUNDAMENTAL DATA PROTECTION PRINCIPLES AND SPECIFIC ARTICLES

---

5 See, Council of Europe risk and impact assessment framework of AI systems from the point of view of human rights, democracy and the rule of law (HUDERIA); Also, for structural implications described in the Guidance Note on Generative AI implications for Freedom of Expression by the Council of Europe MSI-AI Committee.

6 See, Expert Report on "Privacy and Data Protection Risks in Large Language Models (LLMs), pp. 11-20.

7 Article 16, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225).

(1) Designation of Roles and Responsibilities

Throughout their lifecycle, LLM and LLM-based systems operate on massive datasets, which creates chain of data processing activities. Identification of relevant actors throughout this chain is crucial to understand the concentration of data controllership and allocation of decision-making powers under the Convention 108+.

Roles and responsibilities of relevant actors involved in the processing activities depend on the processing actually carried out and should be assessed on a case-by-case basis. They may be distributed or interdependent across several actors throughout the lifecycle of LLM ecosystems, from model development to deployment and user interaction. Their determination should reflect factual decision-making power over purposes and means rather than contractual or technical labels.

Agentic arrangements may additionally involve an agent developer or orchestrator, a deployer, providers of tools and data services, operators of other agents, and the person or organisation on whose behalf the system acts. Descriptions such as “agent”, “assistant” or “autonomous” do not determine legal status under the Convention. The use of an AI agent does not alter roles and responsibilities of the natural or legal persons, public authorities, services, agencies or other bodies that determine the purposes and means of processing, or process personal data on their behalf.

Where several actors determine purposes or means, the existence and scope of joint controllership should be assessed in accordance with the Convention and applicable law. In such assessments clear distinctions should be made between specific actors based on the capacity or the power of determining the main characteristics and conditions of the data processing. Arrangements between actors should allocate in particular, without prejudice to data subjects’ rights, responsibility for authorising tools and permissions, defining memory and retention, maintaining relevant action records, handling rights and breaches, managing downstream processors and transfers, and suspending the agent. Contractual allocation should not be treated as conclusive where it does not correspond to the factual circumstances.

The determination of data controllers and processors (where relevant joint controllers and sub-processors) and allocation of respective roles and responsibilities among them is directly intertwined with the accountability and demonstrability of the compliance with the Convention (Article 10).

Recommendations:

- *Relationships between relevant actors should be delineated and documented in line with the Convention through clear contractual terms and operational responsibility matrices covering the whole processing lifecycle, including controllers’ instructions to processors, tool and permission governance, memory and retention, security monitoring, incident handling, data subjects’ rights, downstream services and transborder flows.*

(2) Categories of Personal Data

The efficient data governance embarks with clear understanding as to what personal data constitutes and how it can be distinguished from non-personal information, especially, in consideration of initial lifecycle stage of LLM ecosystems. This includes determination of the category of personal data, which is an essential requirement to determine prior to the commencement of the processing activities the lawful purposes of the processing and can also inform the sensitivity of processing activity (Article 6) and furthermore support effective privacy impact assessment efforts (Article 10). Throughout the development and deployment phases it is necessary to assess and monitor the identifiability potential of the processed data to prevent identification of individual.

The inclusion of personal data in LLM training datasets creates risks for model memorisation and their unintended disclosure in outputs. Timely anonymisation and removal of personal data from training datasets – where not feasible, applying pseudonymisation, can reduce the risk of re-identification throughout the whole lifecycle of the system or inadvertent leak of personal data in the system's outputs. Data controllers and where relevant, processors should embed privacy-enhancing designs, [such as synthetic data, federated machine learning models, hard encryption, etc.] in systems architectures and extend best efforts in proactive detection and removal (filtration) of personal data prior to model training.

Throughout the lifecycle and having regard to the application context, LLM-based systems may process, generate, access or infer special categories of personal data within the meaning of Article 6 of Convention 108+[11]. Such processing is permissible only where appropriate safeguards are enshrined in law, complementing those of the Convention. Inferences require the same attention where they reveal information covered by Article 6, including when they are probabilistic or subsequently shown to be inaccurate.

In agentic systems, personal data may also be contained in plans, parameters supplied to tools, retrieved context, communications between agents, intermediate outputs, observations, records of actions, operational records and short- or long-term memory. A technical designation or treatment as an internal system artefact does not remove such data from the scope of the Convention where they relate to an identified or identifiable individual.

Recommendations:

- *Relevant actors should identify and classify the categories of personal data and the phases and components in which they are processed, embed appropriate safeguards from the earliest stage and apply privacy-enhancing techniques where suitable. The use of synthetic data should not be treated as automatically anonymous or risk-free: linkability, memorisation, disclosure and the possibility*

[11] Special categories of data are: genetic data; personal data relating to offences, criminal proceedings and convictions, and related security measures; biometric data uniquely identifying a person; personal data for the information they reveal relating to racial or ethnic origin, political opinions, trade-union membership, religious or other beliefs, health or sexual life.

of singling out individuals should be assessed in a given context in line with Convention 108+[12].

- *Controllers and, where applicable, processors should exclude special categories of personal data from collection, curation, retrieval, memory and tool access by default unless their processing is necessary for a specified purpose, rests on a legitimate basis laid down by law and is accompanied by appropriate safeguards under Article 6 complementing those that have been put in place for non-special categories of personal data. They should also prevent unnecessary sensitive inferences and verify the provenance, uncertainty, relevance and accuracy of any necessary inference before it informs an action or decision.*

(3) Legitimacy of Data Processing

LLM-based systems rely on training with large volumes of data to optimise their performance. It is essential that each stage of operation — including data collection, processing, and deployment — is undertaken based on an appropriate legal ground and in compliance with requirements for the legitimate processing under Convention 108+.

In a lifecycle context, each controller should identify and document the legitimate basis laid down by law for each processing purpose and stage. Depending on the circumstances, processing may be based on the data subject's free, specific, informed and unambiguous consent or on another legitimate basis laid down by law, in accordance with Article 5, paragraph 2, of Convention 108+. The availability of a basis should be assessed separately for model creation, post-training, system integration, deployment, interaction, memory, evaluation and any later model improvement.

Difficulties in obtaining valid consent for model creation, training, post-training or fine-tuning do not, in themselves, justify the processing of publicly accessible personal data. For each purpose and lifecycle stage, the controller should identify and document a legitimate basis laid down by law in accordance with Article 5, paragraph 2, of Convention 108+. The commonly deployed techniques for data collection and curation ─ for instance, web scraping, [13] utilisation of proprietary data, data from user interactions, secondary and inferential processing ─ at the initial or advanced stages of the lifecycle of the system can create tensions with legal requirements set by Article 5 of Convention 108+.

Public accessibility of data does not automatically create a legal basis for the processing, while particular data sources must be excluded from processing. Considering the scope, gravity and long-term effects on data subjects' rights, such as repurposing of publicly available data, challenges related to individual's awareness and control over their data, reasonable expectation along with the ability to exercise

[12] Article 19 of the Explanatory Report of the Convention provides that "Data is to be considered as anonymous only as long as it is impossible to re-identify the data subject or if such re-identification would require unreasonable time, effort or resources, taking into consideration the available technology at the time of the processing and technological developments."

[13] EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0, 2026.

their privacy rights,[14] such processing may fail to demonstrate overriding legitimate interest(s) of the data actors within the scope of Article 5, paragraph 2, if not demonstrated: (a) Lawfulness of the pursued interest[15]; (b) Necessity of processing with respect to the pursued interest; (c) Proportionality of the pursued interest with respect to the rights and freedoms at stake.

Personal data obtained through user interactions or agentic workflows should not be repurposed for performance refinement, model improvement or further training unless the controller has established a compatible specified purpose and an appropriate legitimate basis under the Convention and applicable law. Where data subjects' consent is relied upon, it should be free, specific, informed and unambiguous, expressed through an affirmative opt-in and capable of being withdrawn as easily as it was given. The controller should ensure that withdrawal mechanisms are effective and that subsequent processing is brought into conformity without affecting processing lawfully carried out before withdrawal.

Each retrieval, memory operation, invocation of a tool, disclosure or delegation entailing the processing of personal data should remain within the documented purpose and legitimate basis. A broad task instruction cannot legitimise unrelated downstream processing, the autonomous adoption of a new purpose or recipient, or access to personal data relating to persons other than the user beyond what is necessary and lawful.

These considerations are amplified in case of processing personal data related to children and individuals belonging to vulnerable groups, where proper designs, including age-verification practices, should be enabled by default.

Recommendations:

- *Data processing activity should be designed in a manner that ensures the traceability of each operation, adherence to legal basis and legitimacy requirements established by Convention 108+ and incorporates proper safeguards against of violations of data subject's rights. Throughout system's lifecycle, the data controllers and, where relevant, processors should map out, document processing activities and ground on proper legal basis.*
- *Controllers and processors should refrain from unlawful processing practices, including unsolicited marketing, unlawful profiling or behavioural surveillance and should assess the corresponding impacts on privacy, data protection and other human rights.*

(4) Application of the Basic Principles for the Protection of Personal Data

[14] Satakunnan Markkinapörssi Oy and Satamedia Oy v. Finland [GC], 2017, § 137; L.B. v. Hungary [GC], 2023, § 103.

