Strasbourg,12 May 2026 T-PD(2025)3rev3

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

Introduction

Emerging technologies create significant opportunities for individuals, organisations and public authorities. At the same time, they may generate serious risks for the effective enjoyment of human rights, the functioning of democracy, and the observance of the rule of law. In recent years, the pace of technological development has increasingly reshaped the conditions under which private life is exercised, and personal data are processed, requiring renewed attention to the safeguards that protect individuals in digital environments. In this evolving landscape, 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 rights and freedoms.

LLMs constitute a category of artificial intelligence (AI) designed to process and generate human language at scale. Trained on vast amounts of data and embedded in increasingly complex digital infrastructures, they can be used to automate a wide range of tasks, from content generation and summarisation to decision support and automated interaction. Because their development and operation depend on the large-scale processing of data, which may include personal and sensitive information, LLMs raise significant privacy and data protection considerations. Their technical characteristics, including opacity, probabilistic outputs, and continuous adaptation, may challenge traditional safeguards and complicate the effective application of established data protection principles.

Risks may arise at different points in the lifecycle of LLM-based systems, from model development to system deployment and user interaction. These risks may include unintended retention or reproduction of personal data, insufficient transparency regarding data processing, and difficulties in ensuring meaningful human oversight and accountability. When integrated into automated or semi-automated decision-making environments, such systems may also affect individuals' ability to exercise their rights effectively.

Beyond individual-level impacts, the widespread deployment of LLM-based systems may also have broader societal implications, including effects on autonomy, identity formation, democracy, public discourse and trust in digital environments.

As LLMs continue to expand across sectors such as recruitment, education, healthcare, and public administration, ensuring effective privacy and data protection safeguards becomes increasingly complex. This evolving technological landscape reinforces the urgency of reaffirming and operationalising the principles of Convention 108+ in a manner that remains robust, adaptive, and future proof.

These Guidelines are intended to support the application of 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", "Convention

108+") in the context of LLM-based systems. They supplement, and are not intended to replace, limit or modify, the provisions of Convention 108+. Their purpose is to provide practical guidance on how the Convention's existing principles and obligations apply across the lifecycle of LLM-based systems. In doing so, they seek to promote respect for the right to privacy and the protection of personal data, as enshrined in the European Convention on Human Rights (Article 8). The Guidelines build on the previous work of the Consultative Committee of Convention 108 and related Council of Europe guidance on data protection and emerging technologies.[1]

Article 10 of Convention 108+ is of relevance to these Guidelines. In particular, Article 10.2 requires controllers and, where applicable, processors to examine the likely impact of intended data processing on the rights and fundamental freedoms of data subjects prior to the commencement of such processing, and to design the processing in such a manner as to prevent or minimise the risk of interference with those rights and freedoms.

In the context of Agentic and LLM-based systems, this obligation requires a lifecycle-based assessment of privacy and data protection risks, taking into account both model-level risks, such as training data selection, memorisation, extraction and generation of inaccurate personal data, and system-level risks, such as integration, deployment, user interaction, downstream use, monitoring and post-deployment adaptation.

These Guidelines are also grounded in the human-centric approach reflected in the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225) ("the Framework Convention"), which affirms the need to safeguard human rights, democracy and the rule of law throughout the lifecycle of AI systems, including privacy and personal data protection.

Building on the previous work of the Consultative Committee of Convention 108 and related Council of Europe guidance on data protection and artificial intelligence, these Guidelines provide a coherent and forward-looking framework for identifying, assessing and addressing privacy and data protection risks in LLM-based systems.

Section 1 Purpose and Scope

These guidelines address the privacy and data protection risks associated with the use of LLM-based systems, particularly as they relate to the rights and principles enshrined in Convention 108 and its modernised version, Convention 108+. Such risks may emerge across different phases of the LLM lifecycle of LLM-based systems, including model development, post-training adaptation, integration, deployment, user interaction, monitoring and decommissioning, within broader digital and agentic architectures and may implications not only for data protection, but also for private life, dignity, and autonomy.

The Guidelines present an evidence-based understanding of how LLM-based systems and agentic AI may interfere with individuals' rights and propose a structured methodology to Identify, Assess, Mitigate, and Monitor (IAMM) these risks in real-world settings across diverse stakeholder groups. The IAMM risk assessment methodology is grounded on Article 10.2 of Convention 108+, which stipulates, that where applicable, the data controller should examine the likely impact of intended processing activities on the rights and fundamental freedoms of individuals and should design processing operations in a manner that prevents or minimises risks to those rights and freedoms. The IAMM methodology reflects the principle of proportionality and supports the implementation of accountability obligations under the Article 10.2 of the modernised Convention throughout the lifecycle of LLM-based systems. It is intended to assist stakeholders in identifying and addressing privacy and data protection risks in a structured, iterative, and context-sensitive manner.

The Guidelines adopt a lifecycle-based and risk-based approach reflecting the dynamic and evolving nature of LLM-based systems and recognising that privacy and data protection risks may emerge, evolve, accumulate, or propagate across different stages of the lifecycle. They further recognise that these risks may arise not only from the foundation model itself, but also from the broader technical, organisational, and operational environments in which such models are integrated and deployed.

The Guidelines are grounded in the principles and safeguards established under Convention 108+, including lawfulness, fairness, purpose limitation, proportionality, data minimisation, transparency, accountability, data security, and the protection of the rights and freedoms of data subjects. They also reflect the broader human rights framework of the Council of Europe, including Article 8 of the European Convention on Human Rights ("ECHR"), and the lifecycle-oriented governance approach reflected in the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225) ("Framework Convention on AI").

These Guidelines supplement, and are not intended to replace, limit, or modify, the provisions of Convention 108+. Rather, they aim to clarify how the Convention's principles and obligations apply in the specific technical and organisational context of LLM-based systems and related AI architectures.

1.1. Context and Purpose

LLM-based systems are rapidly transforming the digital landscape by enabling new forms of interaction, automation, and information processing. However, these developments introduce privacy

and data protection risks that challenge traditional legal and technical safeguards. These include, but are not limited to, the inadvertent memorisation and reproduction of personal data, susceptibility to manipulation during inference, and the broader erosion of private life through synthetic identities, profiling, opacity in automated decision-making processes and the extensive use of personal data to support adaptive and action-oriented functionalities within agentic systems.