[15] Based on EDPB guidelines on legitimate interest, a legitimate interest should comply with the following cumulative criteria according to which the interest should be regarded as lawful, the interest should be clearly and precisely articulate and be real and present, EDPB Guidelines 1/2024 on Processing of Personal Data based on Article 6(1)(f) GDPR, Version 1.0, 2024, para. 28-30.

a. Principle of Necessity and Proportionality

Processing should be necessary and proportionate in relation to a specified, explicit and legitimate purpose. Controllers should carefully consider the interests of all persons likely to be affected; the development and use of new technologies entail a particular responsibility to strike a fair balance among those interests. In accordance with Article 5, paragraph 1 of Convention 108+, processing should not cause unlawful interference with the rights and fundamental freedoms of data subjects and should be limited to what is provided by Article 11 of the Convention as lawful exceptions. Controllers should assess whether the processing is suitable and essential for the purpose and the if the purpose can reasonably be achieved by less intrusive means and should apply this assessment throughout the lifecycle, in close conjunction with data minimisation.[16] Processors should provide the information and assistance necessary for these assessments, where applicable.

The principle of necessity and proportionality should be demonstrated throughout the entire lifecycle of the foundation models and LLM systems by continuous evaluation and reflection of the fair balance between competing interests. This effort includes clearly articulating the legitimate purpose, assessing less intrusive alternatives in relation to achieving equivalent outcomes, limiting the use of personal data to what is strictly necessary for training, fine-tuning and adaption, and ensuring ─ by default ─ that personal data processed during user interactions is no more than necessary for providing outputs.

b. Fairness and Transparency

Fair and transparent processing of personal data, as set out in Article 5, paragraph 4, litterae “a” of Convention 108+, applies throughout the lifecycle of LLM and LLM-based models. Data actors should ensure that data processing neither result in unlawful interference with data subject’s rights and freedoms, giving a rise to a wider adverse consequence (such as discriminatory practices and bias). Data controllers, and where applicable, processors should strike a fair balance between all interests concerned, whether public or private, and the rights and freedoms at stake and assess whether the personal data has been collected and processed in a manner that respects the reasonable expectations to privacy of data subjects.

For agentic systems, necessity and proportionality should be assessed in relation to each relevant retrieval, inference, addition of information to memory, invocation of a tool, disclosure and action, as well as for the system as a whole. The generality or adaptability of an agent does not justify open-ended purposes, indefinite operation or access to personal data and tools which may merely be useful for future tasks. The least intrusive effective degree of autonomy and narrowest effective range of permitted actions should be selected.

The transparency obligations under Article 8 and the accountability requirements under Article 10 apply throughout the lifecycle of an LLM-based system. The opacity of training pipelines, post-training adaptation, retrieval mechanisms and downstream

[16] EDPB Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models, 2024, para. 70-75.

integrations may make it difficult for data subjects to understand the nature and sequence of processing operations. Controllers and, where applicable, processors should therefore implement effective transparency measures that make relevant processing intelligible and support meaningful explanations. In the broader AI governance context, transparency entails openness and clarity regarding governance, relevant system operations and decision-making processes for the actors and stakeholders concerned.[17] This includes providing appropriate information about the data used throughout the system's lifecycle and the measures taken to protect personal data.[18] Where relevant, AI-generated content should be identified in a clear and effective manner.[19] The fair processing entails, inter alia that data controllers strive to reconcile the fundamental values of the respect for privacy and the free flow of information across borders and between peoples, do not rely on power imbalances when determining the main characteristics and conditions of data processing and ultimately respects everyone's right to informational self-determination.

Pursuant to Article 8 of Convention 108+, transparency of processing requires clear and user-friendly communication of identity of data actors involved in processing, adequate and relevant information regarding stages of processing activities, datasets, including, their sources, legal basis and purposes of processing, categories of personal data processed. Where interactions with an LLM may be retained, analysed or re-used for model improvements, this should be communicated in an accessible manner, prior to processing. In the context of LLM-based systems, data controllers should ensure that all data recipients are identified and communicated along with information regarding internal reasoning processes and any changes to retention periods, automated decision-making, cross-border data flows or new processing purposes. Data controllers should provide meaningful explanations regarding the role of the LLM models in the processing activity, the nature and purposes of the processing, the principal characteristics and limitations of the system, as well as the safeguards implemented to reduce risks to individuals. Transparency should be supported with proper documentation aiming at informing all data subjects categories (including potential users).

Under Article 5, paragraph 4, litterae "a" and Article 8 of Convention 108+, transparency should be operational where it supports fair and transparent processing and equips data subject with proper awareness regarding the privacy-related risks, available safeguards, accessible remedies and the exercise of their rights. Privacy notice regarding the data processing should be designed with user-friendly interfaces, providing contextual information and practical mechanisms that enable individuals to understand how processing operations may affect them and how their personal data are used across the lifecycle of LLM-based systems.

c. Purpose Limitation and Data Minimisation

[17] Article 8, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225).

[18] Explanatory Report, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), para 57.

[19] Article 8, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225).

Broad data collection at the initial stage of model creation may conflict with purpose limitation and data minimisation principles as enshrined in Article 5 of Convention 108+. Data controllers should clearly define and document the specific, explicit and legitimate purpose for which personal data is processed at each phase of the system lifecycle (Article 5, paragraph 4, litterae “b”). Assessment of the legitimacy of purpose should be supported by balancing rights, freedoms and interests at stake is made in each instance; the right to the protection of personal data on the one hand, and the protection of other rights on the other hand.[20]

In consideration of the model development, adaptation and improvements, the compatibility of further processing with original purpose should be continuously assessed by taking into account, *inter alia*, the relationship between the original and subsequent purposes, the context in which the data was initially collected, the reasonable expectations of data subject, the category of personal data and its sensitivity, along with the potential adverse impacts on data subject’s rights and freedoms. Data initially collected for limited and context-specific purposes may subsequently be repurposed, aggregated, or re-used for optimisation, behavioural prediction, personalisation, or inferential analysis in ways that exceed the original purposes of collection or the reasonable expectations of data subjects. Where LLM-based system operates within the scope and purpose of an existing processing activity, no separate legal basis is required, however, if additional processing is being carried out for new purposes, data controllers should identify an appropriate legal basis under Article 5, para 1 of Convention 108+.

Data minimisation should be implemented by design at the initial stage of LLM ecosystems and observed throughout its whole lifecycle. LLM systems should be designed to access and process only the personal data strictly necessary to achieve specific purpose, while considering privacy-preserving alternatives (for instance, by processing synthetic or anonymised data, using pseudonymised data where possible, filtering special category of personal data, etc.). Principles of purpose limitation and data minimisation should be implemented in a manner capable of addressing not only direct data collection practices, but also increasingly advanced forms of inferential processing, multimodal aggregation, behavioural prediction, and adaptive profiling.

Appropriate technical and organisational measures should minimise processing so that personal data are adequate, relevant and not excessive in relation to the legitimate purpose pursued, in accordance with Article 5, paragraph 4, litterae “c”. This requires an assessment of the categories, sources, volume, granularity, duration and recipients of personal data at each relevant lifecycle stage.

In an agentic workflow, purpose limitation and data minimisation should be technically enforceable for each retrieval, message, memory operation, invocation of a tool and disclosure. Open-ended access to entire mailboxes, drives, contact lists, databases or accounts should not be granted where access limited by field, record, purpose, time or task would suffice. Delegation to another agent or service should not silently broaden the purpose, categories of personal data, recipients or permitted actions.

[20] Explanatory Report to the Protocol Amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, para 48.:

d. Storage Limitation

Pursuant to Article 5, paragraph 4, litterae “e”, where personal data is no longer required to achieve the identified purposes, it should be securely deleted or preserved in a form which does not permits identification or re-identification of data subject. Data controllers should establish documented retention periods for specific categories of personal data processed, the purpose(s) for which data is retained and criteria governing deletion or irreversible de-identification, which requires iterative review and evaluation.

Particular attention should be given to risks begotten from data memorisation, as foundational models may memorise and reproduce personal data represented in their training data. Additionally, data memorisation challenges data quality standards and affects data subject’s ability to effectively exercise their rights under Article 9 of Convention 108+. These concerns are amplified in relation with LLM-based system’s memory functions. Data controllers should be able to demonstrate that memory systems and memorisation risk are adequately assessed, addressed through mitigation mechanisms and undertake a subsequent extraction of memorised data. Where data removal is not possible due to technical limitations, data controllers may consider implementation of additional safeguards (such as model retraining, machine unlearning techniques, etc.) in response to potential risk impacts.

In the context of user interactions, controllers should ensure that prompts, conversation history and generated outputs containing personal data are retained only where necessary for specified purposes, such as compliance with a legal obligation, security or accountability. Processing should be lawful, fair and transparent, and retention should not be made conditional merely on a data subject’s presumed knowledge.

Agentic systems may use context windows, working or session memory, persistent user profiles, organisational memory, vector stores, caches and external records. Controllers should determine separately the purpose, legitimate basis, access conditions and retention period for each memory layer. Persistent memory should be disabled by default unless continuity is necessary for a specified purpose and the resulting risks are appropriately mitigated; for high-impact tasks, purpose and user or case segregation should be applied by design.

Where applicable, data subjects should be able to locate, inspect, rectify and erase personal data held in memory. Deletion and rectification measures should address relevant conversation histories, embeddings, vector stores, caches, replicas, shared memories and downstream recipients, and should prevent outdated or erased data from being reintroduced from stale copies. Memory entries should, where appropriate, carry provenance, time, purpose, quality and expiry information.[21;22]

Logging should support accountability and security without becoming an additional source of disproportionate processing. Relevant events, including data sources

[21] AEPD, Agentic Artificial Intelligence from the Perspective of Data Protection, version 1.2, 2026, sections concerning short- and long-term memory and the exercise of data-subject rights.

[22] *Wang Bo et al.,* Unveiling Privacy Risks in LLM Agent Memory, Proceedings of ACL 2025, DOI: 10.18653/v1/2025.acl-long.1227, https://aclanthology.org/2025.acl-long.1227/.

accessed, agents and tools invoked, disclosures and modifications, recipients, delegated authority, human approvals, outcomes and errors, should be recorded in a retrievable form subject to strict access and retention controls. Internal model reasoning traces should not be retained or disclosed by default; where they contain personal data, they remain subject to the Convention and may themselves create security and privacy risks.[23]

e. Data Quality:

Data controllers should ensure that training datasets remain in conformity with the data quality requirements set by Article 5, para. 4, litterae “d” of Convention 108+. In particular, “personal data undergoing processing should be accurate and, where necessary, kept up to date”, otherwise, data controllers are required to delete inaccurate data and refrain from further processing. In the context of LLM ecosystems, this requirement applies to all stages of its lifecycle, covering the system’s outputs, inferential processing practices, and downstream uses of probabilistic information concerning identifiable individuals.

Data controllers and processors should rely on accurate, reliable and representative data, considering the long-term impact of training data lacking these qualities. In particular, LLM-based system’s training data may include biases, which can be further reflected in discriminatory or inaccurate outputs. Associated risks encompasses, but are not limited to, inaccurate statements about identifiable individuals, including false allegations, fabricated or misleading contextual information, “hallucinations” or the generation of plausible, but inaccurate personal information in outputs, which may have long-term adverse impacts on data subject’s rights and freedoms; bias amplification, where underlying statistical associations reproduce or reinforce unfair or discriminatory patterns relating to individual(s) or groups.

By design, data controllers and, where applicable, processors should implement effective measures supporting data content verification and ensuring its accuracy as well as representativeness. This furthermore entails using filtration techniques in inputs and outputs, indexing of data resources, excluding unverified sources from training datasets (especially relevant in terms of data scaping)[24], iterative evaluation and monitoring (testing) system’s outputs, including inferences. Given the probabilistic abilities of LLM-based systems, meaningful monitoring in this regard refers to human oversight and ability of the proactive intervention, where needed. Transparent processing, in this context, entails contributing to data subject’s awareness towards system’s probabilistic outputs and its consequential effects. Outputs and inferences should be tested and monitored throughout operation. Where required by the risks, human reviewers should be able to identify errors and intervene effectively. Data subjects should receive intelligible information about the probabilistic nature and relevant limitations of system outputs where this is necessary to ensure fair and transparent processing.

---

[23] *Green Tommaso et al.,* Leaky Thoughts: Large Reasoning Models Are Not Private Thinkers, Proceedings of EMNLP 2025, DOI: 10.18653/v1/2025.emnlp-main.1347, https://aclanthology.org/2025.emnlp-main.1347/.

[24] EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0, 2026, para. 39–42.

Where an output, inference, memory entry or action record constitutes personal data, accuracy safeguards should reflect its intended use and consequences. Before an agent uses probabilistic or externally retrieved information to modify data, communicate with a third party or take an action affecting a person, the information should be verified to a degree proportionate to the risk. Systems should be designed to detect, contain and correct cascading inaccuracies between agents and downstream services.

Recommendations (Regarding applicable principles and data quality standards):

- *A clear allocation of roles and responsibilities among controllers, processors and other relevant actors is essential to the legitimacy of processing and the effective application of data quality requirements. Where relevant, contractual and other arrangements should specify responsibility for data collection, curation, quality assurance and correction during model creation, fine-tuning and subsequent development. Such arrangements should support compliance with purpose limitation, data minimisation and data quality throughout the lifecycle, without altering the allocation of responsibility that follows from the factual circumstances.*
- *Data processing in LLM ecosystems should strictly adhere to fair and transparent processing, where principles of necessity, proportionality, and data minimisation are respected, ensuring that legitimate purposes are clearly articulated, the processing of personal data is essential, less intrusive alternatives are considered, and personal data use is limited throughout the system's lifecycle. Engaging stakeholders and continuously reflecting on the balance between human rights and fundamental freedoms being it public or private domainis essential for responsible data governance.*
- *Data controllers should ensure transparent and fair processing of personal data throughout the LLM system's lifecycle by providing clear, user-targeted privacy notices and transparent notification and effective transparency measures, thereby empowering data subjects to understand, manage and exercise their rights and safeguards. This approach upholds reasonable expectations, mitigates unlawful interference and prevents wider adverse consequences, such as discrimination.*
- *Data controllers should set and regularly review documented retention periods and deletion criteria for personal data, ensuring that memorisation risks are mitigated and user interaction data is only retained where strictly necessary and transparently communicated, along with repurposing for model development permitted only if justified by a legal basis and appropriate safeguards.*
- *Data controllers should ensure, and processors should assist them in ensuring, that personal data used to train or operate LLM-based systems are accurate, reliable and, where relevant, representative in relation to the purpose of processing. Appropriate measures should include source and provenance checks, validation, filtering, periodic testing and meaningful human review proportionate to the context and risks. Where necessary to*

ensure fair and transparent processing, data subjects should receive intelligible information about the probabilistic nature and relevant limitations of outputs.

V. GENERAL GUIDANCE ON DATA SUBJECT'S RIGHTS

The technical characteristics of LLM ecosystems may hinder the effective exercise of data subjects' rights. Controllers should therefore implement, and processors should assist them with, appropriate technical, organisational and procedural measures to give effect to the rights set out in Article 9 of Convention 108+. Technical difficulty alone should not be treated as grounds for restricting those rights. Any exception or restriction should comply with Article 11 of the Convention[25]. Measures should be established from the earliest lifecycle stage and should take account of the context of use, the categories and sensitivity of personal data, and the particular needs of children and *individuals belonging to vulnerable groups.*.

(1) Right Not to Be Subject to Automated Decision-Making

LLM-based systems are increasingly integrated into automated and semi-automated decision-making environments. Article 9, paragraph 1, litterae "a" guarantees data subject's right not to be subject to a decision significantly affected them based solely on an automated processing of data without consideration of their views. In the broader AI risk assessment and management frameworks, engagement process of all stakeholders is an element of a paramount importance.[26] For privacy-risk governance efforts, Convention 108+ specifically mandates consideration of data subject's interests. This entails data subject's opportunity to contest, challenge and substantiate the alleged inaccuracy of the processing or other factors impacting their rights and freedoms.[27]

Controllers should ensure meaningful and timely human oversight of automated decisions and agentic actions. Reviewers should have the competence, authority, independence, time and information necessary to understand the relevant factors, withhold approval, take the individual's views into account, pause or override the system and change or reverse the result where appropriate. Measures should address automation bias, alert fatigue and over-reliance. Even where Article 9, paragraph 1, litterae "a", is not engaged, the other requirements of the Convention, including Articles 5, 7, 8, 9 and 10, remain applicable to processing and actions affecting individuals.

(2) Right to Access

[25] For more detail, please see: https://rm.coe.int/t-pd-2021-7rev13-interpretation-of-general-principles-article-11-c108-/1680b6c146

[26] See, Article 16, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225); Council of Europe risk and impact assessment framework of AI systems from the point of view of human rights, democracy and the rule of law (HUDERIA).

[27] Explanatory Report to the Protocol Amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, para 75.

Article 9, paragraph 1, litterae “b”, guarantees every individual the right to obtain, on request, at reasonable intervals and without excessive delay or expense, confirmation of the processing of personal data relating to him or her, the communication in an intelligible form of the data processed, all available information on their origin and on the preservation period, as well as any other information that the controller is required to provide in order to ensure the transparency of processing in accordance with Article 8, paragraph 1. In the context of agentic systems, mechanisms for exercising the right of access should be capable of identifying and retrieving relevant personal data contained in prompts, retrieved context, short- and long-term memory, vector stores, records of actions and connected services, as well as tracing disclosures of such data to tools or other agents. Data controllers should establish appropriate coordination arrangements with processors and other entities involved in the processing, in accordance with their respective roles and responsibilities, to ensure that data subjects can exercise their rights effectively without having to identify and contact each entity in the processing chain separately.