Persistent governance challenges include difficulties in identifying accountable actors within the value chain in layered AI ecosystems, ensuring meaningful transparency in probabilistic outputs, and guaranteeing the effective exercise of data subjects' rights.

The opacity of training datasets, inference processes, downstream integrations and adaptive system configurations may limit individuals' ability to know whether, how, and for what purposes their personal data have been processed. The absence of observable and accessible data traces, documentation, and effective user interfaces may further limit individuals' ability to exercise their rights in practice, including rights related to access, rectification, objection, and meaningful information concerning automated processing.

All this raises significant challenges for the practical implementation of core principles of Convention 108+, including fairness, transparency, accountability, purpose limitation, effective remedies and safeguard for data subjects' rights. It also raises serious questions as to whether the principles and obligations of Convention 108+ can be effectively upheld in practice without complementary technical and organisational measures adapted to the specific characteristics of LLM-based systems.

Given the extensive deployment of LLM-based and agentic systems across diverse contexts, there is an increasing need to ensure the effective protection and exercise of data subjects' rights under Convention 108+ in light of these technological developments.

The purpose of these Guidelines is therefore to provide a comprehensive approach to privacy and data protection risks under Convention 108+ in the context of LLMs and LLM-based systems. They:

- clarify the application of Convention 108+ principles across the LLM lifecycle of LLM-based systems;
- identify key privacy and data protection risks arising at different lifecycle stages;
- provide a structured methodology for identifying, assessing, mitigating, and monitoring such risks;
- clarify the roles and responsibilities of relevant stakeholders, including providers, deployers, public authorities, supervisory authorities, and end-users; and
- promote coherence with existing Council of Europe instruments and risk-assessment methodologies.

The Guidelines adopt a lifecycle-based methodology reflecting the dynamic and evolving nature of LLMs and LLM-based systems and recognising that privacy and data protection risks may emerge, evolve or propagate across different stages of the lifecycle (Section 2).

Given the Consultative Committee's role and its previous work in interpreting Convention 108+ in the context of emerging technologies, these Guidelines clarify the application of the Convention's

foundational legal principles and safeguards to LLMs and LLM-based systems throughout their lifecycle (Section 3).

The Guidelines adopt a stakeholder-specific approach when elaborating on concrete obligations or responsibilities for relevant actors including providers, deployers, users, and regulators. They explain how LLMs-based and agentic AI systems may interfere with individuals' privacy and data protection rights and illustrate corresponding mitigation strategies (Section 4).

Overall, the Guidelines provide an up-to-date, structured, research- and risk-informed framework for understanding and governing privacy and data protection risks associated with LLMs-based systems. They lay the groundwork for governance tools that operationalise both the principles and the binding obligations of Convention 108+.

A central feature of the operational application of Convention 108+ in this context is the structured approach to privacy risk governance based on four interrelated steps: "Identify", "Assess", "Mitigate" and "Monitor" ("IAMM approach") set out in Section 4. This methodology is designed to support diverse stakeholders in implementing Convention 108+ obligations in real-world settings, ensuring that emerging technologies remain aligned with fundamental rights, democratic values, and the rule of law.

The Guidelines contribute to the standard-setting efforts of the Council of Europe in advancing an integrated approach to data protection and AI governance through promoting a proactive, rights-based approach to innovation while safeguarding the principles of transparency, accountability, and human dignity. They provide transversal guidance on governance and accountability mechanisms that promote cross-functional collaboration and continuous oversight (Section 5).

These Guidelines do not establish a separate assessment instrument. Rather, they provide LLM-specific guidance that can inform privacy and data protection impact assessment obligations and practices under Article 10 of Convention 108+ and, where relevant, broader human rights, democracy and rule of law impact assessment methodologies developed within the Council of Europe, including HUDERIA. Their added value lies in translating these obligations and methodologies into the specific technical and organisational context of LLM-based systems and their lifecycle. (Section 5).

1.2 The IAMM Approach and Its Steps

IAMM Approach

These Guidelines aim to support organisations and other relevant stakeholders in managing privacy and data protection risks across the lifecycle of LLM-based systems and agentic architectures, from data collection and model development to deployment, monitoring, and post-market oversight. They respond to the need for a coherent and internationally aligned approach to privacy risk governance in the context of rapidly evolving AI systems.

The IAMM approach should not be understood as a mechanism for legitimising processing operations solely because a risk assessment has been carried out. A risk assessment is not an end in itself and does not automatically resolve data protection concerns.

In line with Convention 108+, the assessment may lead to the conclusion that certain processing operations, data sources, model capabilities, design choices or deployment contexts are incompatible with the Convention's requirements. In such cases, mitigation measures may not be sufficient, and the processing should not be initiated, or should be suspended, discontinued, substantially redesigned, or excluded from the intended use.

The IAMM approach includes:

- Establishing a common taxonomy (e.g., clarifying the notion of personal data in the context of LLMs) in Section 2.1.
- Analysing known and emerging privacy and data protection risks across the LLM ecosystem, in Section 2.2.
- Examining how the Principles and Articles enshrined in Convention 108+ and their implementation and safeguards are challenged by latest technological developments, in Section 3.
- Analysing risk management tools, including privacy- and human rights-centred assessment methodologies like the HUDERIA, in Section 2.2 and 4.
- Engaging stakeholders to prioritise piloting and validate the feasibility and relevance of proposed mitigation strategies in Section 4.
- And grounding best practices in real-world constraints and requirements from businesses, policymakers, and AI practitioners, in Section 5.

The underlying premise of the IAMM approach is that privacy and data protection risks in LLM-based systems cannot be adequately addressed through ad-hoc organisational practices or existing compliance tools alone. Instead, a structured, lifecycle-based methodology is needed to identify, assess, and mitigate risks at both the model and system level in a dynamic, and iterative manner.