Article 9, paragraph 1, litterae “c”, guarantees every individual the right to obtain, on request, knowledge of the reasoning underlying data processing where the results are applied to him or her. In LLM-based and agentic systems, compliance should not be reduced to disclosure of source code, model parameters or internal model reasoning traces. Controllers should provide meaningful, intelligible and context-specific information on the principal factors, personal data and sources relied upon, the role of the model, tools and human intervention, and the sequence of relevant actions leading to the result, together with its significance and foreseeable consequences. Explanations generated by the same system should be validated and should not substitute for documented evidence of the processing.

For agentic workflows, the information should cover relevant uses of tools, delegations, permissions, external data sources and human approvals where they affected the result and should identify the actor responsible at each relevant stage. Any limitation necessary to protect the rights and freedoms of others, security or legally protected secrets should be lawful, necessary and proportionate and should not render the right ineffective. Operational records and other records documenting the provenance and movement of personal data should be distinguished from internal model reasoning traces. Such traces should not be regarded as a reliable audit record or as sufficient, in themselves, to satisfy the right under Article 9, paragraph 1, litterae “c”; their retention or disclosure should be limited to what is lawful, necessary and proportionate.[28]

(3) Right to Rectification, Erasure and Objection

Article 9, paragraph 1, litterae “d”, of Convention 108+ guarantees every individual the right to object at any time, on grounds relating to his or her situation, to the processing of personal data concerning him or her, unless the controller demonstrates legitimate grounds for the processing which override his or her interests or rights and fundamental freedoms. Article 9, paragraph 1, litterae “e”, further guarantees the right to obtain, on request, free of charge and without excessive delay, rectification or

[28] *Turpin M., Michael J., Perez E., Bowman S. R.,* Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, Advances in Neural Information Processing Systems 36 (NeurIPS 2023), DOI: 10.52202/075280-3275;

erasure, as the case may be, of personal data relating to him or her where those data are being, or have been, processed contrary to the provisions of the Convention. The technical characteristics of LLM-based systems may make it difficult to locate, rectify or erase particular personal data in training datasets, retrieval and memory stores, stored outputs, logs and downstream systems. Such technical difficulty does not, in itself, relieve controllers of the obligation to give effect to the rights applicable to the processing for which they are responsible. Data controllers should establish procedures capable of identifying the relevant personal data and determining effective measures for responding to requests. Data subjects should not be required to understand the system architecture or identify the component in which their personal data are held.

Data controllers should take effective measures to give effect to requests for rectification or erasure, objections to processing and, where provided for by applicable law, requests for restriction of processing. Processors should assist controllers in this respect, in accordance with applicable law and the instructions of the controller. Measures should be selected having regard to where and how the personal data are processed. Within the processing for which the controller is responsible, such measures may include correcting or removing personal data from source datasets, retrieval systems, vector stores, memory and caches; suppressing stored outputs containing the data; preventing their reintroduction into the system or further disclosure; communicating rectification or erasure to recipients where required by applicable law; and model editing, fine-tuning or retraining where such measures are necessary, proportionate and demonstrated to be effective. Neither the technical cost of a particular measure nor the controllers' interest in maintaining model performance should, in itself, render an applicable right ineffective.
Where an agent has already disclosed inaccurate or unlawfully processed personal data, modified records on the basis of such data or initiated an action affecting an individual, the controller should ensure prompt human review, take measures to prevent or mitigate further adverse effects, notify recipients where required by applicable law and, where feasible, reverse the action or otherwise remedy its effects. Procedures should also prevent inaccurate, erased or superseded personal data from being reintroduced into the system from persistent memory, out-of-date copies or other agents.

(4) Access to Remedies and Assistance

Pursuant to Article 9, para. 1, litterae "f" and litterae "g", data controllers and where applicable, data processors deploying or operating LLM-based systems should facilitate the effective exercise of data subject's rights by providing accessible procedure for the submission and handling of their requests, cooperating with competent authorities within the meaning of Article 15, and ensuring that data subjects are not deprived of an effective remedy due to operational challenges or technical complexities. In this regard, data subjects enjoy right to lodge a complaint with, request the assistance of and seek for the intervention of the competent authority. They further have access to administrative and judicial remedies in accordance with Article 12.

Recommendations:

- *Without prejudice to any other rights or remedies available under applicable laws, data subjects should be able to exercise their rights in relation to the processing of personal data in the context of LLM-based systems.*
- *Proper implementation of Article 9 of Convention 108+ requires availability of technical and organisational safeguards, procedural measures and accessible mechanisms for exercising data subject's rights, enhanced traceability and documentation practices, clear allocation of responsibilities between data actors, effective and meaningful human oversight procedures, safeguards against excessive inferential profiling, and complaint and redress mechanisms.*
- *Data actors should make appropriate and reasonable efforts to facilitate and give effect to requests made by data subjects in the exercise of their rights. To this end, they should give due regard to the specific needs and best interests of minor, as well as to particular circumstances of individuals belonging to vulnerable groups and should implement effective solutions that ensure the accessible and non-discriminatory enjoyment of their rights.*

VI. GENERAL GUIDANCE ON DATA SECURITY, ACCOUNTABILITY AND TRANSBORDER DATA FLOWS

(1) Data Security

Article 7, paragraph 1 of Convention 108+ requires each Party to provide that data controllers and, where applicable, processors take appropriate security measures against risks such as accidental or unauthorised access to, destruction, loss, use, modification or disclosure of personal data. In the context of LLM-based systems, such measures should cover all relevant components and operations within the processing for which they are responsible, including training, post-training and operational data; retrieval systems and memory; tools and external services; identity and access-management mechanisms; communications between agents; and arrangements for human intervention and oversight. Where a system is capable of initiating actions, the measures should also extend to the components and communications through which those actions are authorised, carried out and recorded.

When determining the appropriate measures, controllers and, where applicable, processors should have regard to the nature and sensitivity of the personal data, the volume, scope, purposes and duration of the processing, the length and complexity of the processing chain and the degree of interconnection between the relevant components. Particular attention should be paid to vulnerabilities arising from the architecture and use of LLM-based systems, including those affecting training data, model memorisation, retrieval mechanisms and application programming interfaces, as well as risks of unintended disclosure or mixing of personal data between contexts, users or sessions. Such risks may be increased where systems are interconnected or operate with a high degree of autonomy and are granted access to external services, data sources or operational environments.

To give practical effect to Article 10, paragraphs 2 and 3 of Convention 108+, the examination of the likely impact of the intended processing should address general risks to information security as well as vulnerabilities specific to the model and to the system as deployed. Appropriate technical and organisational measures should be incorporated from the earliest stages of design and maintained throughout the period of operation. Where relevant, default configurations should limit access to personal data, tools, accounts, external services and system functions to what is necessary for the specified purposes. The assessment, the measures selected and the reasons for their selection should be documented so that compliance can be demonstrated in accordance with Article 10, paragraph 1. The adequacy and effectiveness of the measures should be tested at appropriate intervals and kept under review in the light of evolving risks and vulnerabilities.

The examination should also address risks arising from the authority and access granted to the system. Such risks may include[29]:

(a) Risks to the confidentiality, integrity, availability and quality of training, post-training and operational data, including the use of unreliable or compromised sources, data poisoning and backdoor attacks capable of affecting model behaviour, personal data or system outputs;

(b) Model inversion, membership inference, the extraction of memorised data and other attacks capable of revealing personal data or enabling inferences to be drawn about individuals from training data, retrieval repositories, memory or records of interactions;

(c) Jailbreak attacks, compromising security restrictions or instructions and prompt injections[30], manipulating model to access and reveal data and direct or indirect prompt injection, including instructions embedded in webpages, files, messages, retrieved content or outputs from tools, which may circumvent applicable safeguards, divert the system from the authorised purpose or task, or cause unauthorised processing or action;

(d) The poisoning of memory, retrieval repositories, outputs from tools or communications between agents, including attacks designed to be activated at a later stage or to propagate manipulated information across agents;

(e) The theft or misuse of credentials; false representation or misuse of the identities of users, agents or services; misuse of delegated authority; unauthorised increases in permissions; and the granting or delegation of access or authority which is broader than necessary or has not been documented;

(f) The unauthorised or unintended use of tools or external communications, including the modification, deletion, transmission, extraction or disclosure

[29] For the general classification of privacy treats against LLMs and LLM agents, see: *Yan B., Xu M., Dong Y., Ren Zh., Cheng X.*, On Protecting the Data Privacy of Large Language Models (LLMs) and LLM agents: A Literature Review, High-Confidence Computing Volume 5, Issue 2, 2025.

[30] This includes indirect prompt injections, where interconnected systems are involved (in the case of RAG and external APIs), leading to data leaks.

of personal data through compromised tools, services or supply-chain components; and

(g) The disclosure or unintended mixing of personal data between users or sessions, the propagation of errors or attacks within multi-agent systems, and insufficient means of containing, interrupting or reversing actions capable of significantly affecting individuals.

Measures should be selected on the basis of the risks identified through the examination carried out under Article 10 and should be proportionate to the particular context. Depending on that context, they may include, *inter alia*:

(i) Operating relevant functions in an appropriately isolated environment and restricting the system to approved models, agents, tools, data sources and endpoints which are, where appropriate, subject to version and change control;

(ii) Assigning unique and verifiable identifiers to agents and other system components for authentication, authorisation and accountability purposes; using credentials which are not shared and, where technically feasible, are non-transferable, limited to the permissions necessary for the specified purpose and valid only for a limited period; and establishing distinct permissions to read, create, alter, transmit, delete or execute;

(iii) Validating the data structures, parameters and outputs used when tools are invoked; separating untrusted content from instructions and other control information; and recording the provenance of data and indicators relevant to their reliability;

(iv) Applying controls to outgoing communications, transborder flows and other disclosures of personal data, together with measures against data loss; and setting proportionate limits on the number of steps and on the duration, volume and frequency of operations, as well as on the recipients concerned;

(v) Providing for checks which are independent of the agent or model concerned and for effective human review or prior authorisation of actions capable of significantly affecting an individual; and

(vi) Monitoring the system during operation; revoking credentials without delay where necessary; providing mechanisms through which operation can be interrupted or safely terminated; and, where feasible, reversing actions or remedying their consequences.

Privacy-enhancing measures and technologies, including techniques intended to achieve anonymisation, differential privacy, federated learning, homomorphic encryption and other forms of secure computation, may contribute to preventing or minimising particular risks where their suitability and effectiveness have been established in the specific context. No such measure or technology should be presumed, by virtue of its designation or technical characteristics alone, to render personal data anonymous, eliminate the risk of re-identification or disclosure, or

ensure that processing complies with the Convention. Its effectiveness and limitations should be documented and kept under review.

Compliance with the security requirements of the Convention should not rely solely on instructions provided to the model, general behavioural constraints or the model's capacity to refuse requests. Such mechanisms should, where necessary, be supplemented by technical and organisational controls whose operation does not depend on the model.

Organisational measures should include arrangements for the prevention, detection and management of incidents; clearly defined responsibilities and escalation and decision-making procedures; the effective involvement of the data protection officer where one has been designated; staff training and awareness-raising concerning data-protection and security risks; and periodic testing of the system as deployed and of the entire relevant processing chain.

Arrangements for the management and notification of data breaches should enable the controller to comply with Article 7, paragraph 2 of Convention 108+. Procedures should enable data breaches and other security incidents to be detected, contained and assessed without delay; preserve only such information as is necessary and proportionate for their investigation; provide for the examination carried out under Article 10 and the safeguards applied to be reviewed; and enable the controller to determine without delay whether a breach may seriously interfere with the rights and fundamental freedoms of data subjects. Where that threshold is met, at least the competent supervisory authority should be notified without delay. Data subjects concerned and other controllers should also be notified where required by applicable law.

Recommendations:

- *Data controllers and, where applicable, processors should implement appropriate and adaptive technical and organisational security measures throughout the lifecycle of an LLM-based system. The measures should address existing and emerging vulnerabilities, including personal data breaches, data and memory poisoning, prompt injection, inference attacks and unauthorised tool use. They should be integrated by design and by default, informed by the assessment of risks and likely impacts, documented, regularly tested and supported by effective incident-prevention and response arrangements, in accordance with Articles 7 and 10 of Convention 108+.*
- *Competent authorities may adopt and implements preventive initiatives, including the awareness raising initiatives, sector-specific policies and guidance and other appropriate measures, aiming at fostering compliance with standards of Convention 108+ and applicable laws.*

(2) Privacy Risk Assessment and Privacy by Design

Article 10 of Convention 108+ requires controllers and, where applicable, processors to take appropriate measures to comply with the Convention and to be able to demonstrate compliance. Controllers should examine the likely impact of intended processing on the rights and fundamental freedoms of data subjects before commencing it and should design processing so as to prevent or minimise the risk of interference with those rights. Measures should be maintained throughout the relevant processing stages. Considering the privacy-related risks introduced by the operational architecture and adaptive, interconnected nature of LLM-based system, which induces the sensitivity of processing activity, *ex ante* privacy risk assessment, regardless of the potential risk level[31], should be seen as one of the primary obligations, mandated under the Convention 108+ and to be undertaken prior to actual commencement of processing. In this regard, data protection impact assessment should be understood as contributory efforts, capable of concluding re-useable evidentiary and accountability outputs for a broader AI system's governance in light of human rights, democracy and the rule of law[32] and without prejudice to impact assessment frameworks developed based thereof.

In the context of LLM-based system, privacy risk assessment should identify, assess and mitigate both LLM model-level and system-level privacy risks, capturing data processing chain throughout the whole lifecycle. According to Article 10, paragraph 2 assessment should be performed at the earliest stage of model creation. In the context of LLM ecosystems, this obligation should be understood as continuous and iterative effort leading to up-to-date evaluation of privacy risks and effectiveness along with proportionality of corresponding mitigation measures. Data controllers, and where applicable, processors, should design and implement organisational measures for iterative review of conducted assessment and deployed mitigation solutions.

Risk assessment efforts should start by mapping out processing operations and case-by-case identification of roles and responsibilities amongst involved data actors throughout the system's lifecycle. Respective data actor (e. g., system developers, suppliers, deployers) ─ within the scope of their role as a controller, should conduct an independent risk assessment for which processing they are responsible. The *ex ante* consideration of the system's application context(s) (e. g, domains and sectors) will allow for prior identification of the sensitivity and gravity of the processing activity. When assessing risk impacts, data controllers, and where applicable, processors should contextually analyse the scope, nature, volume and purpose of the processing operations[33], with observance of fair balance between all interests concerned (principle of proportionality) and in consideration of affected stakeholders' views (e. g., data subjects' engagement in case of automated decision-making).

Data controllers should document identified risks, the measures adopted and the resulting residual risk and should be able to demonstrate that the processing is compatible with the Convention. Risk acceptance alone does not demonstrate

[31] Noting that data processing does not need to likely result in a high risk to individual's rights and freedoms to give a rise to privacy risk assessment obligation under Article 10 of Convention 108+.

[32] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225); Council of Europe risk and impact assessment framework of AI systems from the point of view of human rights, democracy and the rule of law (HUDERIA).

[33] Explanatory Report to the Protocol Amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, para 85, 88, 89.

compliance. Where appropriate, controllers should consult the Data Protection Officer and the competent supervisory authority, in accordance with applicable law. Competent authorities within the scope of Article 15 of Convention 108+ are encouraged to develop sector-specific risk assessment methodologies and templates to facilitate privacy risk identification and management efforts across diverse application context(s) of the LLM-based systems.

Under Article 10, paragraph 2, data controllers and, where applicable, processors should embed data protection by design and by default throughout the system lifecycle, including in architecture, interfaces, default settings and governance. For agentic systems, controls independent from the model should technically enforce task and purpose boundaries; limit access and permissions to what is necessary for the specified purpose and assigned task; ensure segregation and expiry of memory; restrict operation to approved tools and destinations; establish points of effective human review or prior authorisation; limit permitted actions; provide means of suspension or safe termination and, where feasible, reversal or remedy; support effective handling of data subjects' rights and traceability; and ensure secure decommissioning. The mitigation mechanisms should offer technical and organisation measures by design, by default allowing for mitigating identified risks and embedding privacy-enabling solutions, such as data filtration, data minimisation mechanism and privacy-enhancing technologies, at the initial stage of processing. Data actors should consider the nature, likelihood and severity of identified risks, when developing respective mitigation strategies. This includes implementing measures to facilitate the exercise of data subject's rights and redress mechanisms for potentially affected individuals, from the outset.

Where identified risks cannot be adequately prevented or mitigated, data controllers, and where relevant, processors should modify, redesign, restrict, suspend or abandon certain data processing operations, functionalities, deployment environments or system architectures in order to ensure compliance with Convention 108+. Specific considerations should be given to appropriate safeguards for sensitive data processing under Article 6. In support of broader accountability obligation and demonstrability of the compliance with Convention, data actors should be encouraged to periodically publish risk assessment reports, which, *inter alia*, contribute to transparency obligations under Article 8 of Convention 108+.

Controllers and processors should use continuous evaluation, oversight, audit and reassessment to monitor residual and emerging risks throughout the lifecycle. Reassessment should occur where a significant change to the model, system instructions, orchestration, tools, data sources, memory, permissions, degree of autonomy, agent composition, recipients, geographic routing or deployment context affects the purposes, means or risks of processing.