Section 4 outlines a privacy risk management framework aligned with the principles of Convention 108+, adapted to the technical and organisational realities of Generative AI. Its scope includes the identification of privacy and data protection risks across distinct phases of development and deployment, a two-tiered assessment of risks at both the model- and composite system-levels, and an initial mapping of mitigation and monitoring measures within a framework structured around four core components: "Identify", "Assess", "Mitigate" and "Monitor" ("IAMM"):

1. Identify privacy and data protection risks under Convention 108+. This first step involves determining which risks may arise at different stages of the lifecycle of LLM-based systems, how such risks emerge, and which safeguards may be required in light of Convention 108+, the protections enshrined in Article 8 of the European Convention on Human Rights, and the risk and data protection obligations reflected in the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rules of Law ("Framework Convention on AI"), in particular Article 11 on privacy and personal data protection and Article 16 on a risk and impact management framework.
2. Assess risks through a two–tiered methodology, distinguishing between model-level and system-level risks as well as the respective roles and responsibilities of stakeholders throughout the lifecycle. This approach supports the assessment of risks in light of the

foundational principles laid out in Chapter II – Basic principles for the protection of personal data of Convention 108+, which include lawfulness, fairness, purpose limitation, data minimisation, transparency, safeguards for data subject's rights, data security and the protection of sensitive data. The objective is to equip stakeholders with tools to evaluate risks arising from different steps of the lifecycle, including risks related to training data, memorisation, inferencing, hallucinations, system integration, user interaction, downstream adaptation and third-party components.

3. Mitigate identified risks through appropriate technical and organisational measures proportionate to the nature, likelihood, and severity of the risks identified. Such measures may include governance controls, privacy-enhancing technologies (PETs), data protection impact assessment practices, monitoring mechanisms, and safeguards adapted to the operational context of LLM-based and agentic systems. The Guidelines recognise both the potential value and the limitations, trade-offs, and governance implications associated with such measures in practice.
4. Monitor residual and emerging risks throughout the lifecycle of LLM-based systems through continuous evaluation, oversight, auditing, and review mechanisms.

1.3 Target Audience and Stakeholders Across the LLM Ecosystem

These Guidelines are addressed to a broad range of stakeholders involved throughout the lifecycle and ecosystem of LLM-based systems, including start-ups, major technology providers, technology auditors, private and public deployers, regulatory authorities, civil society organisations, and research institutions. The diversity of actors involved reflects the distributed nature of roles, responsibilities, and governance challenges associated with the development, deployment, and use of LLM-based systems.

The roles and responsibilities are distributed and interdepended across multiple actors throughout the lifecycle of LLM-based systems, from model development to deployment and user interaction. Effective privacy and data protection governance therefore requires a multi-stakeholder approach reflecting distinct legal, technical, and operational responsibilities.

– States and Governments

Convention 108+ establishes a technologically neutral, principle-based framework for addressing privacy and data protection challenges arising from emerging technologies. States and Governments retain primary responsibility for establishing and maintaining legal and institutional frameworks that safeguard individuals' right to privacy and data protection. These Guidelines support the development, implementation, and promotion of context-sensitive privacy risk governance frameworks for LLM-based systems and contribute to regulatory coherence across sectors and jurisdictions.

– LLM Providers (Model Developers and System Designers)

A lifecycle approach grounded in Convention 108+ and the Framework Convention on AI supports the integration of privacy and data protection principles throughout the development and use of LLM-based systems, rather than introducing safeguards only at later stages. Model developers and system

designers should implement privacy by design and by default measures during model development, including during data collection, pre-processing, training, and post-training adaptation. These Guidelines support providers in understanding relevant data protection principles and obligations established by Convention 108+, as well as in conducting risk assessment and governance activities throughout the lifecycle of LLM-based systems.

– Deployers (System Integrators and Service Providers)

System integrators and service providers play a central operational role in the deployment and use of LLM-based systems. Even where LLM-based systems or components are provided by third parties, deployers may retain decision-making powers regarding specific data processing operations and therefore bear corresponding responsibilities under Convention 108+, depending on their role and degree of influence over the processing. In certain contexts, responsibilities may also be shared among multiple actors in accordance with the applicable interpretation of Article 2(d) of Convention 108+. These Guidelines support deployers in assessing privacy and data protection risks prior to deployment, implementing appropriate safeguards, and adopting relevant technical and organisational measures, including privacy-enhancing technologies (PETs) where appropriate.

– Industry

Industry actors contribute to the development of technical standards, operational best practices, and governance mechanism that may support greater consistency across sectors. These Guidelines may assist industry stakeholders in aligning privacy risk management practices and implementing lifecycle-based governance approaches for LLM-based systems.

– Civil Society Organisations

Civil Society Organisations contribute to awareness raising, public scrutiny, and independent assessment of privacy and data protection risks associated with LLM-based systems. They may also facilitate dialogue and feedback mechanisms between stakeholders, regulators, and affected individuals. These Guidelines support civil society organisations by providing a structured framework for understanding privacy risks and relevant mitigation approaches

– Users (End-Users)

The increasing deployment of LLM-based systems may affect individuals' rights and freedoms, including rights to privacy and personal data protection. These Guidelines aim to improve understanding of privacy-related challenges associated with such systems and provide a structured methodology for "identifying", "assessing", "mitigating", and "monitoring" such risks in real-world settings across different operational contexts.

– Regulators / Supervisory Authorities

Data protection supervisory authorities and other competent authorities play a central role in overseeing the lawfulness of data processing operations throughout the lifecycle of LLM-based systems.

Guidelines support supervisory and regulatory authorities in assessing transparency, accountability, and risk management practices associated with LLM-based systems, as well as in developing or

reviewing relevant oversight and assessment methodologies. They may also contribute to coordination and regulatory coherence across different actors and sectors.

Section 2 Key Concepts and Definitions

This section provides a technical overview of how personal data may be processed within Agentic and LLM-based systems, establishing the conceptual foundational for the proposed risk assessment framework. It clarifies key terminology and examines the privacy and data protection implications of current technological developments, including issues related to explainability and transparency, and their relevance for AI governance.

Rather than providing an exhaustive analysis, the section focuses on essential technical characteristics and operational features that are relevant for privacy and data protection risks, including how data may be ingested, transformed, organised, retained and potentially memorised by such systems. It also addresses risks associated with textual input data and their representation within LLM-based systems. By clarifying relevant technical concepts, system behaviours, and the ways in which these systems operate internally, this section seeks to demonstrate that a sufficiently informed understanding of the technical functioning of LLM-based systems is essential for developing effective privacy safeguards and governance mechanisms grounded in real system behaviour.