Recommendations:

- *Controllers should assess the likely impact of intended processing at the earliest stage and before processing begins, in accordance with Article 10 of Convention 108+. The assessment should identify and evaluate both model-level and system-level risks and define measures to prevent or minimise*

interference with data subjects' rights and fundamental freedoms. It should be documented, kept under review throughout the lifecycle and renewed where significant changes occur. Processors should provide the information and assistance necessary within the scope of their role and instructions.

- *Data controllers, and where applicable, processors should determine level of identified risk impacts by considering the context-based analysis of the scope, nature, volume and purpose of the processing operations, with an observance of fair balance between all interests concerned and consideration of affected individual(s) views.*
- *Corresponding mitigation strategies should be drawn up to offer proper technical and organisation measures by design and by default allowing for addressing identified risks and embedding privacy-enabling solutions in LLM-based system architectures. Proper risk impact assessment framework involves periodic review of adequacy of the implemented solutions by taking into account the nature, scope, sensitivity and potential adverse impacts of the processing activities. As part of accountability and transparency, data actors should be encouraged to publish initial risk assessment, as well as subsequent re-assessment reports.*
- *Competent authorities are encouraged to render consultations and adopt sector-specific risk assessment methodologies and templates to facilitate privacy risk identification and management efforts across diverse application context(s) of the LLM-based systems.*

(3) Transborder Flows of Personal Data

Transborder flows of personal data in LLM ecosystems may arise at several layers of the lifecycle, including through distributed infrastructure, external cloud services, hosted retrieval-augmented generation components, vector databases, tool services, monitoring providers or agents operated in another jurisdiction. Remote access to a tool, dynamic routing and delegation between agents may constitute a transfer or disclosure even where the user experiences a single interface. Each relevant operation should therefore be identified and assessed according to its factual data flow. Additional risk dimension in regard to appropriate level of protection is created when data transfers occur in jurisdictions of a State which is not a Party of Convention 108+. Within the meaning of Article 14 of Convention 108+, in such environments, transborder flow should be grounded in appropriate level of protection, and only for limited situations, per the derogations afforded by Article 14, paragraph 4.

Subject to safeguards afforded in Article 14 of Convention 108+, each transfer within an LLM system chain should be analysed separately and distinctively. Relevant risks should be addressed through privacy risk assessment, informing on the potential adverse impacts of such transfers. Data actors should assess and consider the adequacy of appropriate level of protection secured by the law or by *ad hoc* or approved standardised contractual safeguards to ensure continuity of adequate protection guaranteed by Convention 108+, irrespective of supply chains, allocation of roles and responsibilities, throughout the lifecycle of processing.

For data exports outside of the jurisdiction of the Parties, the use of CoE Model Contractual Clauses[34] as a tool for ensuring an appropriate level of protection recognised by the Parties of the Convention could be considered.

Recommendations:

- *Data controllers, and where applicable, processors should implement and maintain data governance arrangements allowing for accurate identification of data processing chain and involved actors throughout the lifecycle of the system. This includes mapping and documenting information related to transborder flows and processing chain associated with the operation of the system. Data actors should conduct appropriate risk assessment of intended transfer to determine whether the legal and technical environments of data recipients may affect the appropriate level of protection, notably by non-adherence to international data protection standards as defined by the Convention or third party access to the data in a given jurisdiction.*
- *Data controllers and where relevant, processors should implement contractual mechanisms for establishing enforceable obligations and appropriate level of protection for demonstrating compliance with Convention 108+. In this context, data actors are encouraged to rely on Council of Europe Model Contractual Clauses (CoE MCCs).*
- *Data controllers and where relevant, processors should ensure that data subjects are duly informed, in a clear and accessible manner, where their personal data may be transferred to another jurisdictions, along with the safeguards applicable to such transfers and other relevant information. Specific considerations should be given to transfers involving special categories of personal data or affecting certain categories of data subjects (such as children and individuals belonging to vulnerable groups).*

VII. ADDITIONAL RECOMMENDATIONS CONCERNING AGENTIC LLM-BASED SYSTEMS

Agentic systems may transform a single request into multiple processing operations, including retrieval, inference, storage in memory, invocation of tools, communication with other agents and actions affecting persons who have not interacted with the system. Communications between agents, shared memories, parameters supplied to tools, observations and intermediate outputs may constitute processing of personal data and should not be treated as merely internal technical operations. The categories of personal data, recipients and jurisdictions involved may change during execution. A user's instruction to an agent, or acceptance of an agentic feature, should not in itself be treated as consent to all subsequent processing and does not, by itself, constitute consent to the processing of personal data relating to other persons.

[34] Convention 108 Committee (T-PD) - Data Protection

These characteristics may amplify risks of excessive collection, cross-context linkage and inferential profiling; persistent or inaccurate memory; unauthorised disclosure through tools; identity or credential misuse; propagation of errors between agents; and actions capable of producing significant or irreversible effects when based on probabilistic outputs. Access to accounts, files, communications, sensors or external services may further increase informational asymmetry and reduce the practical ability of data subjects to foresee, understand and exercise control over processing.[35;36]

Controllers and, where applicable, processors should define in advance and restrict by design the scope of an agent's task, its access to personal data, its permissions and tools, the forms of memory it may use and the duration of its operation. They should limit access to what is necessary for the specified purpose and assigned task; distinguish permissions to read from permissions to create, modify, communicate, delete or execute; require effective human review or prior authorisation before actions capable of producing significant or irreversible effects; and provide effective means to suspend or safely terminate the system and, where feasible, to reverse actions or remedy their consequences. Safeguards should not depend exclusively on instructions expressed in natural language or on the model's ability to refuse a request.

Recommendations for controllers and, where applicable, processors concerning agentic LLM-based systems:

a) Controllers and, where applicable, processors should establish, maintain and periodically verify a documented inventory of the system and a map of personal data flows throughout the processing chain.[37]

- *The inventory and map should identify the models and their relevant versions; agents and orchestration components; retrieval components and data sources; temporary, persistent and shared forms of memory; tools and external services; unique and verifiable technical identifiers, credentials and permissions used for access-control purposes; communications and delegations between agents; recipients; locations of processing; transborder flows of personal data; and the points at which effective human review, authorisation or intervention is required or available.*
- *Each component should be linked to the natural or legal person responsible for its operation and to its role in the processing.*
- *The documentation should cover relevant intermediate processing operations and the range and nature of actions that the system is authorised, or can reasonably be expected, to perform. It should not be limited to the initial input and final output.*

---

35 Information Commissioner's Office, ICO Tech Futures: Agentic AI — Data protection and privacy risks, 2026

36 *Beduschi A.,* Data protection in the era of agentic artificial intelligence, Computer Law & Security Review, vol. 61, 2026, 106342, DOI: 10.1016/j.clsr.2026.106342.

37 Cyber Security Agency of Singapore, Securing Agentic AI – An Addendum to the Guidelines and Companion Guide on Securing AI Systems, 17 June 2026, Chapters 2 and 3.

b) Before enabling a relevant processing operation, controllers should determine and document its specified, explicit and legitimate purpose and the applicable legitimate basis laid down by law, in accordance with Article 5 of Convention 108+.

- *A technical objective assigned to an agent, an instruction from a user or the availability of personal data through an authorised tool should not, in itself, be treated as a purpose or legitimate basis for further processing.*
- *Purpose limitation, necessity and proportionality, data minimisation and accuracy should be assessed and enforced, as appropriate, in relation to inputs, retrieved information, inferences, information sent to and received from tools, communications between agents, memory operations, and any disclosure or other action.*
- *Particular attention should be paid to inferred data, special categories of personal data and personal data relating to persons who have not interacted with the system.*
- *Processors should provide controllers with the information necessary to carry out and demonstrate these assessments.*

c) Controllers and, where applicable, processors should define, document and enforce verifiable limits for the operation of each agent.[38]

- *Those limits should specify its task, duration of operation, authorised personal data, tools, recipients, locations of processing and the range and nature of the actions that it is permitted to perform.*
- *Access by each agent to personal data, systems and tools should be attributable, denied by default, and limited to what is necessary for the specified purpose and the assigned task. Permissions to read should be distinguished from permissions to create, alter, transmit, delete or execute, and should be granted separately. This may be achieved, for example, by assigning distinct and verifiable technical identifiers to individual agents.*
- *Credentials used by or on behalf of an agent should be limited to what is necessary for the assigned task and capable of being revoked without delay. They should not be included in prompts, in general-purpose memory or in records which are not subject to appropriate protection.*
- *An agent should not be able to increase its own permissions or confer broader permissions on another agent solely on the basis of a request generated by an LLM-based model.*

[38] Spanish Data Protection Agency (AEPD), Agentic Artificial Intelligence from the Perspective of Data Protection, version 1.2, February 2026, sections VII.E and VII.F; Infocomm Media Development Authority, Model AI Governance Framework for Agentic AI, version 1.5, published 20 May 2026 and updated 5 June 2026, sections 1.1.3, 2.1.2, 2.2.2 and 2.3.1; Eugene Bagdasarian et al., "AirGapAgent: Protecting Privacy-Conscious Conversational Agents", Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, pp. 3868-3882, DOI: 10.1145/3658644.3690350.

- *Authorisation requirements, points of effective human review or prior approval, means of suspending or safely terminating the system, and means of reversing actions or remedying their consequences where feasible should be proportionate to the range and nature of permitted actions; the degree of autonomy; the duration and context of operation; and the likelihood, severity, cumulative nature and reversibility of possible effects.*

d) Controllers should adopt and implement documented rules for each form of memory used by the system, including conversational context, persistent user memory, shared memory, databases used to retrieve information, including vector stores, caches and operational records.[39]

- *The rules should specify what information may be stored, retrieved, combined, corrected or disclosed; its source and reliability; the persons or agents permitted to access it; and the applicable retention and deletion periods.*
- *Measures should prevent the unintended addition of retrieved or generated information to memory, the unintended mixing of personal data relating to different users or contexts, and further use for incompatible purposes.*
- *Controllers should be able to locate personal data in all relevant forms of memory and to give practical effect to the applicable rights of access, rectification, erasure and objection. Where necessary, corrections or deletions should extend to copies, caches, derived records, downstream agents and processors.*
- *Personal data should not be retained solely because they may improve the future performance of the agent.*

e) In accordance with Article 8 of Convention 108+, controllers should provide data subjects with clear, intelligible and readily accessible information about the use of agentic LLM-based systems.

- *The information should address, as appropriate, the identity of the controller; the legal basis and purposes of the processing; the main categories and sources of personal data; the use and duration of memory; the types of tools, services and other agents involved; recipients and transborder flows of personal data; the types of action that the system may initiate; the points at which effective human intervention is available; and the means of exercising data subjects' rights.*
- *Information and, where appropriate, confirmation should be provided at the time when an agent seeks additional access or proposes an action that the person would not reasonably expect.*

---

[39] *Wang B. et al.,* Unveiling Privacy Risks in LLM Agent Memory, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025, pp. 25241-25260, DOI: 10.18653/v1/2025.acl-long.1227; *Liu J. et al.,* Topology Matters: Measuring Memory Leakage in Multi-Agent LLMs, Findings of the Association for Computational Linguistics: ACL 2026, pp. 39728-39746, DOI: 10.18653/v1/2026.findings-acl.1980.

- *Where a workflow produces a decision significantly affecting a person based solely on automated processing of personal data, the safeguards of Article 9 should be ensured. In particular, the person should have a genuine opportunity to express their views and to challenge the decision.*
- *Any person entrusted with review should have the competence, information, time and authority necessary to refuse, suspend or override the proposed action. A merely formal or general request for confirmation should not, in itself, be regarded as effective human review.*

f) Controllers and, where applicable, processors should apply appropriate technical and organisational measures to all interactions with tools and external services, memory operations, communications and delegations between agents.[40]

- *The requesting agent and the intended destination should be authenticated and authorised. The purpose and parameters of a request, and any result returned by a tool or service, should be validated before an action capable of producing significant effects is carried out.*
- *External content retrieved information and messages from other agents should be assessed in light of their origin and reliability.*
- *Measures should address direct or indirect manipulation of model instructions, including prompt injection; poisoning of retrieval sources or memory; unauthorised access to, disclosure or extraction of personal data; unauthorised increases in permissions; and the propagation of errors between agents or components.*
- *Where a tool is capable of producing significant or difficult-to-reverse effects, controllers should limit the systems, personal data and destinations accessible through it and should remain able to control the disclosures and transfers it may bring about. Such measures should be proportionate to the severity and reversibility of the possible effects.*
- *Delegation to another agent should not extend the original purpose, access to personal data, permissions, retention periods, locations of processing or requirements for human review.*
- *Security should not depend solely on instructions expressed in natural language, general behavioural constraints of the model or the model's ability to refuse a request.*

g) Controllers and, where applicable, processors should maintain secure and proportionate operational records that enable relevant sequences of processing

---

[40] *Greshake K. et al.,* Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection, Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 2023, pp. 79-90, DOI: 10.1145/3605764.3623985; *Zhan Q. et al.,* InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents, Findings of the Association for Computational Linguistics: ACL 2024, pp. 10471-10506, DOI: 10.18653/v1/2024.findings-acl.624; *Debenedetti E. et al.,* AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents, NeurIPS 2024, DOI: 10.52202/079017-2636; *Chen Zh. et al.,* AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases, NeurIPS 2024, DOI: 10.52202/079017-4136.

operations and actions to be traced and, where necessary, reconstructed, and that support demonstration of compliance.

- *The content and level of detail of the records should be proportionate to the risks of the processing and to what is necessary to trace and, where required, reconstruct the relevant sequences of processing operations and actions. Depending on the context, the records may relate, among other matters, to the agent, model and configuration concerned; the assigned task; the data sources and permissions used; relevant retrievals, uses of tools and communications between agents; recipients and locations of processing; actions attempted, performed, prevented or reversed; and human approvals or interventions. The records should themselves be subject to purpose limitation, data minimisation, access controls, safeguards ensuring their integrity and defined retention periods. They should not routinely reproduce the full content of personal data where references or limited metadata are sufficient.*
- *Internal model reasoning traces should not be regarded as a reliable audit record or, in themselves, as sufficient to provide the knowledge of the reasoning underlying data processing referred to in Article 9, paragraph 1, subparagraph 'c'. Any retention or disclosure of such traces should have a legitimate basis laid down by law, serve a specified purpose, and be necessary and proportionate.*
- *Operational records should support the exercise of data subjects' rights, internal accountability, independent audit and investigation by supervisory authorities.*

h) Before deployment, at appropriate intervals and after significant changes, controllers and, where applicable, processors should test the complete system as deployed and representative sequences of its processing operations and actions. Evaluation of the underlying model alone is not sufficient.

- *Testing should cover normal operation, reasonably foreseeable misuse and relevant adversarial conditions. It should include excessive retrieval of personal data; attempts to manipulate model instructions; unauthorised extraction or disclosure of data; unauthorised use of tools or increases in permissions; disclosure between agents or users; inaccurate or poisoned memory; propagation of errors; failure of human review or intervention; and actions affecting persons who have not interacted with the system.*
- *Measurable criteria for acceptable performance and risk, thresholds requiring escalation, and conditions requiring suspension or termination should be defined in advance.*
- *Where the intended processing is likely to have a significant adverse impact on the rights and fundamental freedoms of individuals, the examination of its likely impact referred to in Article 10, paragraph 2, of Convention 108+ should not be carried out solely by persons directly responsible for the design or deployment of the system. Controllers and,*

where applicable, processors should involve persons with appropriate expertise and sufficient independence of judgement and should document the arrangements put in place to ensure the objectivity of the assessment and to address potential conflicts of interest.

- *After deployment, monitoring should be capable of detecting significant changes in behaviour, unexpected delegation or use of tools, access violations, recurring errors and ineffective human review. Monitoring should not entail unnecessary or incompatible further processing.*

i) Controllers and, where applicable, processors should establish and regularly test procedures for managing incidents involving agentic LLM-based systems.[41]

- *The procedures should enable the controller, without delay, to suspend the operation of any agent involved, contain the effects of the incident, prevent their propagation to other agents or recipients, and preserve the information necessary to assess and investigate the incident. The measures taken should be proportionate to the nature, severity, scale and reversibility of the actual or reasonably foreseeable effects.*[42]
- *The procedures should also enable the controller to identify the individuals and personal data affected, the systems involved and the jurisdictions in which the data have been processed or to which they have been transferred; rectify inaccurate personal data held in memory or downstream records; reverse actions or, where this is not feasible, remedy their consequences; identify and address the underlying cause; and notify other relevant controllers, supervisory authorities and the data subjects concerned where required by the Convention or applicable law.*[43]
- *Processors should inform the controller without delay of any personal data breach or other security incident, and of any significant departure from the specified or reasonably expected operation of the system, that may affect the protection of personal data or the controller's ability to comply with Convention 108+ and applicable law.*[44]
- *The assessment of the system as deployed and of the safeguards applied should be reviewed at appropriate intervals and updated where necessary. A further review should be carried out after an incident, a significant departure from the intended or expected operation of the*

---

41 *Shao Y. et al*., PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action, NeurIPS 2024, DOI: 10.52202/079017-2837; *Zharmagambetov A.* et al., AgentDAM: Privacy Leakage Evaluation for Autonomous Web Agents, Advances in Neural Information Processing Systems 38, NeurIPS 2025, Datasets and Benchmarks Track; Infocomm Media Development Authority, Model AI Governance Framework for Agentic AI, version 1.5, published 20 May 2026 and updated 5 June 2026, sections 2.3.2 and 2.3.3; NIST, Incident Response Recommendations and Considerations for Cybersecurity Risk Management, SP 800-61 Rev. 3, April 2025, DOI: 10.6028/NIST.SP.800-61r3.