2.1 Essential Concepts

2.1.1 A Lifecycle Approach Rooted in the Framework Convention on AI and the LLM Ecosystem: Five Fundamental Steps

An introductory privacy risk analysis of the lifecycle of LLM-based systems can identify the main stages based on the types of data processing involved in the technological ecosystem of LLM-based systems, as illustrated in Figure 1. A lifecycle-oriented overview clarifies the fundamental distinction between foundation LLMs and the broader systems built with and around them to illustrate that different technological, organisational and economic actors may be involved at different stages of the lifecycle, with distinct roles, responsibilities and levels of influence over data processing operations. It also highlights that the nature, scope, sensitivity and context of the data processed, including personal data, may vary significantly across these phases.

The main stages include:

(#1) Model creation, where training data is collected, pre-processed, and models are developed with certain privacy by design considerations.
(#2) Post-training adaptation, where models are instructed, aligned, fine-tuned and adapted to specific tasks or operational purposes.
(#3) System integration, in which LLMs are integrated into applications, services or compound systems, often involving additional adaptation, orchestration mechanisms, or external components, together with appropriate safeguards.

(#4) Operational deployment, referring to the live deployment and operation of the LLM-based system, including active monitoring, governance controls, and post-deployment oversight mechanisms.

(#5) End-user interaction, covering how users interact with LLM-based systems and how autonomous workflows, memory functions or agentic capabilities are configured and managed.

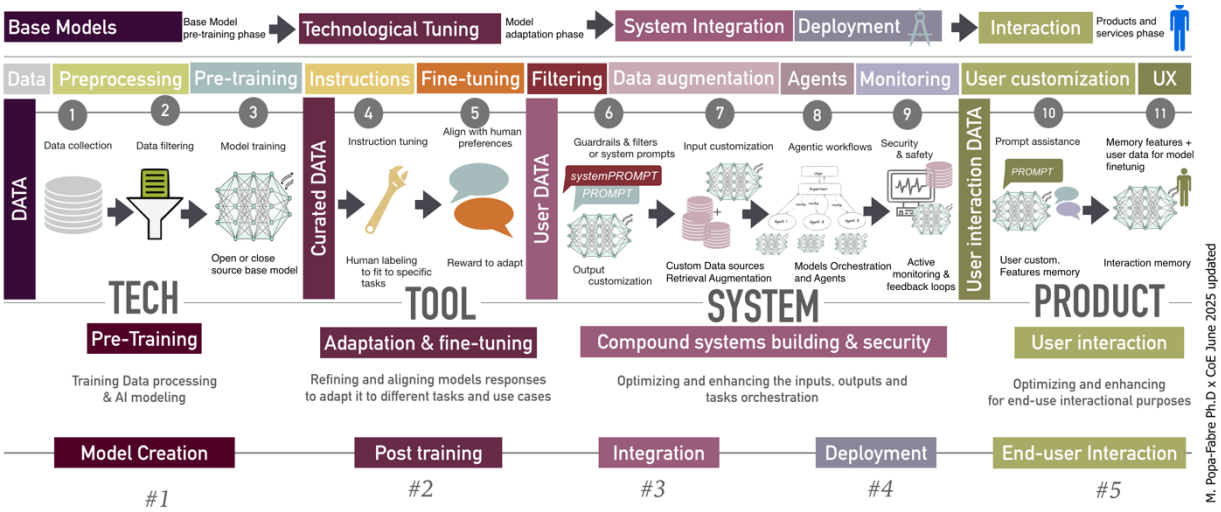

Figure 1: The lifecycle and value-chain of LLM-based systems Illustrative overview of the lifecycle and value-chain of LLM-based systems, reflecting evolving practices in the development and integration of compound AI systems and LLM-based applications.

Foundational stages

While the model pre-training phase of LLMs' development (stage #1) typically relies on the processing of very large and diverse datasets, including data collection and pre-processing (steps 1 to 3), subsequent stages increasingly depend on more curated and context-specific data (stages #2, #3, #5).

Intermediate stages

During post-training adaptation (stage #2), models are refined and aligned for particular purposes, often through the processing of targeted datasets and task-specific inputs. This intermediate phase, which includes fine-tuning, system integration, and preparation for deployment (steps 4 to 9), is particularly significant from a privacy and data protection perspective. It is often during these stages that safeguards may be embedded into system design, governance structures, and operational configurations. Despite its importance, these steps remain insufficiently examined in many privacy and data protection discussions.

Later stages

In later stages of the lifecycle (stage #5), optimisation and customisation practices further shape the functioning of Agentic and LLM-based systems in operational environments (steps 10 to 11). The evolving landscape of integration techniques and compound system and agentic architectures (steps 7 to 11) has enabled standalone models to operate within increasingly complex environments. These include, for example:

- Data augmentation techniques such as Retrieval-Augmented Generation (RAG), which allow models to access external or proprietary data sources in response to user queries;

- Agentic workflows that orchestrate multiple models, tools and APIs to perform multi-step tasks with varying degrees of autonomy.

Deployment advanced stages

At advanced stages of deployment, additional layers of personalisation may be introduced. These frequently rely on the intensive processing of personal data to tailor applications, products, or services to individual users. Such mechanisms may include prompt customisation (stage 10), user intention decoding, and features designed to infer user preferences, habits, or contextual cues in order to dynamically adapt system responses (stage 11). While these developments may enhance functionality and adaptability, they may also expand the scope, volume, sensitivity, and persistence of personal data processing, thereby increasing privacy and data protection risks.

2.1.2 Difference between LLM Models and LLM-based Systems

Together with the nascent agent-based applications for personal assistance (so-called "Personal Intelligence Products"), popular LLM-based applications rely heavily on the intensive use of personal data to create intuitive and highly interactive services. While such functionalities may improve usability and personalisation, they may also raise significant privacy and data protection concerns. In particular, they may blur the boundaries between training, personalisation, and continuous data collection processes, thus increasing the difficulty of determining where, how and for what purposes personal data are processed. In this context, it is essential to distinguish between risks associated with the model itself and those that emerge at the level of the broader system in which the model is embedded. This distinction is not merely technical; it is foundational to design effective, lawful, and context-aware privacy risk management strategies.

Model-level risks

Model-level risks arise from the ways in which LLMs are trained, post-trained, fine-tuned, and architected (stage 1 to 5), including:

- Ingestion of personal data during pre-training (stage #1), often from large-scale web scraping, that may lack sufficient transparency or an appropriate legal basis.
- Memorisation and regurgitation of sensitive or identifiable information derived from training data, potentially conflicting with principles relating to data minimisation and storage limitation.
- Hallucinations or the generation of plausible but inaccurate personal information, which may undermine the principles data quality, notably accuracy under Article 5 of Convention 108+ and may adversely affect individuals' privacy, reputation, or other rights and freedoms.
- Bias amplification, where underlying statistical associations reproduce or reinforce unfair or discriminatory patterns relating to individuals or groups.

System-level risks

System-level risks arise when an LLM is integrated into a broader application environment, often including APIs, interfaces, plug-ins, memory functions, feedback loops (step 6), Retrieval-Augmented Generation (RAG) systems (step #7), agentic workflows (step #8) involving LLM orchestration, and third-party services (stages #3 to #5). In such contexts, privacy and data protection risks are linked not

only to the behaviour of the model itself, but also to how the system is configured, deployed, governed, and used, and by whom. Examples of risks include persistent user profiling or behavioural inference through memory and interaction-history features (step #11).

- Lack of transparency regarding how user data is processed, shared, retained or reused, particularly when LLMs operate in dynamic cloud-based environments.
- Cross-context data leakage, where user data provided in one application context is reused in another (e.g., through shared fine-tuning across products).
- Inadequate user controls, information and or consent mechanisms, especially for secondary uses of data, including personalisation, A/B testing or system optimisation.

Memory functions may store and leverage users' interaction history to improve continuity, personalise responses, enhance user experience and support further system optimisation or product development.

2.1.3. Lifecycle risk approach

At each stage of the lifecycle, data quality, provenance, and availability play a significant role. However, it remains important to distinguish between the privacy and data protection risks associated with LLM base-models and those arising from LLM-based systems operating within specific deployment and operational contexts.

Within the ecosystem of LLM-based systems, the privacy and data protection risk landscape is dynamic, context-dependent, and multi-layered. To be effective, a privacy risk framework for LLM-based systems should:

1. Address both model-level and system-level risks by considering risks associated both with the statistical model itself and with the broader application or operational environment in which it is deployed.
2. Track risks across the different steps of the lifecycle: from pre-training and post-training adaptation to deployment, monitoring, downstream integration, and user interaction.
3. Track response variability through continuous evaluation: recognising that LLMs are probabilistic rather than deterministic[1] systems and that certain privacy and data protection risks may only emerge post-deployment (e.g., through memorisation, regurgitation, prompt injection, inferencing, or output misuse).
4. Incorporate both technical and organisational safeguards recognising that no single technical measure is sufficient, and that effective governance requires combined technical, legal, organisational and user-centred approaches.

Effective privacy and data protection risk management must remain aligned with core principles of Convention 108+, such as *lawfulness*, *fairness*, *transparency*, *purpose limitation*, data minimisation and *accountability*. These principles apply not only to the foundation models themselves, but also to the

[1] LLM-based systems fundamentally differ from traditional software in that they do not have a predefined set of rules yielding a deterministic behaviour where one input corresponds only one output. Probability and prediction are at the core of LLM non-deterministic behaviour, where one input has many different outputs requiring a governance framework to embed a continuous evaluation layer.

broader technical and organisational ecosystems in which LLM-based systems are developed, deployed, and operated and updated.

2.2 Types of Privacy and Data Protection Risks Across the AI Lifecycle and the LLM Ecosystem

2.2.1 Lifecycle and Operational Privacy Risks in LLM-based Systems

LLMs and LLM-based systems may give rise to significant privacy and data protection challenges throughout their lifecycle. Risks may manifest not only through direct data leakage or memorisation, but also through model and system behaviours that facilitate identity manipulation, impersonation, disinformation, or potentially manipulative predictive behaviour. For example, synthetic or AI-generated content and AI-generated identities have proliferated online, reinforcing the difficulty of distinguishing real information and real individuals from artificially generated ones.

Privacy and data protection risks in LLM-based systems should be understood in relation to specific phases of the lifecycle. Each phase presents distinct risk profiles, system exposures, and governance challenges that require context-sensitive assessment:

1. Phase 1: Training data collection and pre-processing
   Risks may arise from uncontrolled large-scale ingestion and aggregation of personal data within training datasets. Publicly available datasets and corpora may contain identifiable or sensitive information, such as contact details, credentials, emails, or other personal data. Without adequate filtering, models can memorise and later reproduce this content, potentially affecting the lawfulness of processing and the principles relating to data quality, accuracy, and data minimisation under Convention 108+.
2. Phase 2: Post-training adaptation and fine-tuning
   Fine-tuning introduces risks, particularly when sensitive or domain-specific data is used without sufficient controls. It may also amplify biases, alter model behaviour in unpredictable ways, or introduce additional privacy and data protection risks. Many deployers rely on fine-tuned models offered as services without full transparency about the post-training process, complicating privacy and data protection assessments, accountability, and governance responsibilities.
3. Phase 3: System integration, API design, and Agentic orchestration
   Risks extend beyond the model itself to include APIs, middleware, Retrieval Augmented Generation (RAG) architectures, external integrations and Agentic orchestration mechanisms. Poorly secured endpoints or integrations may expose private data. These risks reinforce the importance of security-by-design principles architectural oversight and layered technical and organisational safeguards.
4. Phase 4: Operational deployment, monitoring and post-deployment adaptation
   Ongoing data collection and model updates often lack transparency. Feedback data may be reused in ways that reintroduce privacy risks, and users may be unaware of how their inputs are stored or analysed. Articles 10 of Convention 108+ and Article 16 of the Framework

Convention on AI underscore the importance of continued oversight, ongoing risk management, and effective impact assessment processes.