42 Depending on the architecture of the system and the nature of the incident, such measures may include isolating affected components, revoking access credentials, suspending access to tools or restricting communications with other components or external services.

43 Convention 108+, Article 7, paragraph 2, and Article 9, paragraph 1, litterae "e".

44 Convention 108+, Article 7, paragraphs 1 and 2, and Article 10, paragraph 2.

system or the discovery of a serious weakness, and before any change likely to affect the basis or conclusions of the assessment is introduced. Such changes may concern the model, instructions governing system operation, purposes, data sources, memory, tools, permissions, recipients, locations of processing or the degree of autonomy.

- *Where the identified risks cannot be prevented or reduced to a level compatible with the requirements of Convention 108+ and applicable law, the processing concerned should not commence or resume and, if already under way, should be suspended or brought to an end.*

Recommendations for policy and decision-makers:

a) *States bound by the European Convention on Human Rights should, in accordance with their positive obligations under Article 8, and Parties to Convention 108+ should, within the scope of that Convention, establish, maintain and periodically review a clear, accessible and foreseeable legal and institutional framework providing effective protection against unlawful interferences with private life and the right to the protection of personal data arising from LLMs and LLM-based systems. The framework should cover the entire lifecycle, including agentic arrangements developed or operated by private actors, and should provide for enforceable safeguards, independent supervision and accessible remedies. The measures adopted should be appropriate to the nature, severity and foreseeability of the risks and should not be understood as imposing an obligation to prevent every adverse outcome.*[45]

b) *Parties should ensure that domestic legal and policy frameworks enable the respective roles and responsibilities of the natural and legal persons involved in an LLM-based system to be determined according to their actual functions and decision-making power, rather than solely by contractual or technical designations. Describing a system as an AI agent, or increasing its technical autonomy, should neither confer separate legal personality for the purposes of these Guidelines nor interrupt, obscure or dilute the responsibility of controllers and processors. Delegation within a multi-agent system should remain traceable to the persons or bodies responsible for the relevant processing.*[46]

c) *Parties should ensure that controllers and, where applicable, processors examine the likely impact of intended processing and design it so as to prevent or minimise interference with data subjects' rights, in accordance with Article 10 of Convention 108+. In the case of agentic systems, the assessment should*

---

[45] As regards positive obligations under Article 8 of the European Convention on Human Rights, see X and Y v. the Netherlands, 26 March 1985, paragraph 23; Söderman v. Sweden [GC], no. 5786/08, 12 November 2013, paragraph 78; and K.U. v. Finland, no. 2872/02, 2 December 2008, paragraphs 42-49. See also Convention 108+, Articles 4, 12 and 15, and the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), Articles 1, 3, 4 and 11.

[46] Convention 108+, Article 2(d) and (f), and Explanatory Report, paragraph 22; Spanish Data Protection Agency (AEPD), Agentic Artificial Intelligence from the Perspective of Data Protection, version 1.2, February 2026; *Beduschi A.,* Data protection in the era of agentic artificial intelligence, Computer Law & Security Review, vol. 61, 2026, 106342, DOI: 10.1016/j.clsr.2026.106342.

concern the system as deployed and complete sequences of relevant processing operations and actions, rather than only the underlying model or final output. It should address, as relevant, objectives, data sources, retrieval mechanisms and vector stores, memory, tools and external services, permissions and credentials, delegation and communications between agents, recipients and transborder flows of personal data, points of effective human review, authorisation or intervention, reversibility, cumulative and cascading effects, reasonably foreseeable misuse and persons who did not interact with the system. The assessment should be renewed before significant changes. Processing should not commence or continue where it cannot be made compatible with the requirements of Convention 108+.[47]

d) *Parties should ensure that the supervisory authorities established under Article 15 of Convention 108+ remain fully independent and are provided with the financial, human and technical resources, multidisciplinary expertise and powers necessary to examine both model-level and system-level processing. This should include, where necessary and proportionate, access to relevant documentation, records concerning data flows, provenance and relevant processing operations, inventories of agents and tools, risk assessments, evaluation and security-testing results, operational records, contractual arrangements, information on transborder flows of personal data and incident records. Cooperation with authorities competent for AI market surveillance, cybersecurity, consumer protection, competition, equality and human rights should be organised without reducing the competences or independence of data protection supervisory authorities.*[48]

e) *Public authorities should document the necessity and proportionality of an LLM-based system, its expected benefits and any less intrusive alternatives before procuring, piloting or deploying it. Procurement documents and contracts should address, as appropriate, the purposes of processing; allocation of responsibilities; authorised data sources, agents and tools; restrictions on secondary use and further training; sub-processors and third-party services; retention and deletion; memory and permissions; security testing and audit rights; incident notification; transborder flows of personal data; significant changes; support for the exercise of data subjects' rights; and the return or deletion of data upon termination. Contractual confidentiality or trade-secret claims should not prevent effective supervisory access or the provision of*

---

[47] Convention 108+, Article 10; Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), Article 16; *Barberá I.*, AI Privacy Risks & Mitigations - Large Language Models (LLMs), European Data Protection Board Support Pool of Experts, April 2025; *Shao Y. et al.,* "PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action", NeurIPS 2024, DOI: 10.52202/079017-2837; *Zharmagambetov A.* et al., "AgentDAM: Privacy Leakage Evaluation for Autonomous Web Agents", NeurIPS 2025, Datasets and Benchmarks Track. The empirical findings relate to the systems and tasks evaluated and should not be generalised to every agentic system.

[48] Convention 108+, Article 15, and Explanatory Report, paragraphs 126-130; Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), Article 26.

information necessary for individuals to exercise their rights. Systems should not be procured or renewed where compliance cannot be demonstrated.[49]

f) *Parties should ensure the availability of accessible and effective remedies and procedural safeguards for persons affected by LLM-based and agentic systems. These should include timely and intelligible information, the effective exercise of data subjects' rights throughout the processing chain, means of contesting outputs or actions capable of significantly affecting an individual, and effective human review, suspension, correction and, where feasible, reversal or remedy. An individual should not bear the burden of identifying which model, agent, tool or service within a complex chain caused the contested processing. Where human oversight is relied upon, it should be effective rather than nominal: the reviewer should have the competence, authority, information and time needed to refuse, suspend or override an action, and responsibility for systemic design failures should not be transferred to an end user or nominal supervisor.*[50]

g) *States should ensure that competent authorities have effective legal powers to restrict, suspend or prohibit uses of LLM-based or agentic systems that are incompatible with human rights and fundamental freedoms or with the requirements of Convention 108+. In line with Article 16, paragraph 4, of the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Parties to that Convention should assess the need for a moratorium, ban or other appropriate measure in respect of particular uses. Agentic systems capable of initiating significant or irreversible actions should not be authorised where their authority and access cannot be effectively bounded, monitored and interrupted or where affected persons cannot obtain an effective remedy.*[51]

h) *Parties and supervisory authorities should strengthen international cooperation and mutual assistance concerning LLM-based and agentic systems whose models, infrastructure, data sources, tools or operations span several jurisdictions. Cooperation should include the timely exchange of information on serious incidents and emerging risks, coordinated investigations or joint action where appropriate, and effective enforcement of safeguards governing transborder data flows. The distribution of an agentic system across services or jurisdictions should not lead to a lower level of protection or prevent*

---

[49] Council of Europe, Guidelines on Artificial Intelligence and Data Protection, T-PD(2019)01, section III, paragraphs 1-3; Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), Article 16.

[50] Convention 108+, Articles 9 and 12; Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), Articles 14 and 15; *Saar Alon-Barkat S., Busuioc M.,* Human-AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice, Journal of Public Administration Research and Theory, vol. 33, no. 1, 2023, pp. 153-169, DOI: 10.1093/jopart/muac007; *Green B.,* The Flaws of Policies Requiring Human Oversight of Government Algorithms, Computer Law & Security Review, vol. 45, 2022, 105681, DOI: 10.1016/j.clsr.2022.105681.

[51] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), Article 16, paragraph 4.

identification of the controllers and processors responsible for particular processing operations.[52]

Recommendations for data controllers and, where applicable, to processors:

a) *With an objective of complying with legal requirements enshrined in Convention 108+, data controllers and processors, and where applicable joint controllers, should undertake ex ante privacy risk assessment and management efforts based on the LLM-based system's lifecycle approach, allowing for implementation of appropriate mitigation plan with a focus on technical and organisational measures, as well as privacy-enhancing designs throughout its application, development and design and deployment contexts.*
b) *For agentic systems, the assessment should address the range and nature of actions which each agent is authorised to perform and its degree of autonomy; the duration and possible self-initiation of tasks; tools and credentials; persistence and sharing of memory; access to external or untrusted sources; multi-agent delegation, configuration and communication pathways; cumulative disclosures; points of human intervention; reversibility; and cascading or downstream effects.*

[52] Convention 108+, Articles 16 and 17; Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), Article 25.