5. Phase 5: Inference, user interaction and Agentic functionality
   Prompt injection and jailbreak techniques may create significant privacy and security risks during this phase. The inability to consistently predict or trace outputs also raises issues of transparency and accountability. The intensive processing of personal data to create intuitive and highly interactive services may further increase previously identified privacy and data protection risks.

Beyond these lifecycle phases, LLM-based systems may also introduce broader systemic and societal harms, which require a different kind of scrutiny[2].

2.2.2 Data Processing risks in LLM-based Systems: Privacy and Data Protection

While the previous section mapped privacy and data protection risks across different lifecycle phases, this section focuses specifically on the categories of personal data processed within the ecosystem of LLM-based systems. Understanding the types of data processed is essential for assessing the scope, sensitivity, and potential impact of privacy and data protection risks.

The data ecosystem surrounding LLM-based systems is layered, dynamic, and context-dependent. Different categories of personal data may be processed at different stages of model development, adaptation, integration, deployment and operation. These categories may vary significantly in origin, sensitivity, identifiability, provenance and governance implications.

Phase 1: Training and Pre-Training Data
Training datasets may include publicly accessible online content, licensed corpora, proprietary databases, and other aggregated sources. Such datasets may contain:

- Identifiable personal information, such as names, contact details, identifiers or professional affiliations;
- User-generated content, including forum discussions, social media posts, and comments;
- Special categories of data, including where such data are collected or incorporated without adequate filtering or safeguards;
- Embedded confidential or sensitive information, such as credentials (API keys), personal communications or proprietary information.

At this stage, personal data are often processed at scale and without direct interaction with data subjects. The scale and aggregation of such data may increase the potential for unintended retention, inference of sensitive attributes, and reduced transparency regarding data provenance and lawful basis for processing.

---

[2] See for example structural implications described in the Draft Guidance Note on Generative AI implications for Freedom of Expression by the Council of Europe MSI-AI Committee. https://www.coe.int/en/web/freedom-expression/msi-ai-committee-of-experts-on-the-impacts-of-generative-artificial-intelligence-for-freedom-of-expression# {%22265382451%22:[1]}.

Phase 1 & 2: Testing, Evaluation and Post-training Data

During testing, benchmarking, evaluation and red-teaming activities, additional datasets may be introduced. These may include curated personal data, synthetic data derived from real individuals, adversarial prompts, or evaluation of datasets designed to assess system performance. Although often considered technical artefacts, such datasets may still contain identifiable or inferable personal information. Their use may raise questions concerning secondary processing, safeguards in experimental environments, and the distinction between anonymised and pseudonymised data.

Phase 3 & 4: Operational and Interactional Data.

Once deployed, LLM-based systems process data generated through user interaction. These may include:

- User queries and prompts;
- Uploaded documents, files or contextual inputs;
- Retrieved content within architectures such as Retrieval-Augmented Generation (RAG) systems;
- Usage logs, telemetry, and behavioural metadata.

Operational and interactional data are often highly contextual and may reveal preferences, behavioural patterns, intentions, habits, or sensitive personal circumstances. Unlike many training datasets, these inputs are often directly linked to identifiable users and may be processed continuously, retained for system optimisation, adaptation, or monitoring purposes.

Phase 5: Monitoring, Feedback, and Retraining Data

Post-deployment improvement mechanisms frequently rely on feedback data, performance logs, and selected interaction records. Over time, these feedback loops may generate additional datasets derived from user behaviour and system interactions. Such iterative processing may blur the boundaries between initial purposes and subsequent optimisation or retraining activities, raising concerns related to purpose limitation, retention periods, and the allocation of responsibilities among actors within the ecosystem.

Phase 6: Archival, Retention and Deletion Data

At later stages of the lifecycle, decisions may need to be taken regarding the retention, anonymisation, deletion or continued storage of training data, operational logs, and derived datasets. In multi-actor ecosystems, determining responsibility for erasure or rectification, or other data subject rights may become particularly complex, particularly where personal data are embedded in model parameters, logs, memory systems or distributed infrastructures.

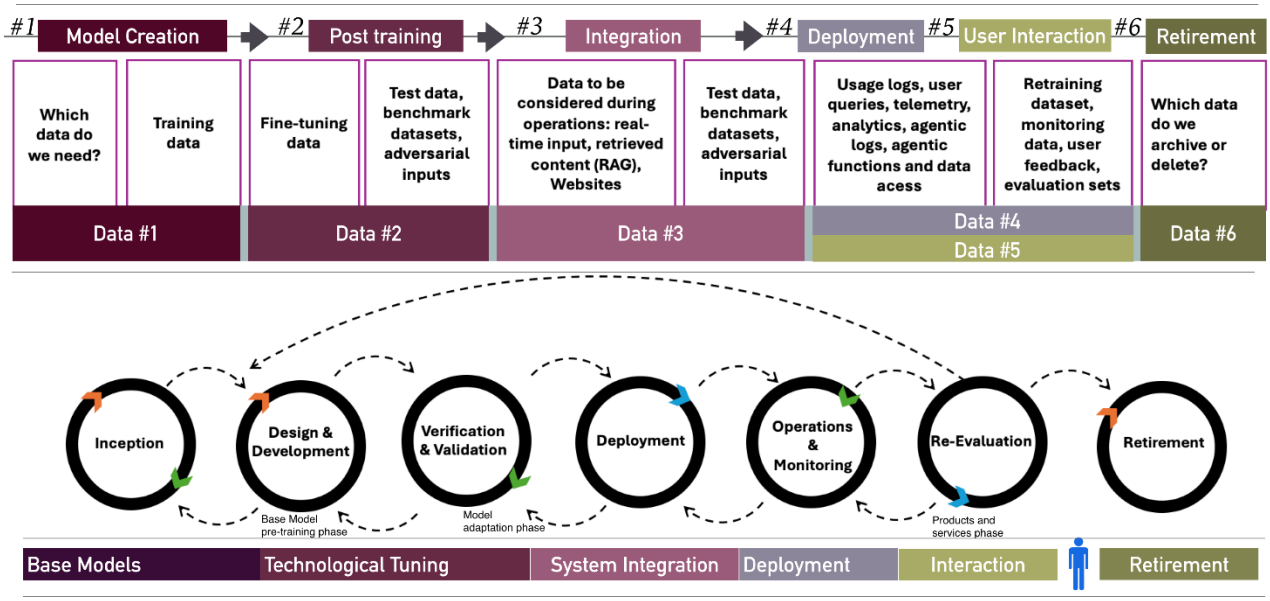

Figure 2: Data Flows across the lifecycle stages of LLM-based systems Illustrates the distribution of privacy risks across different lifecycle stages of LLM-based systems with a focus on data access and flows. (Modified from EDPB's report on Privacy Risks & Mitigations in LLMs. The lifecycle phases of ISO/IEC 22989 are also illustrated in the image.

2.3 Emerging Privacy Risks

As LLM-based systems become increasingly integrated into public and private infrastructures, and progressively combined with other rapidly evolving technologies, new categories of privacy and data protection risks are emerging, that are both contextual and systemic. These developments extend beyond traditional digital processing environments and increasingly involve persistent, adaptive, multimodal, and autonomous systems (e.g., AI assistants and AI agents) capable of interacting continuously with individuals, environments, devices, and other AI systems.

Integration of LLM-based systems with other technologies

Under the framework of Convention 108+, such developments raise important questions regarding the scope, intensity, and persistence of personal data processing, as well as the capacity of existing safeguards to remain effective in increasingly complex technological ecosystems. In particular, the convergence of LLM-based systems with technologies such as neurotechnologies, biometric systems, robotics, physical AI, ambient computing, emotional AI, wearable devices, and autonomous agents may significantly expand the scale and granularity of data collection, behavioural analysis, inferencing, and personalised interaction.

Beyond data processing and data flows solutions

Current governance and risk assessment practices may not yet be fully equipped to address the complexity and evolving characteristics of these systems. Traditional approaches often focus on discrete processing operations or clearly identifiable data flows, whereas emerging LLM-based ecosystems may involve continuous interactions, distributed architectures, interconnected services, autonomous decision-making processes, and ongoing adaptation through feedback loops and environmental inputs.

Beyond personal data protection

In such contexts, privacy and data protection risks may no longer arise solely from the collection or disclosure of personal data, but also from increasingly advanced forms of inferencing, behavioural prediction, emotional analysis, persistent monitoring, identity simulation, and large-scale aggregation of contextual and multimodal information.

Emerging risks may arise from:

- increasingly advanced multimodal systems capable of combining textual, visual, audio, behavioural, geolocation, biometric, and contextual data in ways that enable highly granular profiling, inferencing, prediction, or behavioural manipulation;
- the emergence of persistent Agentic systems and personalised AI assistants capable of continuously collecting, retaining, and analysing user interaction histories, preferences, habits, routines, or contextual signals across multiple services and environments;
- large-scale interoperability and data sharing between interconnected AI systems, platforms, APIs, and autonomous agents, potentially resulting in secondary uses of personal data that exceed the original context or reasonable expectations of data subjects;
- the increasing difficulty of distinguishing between authentic and synthetic interactions, identities, communications, or representations, including where LLM-based systems generate highly realistic content capable of simulating individuals, behaviours, or social relationships;
- emerging forms of inferencing and predictive analysis capable of revealing intimate aspects of private life, including behavioural traits, emotional states, vulnerabilities, intentions, or sensitive characteristics derived from seemingly non-sensitive data.
- the integration of LLM-based systems with neurotechnologies, biometric systems, affective computing, or cognitive inference tools capable of processing highly sensitive behavioural, physiological, emotional, or neural data;
- the deployment of LLMs within physical AI systems, robotics, autonomous devices, wearable technologies, smart environments, or ambient computing infrastructures involving continuous environmental sensing and real-time behavioural monitoring;
- the increasing difficulty of distinguishing between authentic and synthetic interactions, identities, communications, or representations, including where LLM-based systems generate highly realistic content capable of simulating individuals, behaviours, or social relationships.

Amplification of Informational asymmetries

Such developments may amplify asymmetries of information and power between individuals and organisations, while increasing challenges relating to transparency, autonomy, meaningful consent, and the effective exercise of data subjects' rights.

Beyond privacy and data protection

In certain contexts, these risks may also engage broader protections relating to dignity, individual identity and autonomy, psychological integrity, and the right to private life affecting not only the safeguards of Convention 108+, but also the protections enshrined in Article 8 of the European Convention on Human Rights (ECHR), which upholds the right to identity, reputation, and private life.

Section 3 Convention 108+ Principles and Articles Relevant to LLM-based Systems

This section focuses on the core data protection principles set out in Chapters II and III of Convention 108+, with particular attention to their relevance in the context of LLM-based systems and Agentic architectures.

It begins by analysing the foundational principles established in Chapter II, including lawfulness and fairness, purpose limitation, data minimisation, accuracy, transparency, accountability, data security, the protection of special categories of data, and the rights of data subjects.

The objective is twofold. First, the section highlights the specific tensions and challenges that LLM-based systems and Agentic frameworks may generate under these principles, including issues such as data repurposing, hallucination phenomena, inferential processing of personal data, large-scale behavioural prediction, and risks associated with synthetic data and multimodal aggregation.

Second, it clarifies and contextualises how these principles apply across the lifecycle of LLM-based systems, including during training, post-training adaptation, system integration, deployment, and post-deployment monitoring.

Through this analysis, the section aims to provide a principled legal foundation for the lifecycle-based risk governance approach developed in subsequent sections.

3.1 Understanding the Principles of Convention 108+ in the Context of Evolving LLM-based and Agentic Systems

3.1.1 Data Security, Accuracy, Transparency, and Accountability in LLM-based and Agentic Systems

Agentic LLM-based systems amplify traditional data security, transparency, and accountability challenges. The orchestration of APIs, retrieval pipelines, plugins, external tools, memory functions, and interconnected services expands the attack surface and increases the risks of unauthorised access, cross-context data leakage, secondary processing, or unintended onward transfers of personal data. In such layered and distributed environments, determining which actor acts as controller, joint controller, or processor may be complex, thereby complicating the allocation of responsibilities under Convention 108+.

The opacity of inference processes, post-training adaptation, and downstream integrations may undermine transparency obligations and limit the effective exercise of data subjects' rights. In particular, where outputs rely on probabilistic generation, dynamic retrieval from multiple sources, external API calls, and continuous contextual adaptation mechanisms, ensuring traceability, auditability, and meaningful explainability becomes technically and organisationally challenging.

Loss of autonomy and meaningful control

The increasing reliance on predictive and inferential processing in LLM-based systems raises concerns regarding individuals' loss of autonomy and meaningful control over their personal data. Where interactional AI systems infer preferences, behavioural patterns, emotional states, vulnerabilities, or sensitive attributes from limited interaction data or multimodal signals, individuals may become subject to profiling and behavioural prediction practices that are neither sufficiently transparent nor easily contestable. Such dynamics may erode informational self-determination and undermine individuals' ability to meaningfully understand, influence, or challenge how their data are interpreted and operationalised within automated environments.

Hallucination and inaccurate personal information

Hallucination phenomena and documented inaccuracy introduce distinct privacy and data protection risks. LLMs-based systems may generate plausible but inaccurate statements about identifiable individuals, including false allegations, fabricated biographical details, misleading contextual information, or synthetic associations between individuals and events. Even where such outputs are entirely generated, they may still qualify as personal data where they relate to an identifiable individual and affect that person's rights, reputation, or interests.

Such risks engage the principles of accuracy and data quality under Article 5 of Convention 108+ and raise questions concerning rectification, transparency, accountability, and effective remedies, particularly where generated outputs are reproduced, shared, or relied upon in downstream systems.

Synthetic data and synthetic identity risks

Although synthetic data are often presented as a privacy-enhancing solution, they may still encode statistical patterns derived from real individuals or reproduce sensitive correlations embedded in training datasets. In certain circumstances, synthetic outputs may facilitate re-identification when combined with additional information or linked datasets.

Moreover, the increasing use of synthetic identities, deepfake technologies, AI-generated personas, and realistic simulation systems may facilitate impersonation, fraud, manipulation, or deceptive interaction practices, thereby affecting individuals' dignity, identity, reputation, autonomy, and private life.

Automated decision-making and accountability

LLM-based applications are increasingly integrated into automated or semi-automated decision-making environments, including recruitment, education, insurance, healthcare, customer management, law enforcement support, and content moderation. Where such systems operate without meaningful human oversight or effective safeguards, they may undermine protections reflected in Article 9.1 litterae "a" of Convention 108+, which safeguards individuals from decisions significantly affecting them based solely on automated processing.

Deceptive or manipulative interface designs, dark patterns, anthropomorphic interaction strategies, and limited access to meaningful redress mechanisms may further amplify these risks. Such structural issues may extend beyond traditional data protection concerns and have broader implications for democratic participation, media integrity, informational autonomy, and public trust.

3.1.2 Lawfulness and Fairness of Processing: Inferencing, Data Proxies, and Reconstruction of Private Life

The multimodal aggregation and predictive capabilities of LLM-based systems increasingly enable the inferencing of personal data and the reconstruction of detailed aspects of individuals' private lives, identities, behaviours, preferences, and vulnerabilities through the processing of data proxies. Through advanced simulation techniques, such systems can better capture the multifaceted nature of human behaviour with implications not only for data protection, but also for private life, dignity, and autonomy. Analysing these recent developments is essential to forge future-proof governance and risk management measures that can address emerging privacy risks that address all considerations relating to the right to privacy.

Rather than relying solely on directly provided personal information, such systems may combine behavioural, contextual, biometric, geolocation, interactional, and environmental signals and data proxies to infer highly sensitive information concerning individuals and can thus reconstruct a comprehensive picture of individuals' personal lives, identities and behaviours[3].

Behavioural prediction, predictive data-caging and fairness, lawfulness and proportionality
Recent research demonstrates that LLM-powered systems are increasingly capable of identifying patterns across multiple modalities of information and using these patterns to simulate, predict, or influence user behaviour. Such developments may raise new privacy and autonomy risks by:

1. Enabling the fusion of diverse data modalities to capture the multifaceted nature of users' behaviours on e-commerce platforms and predict product selection (e.g., 43.09% boost in Sports category accuracy on Walmart).[4]
2. Providing adaptable needs anticipation by simulating users' intentions and behaviour in LLM-based Agentic Recommender Systems[5] that significantly boost users' intentions and next-purchase predictions.
3. Simulating election prediction[6] through simple prompting techniques based on publicly available demographics and social science surveys (e.g., 2024 US presidential election).

---

[3] Data proxies processed by LLMs can also yield large scale online deanonymisation, see Lermen, S., Paleka, D., Swanson, J.P., Aerni, M., Carlini, N., & Tramèr, F. (2026). Large-scale online deanonymization with LLMs. ArXiv, abs/2602.16800.

4 Ma, L., Li, X., Fan, Z., Xu, J., Cho, J.H., Kanumala, P., Nag, K., Kumar, S., & Achan, K. (2024). Triple Modality Fusion: Aligning Visual, Textual, and Graph Data with Large Language Models for Multi-Behavior Recommendations. ArXiv, abs/2410.12228.

[5] See recent review of the literature of LLM-based Agentic Recommender Systems (LLM-ARS): Huang, C., Yu, T., Xie, K., Zhang, S., Yao, L. and McAuley, J., 2024. Foundation models for recommender systems: A survey and new perspectives. arXiv preprint arXiv:2402.11143.

[6] Jiang, S., Wei, L., & Zhang, C. (2024). Donald Trumps in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models.

4. Powering AI agent simulations of surveys' responses that correspond to accurate human response patterns in sociological and psychological matters or economic games.[7]

These developments illustrate how emerging LLM-based architectures may create highly adaptable and continuously updated behavioural representations of individuals through the aggregation of heterogeneous data sources and inferential processing techniques. Such configurations may result in forms of persistent and predictive profiling that significantly expand traditional understandings of personal data processing.

This dynamic can be understood as a form of "predictive data-caging", whereby systems continuously model, anticipate and nudge individuals' behaviours, preferences, intentions[8], and reactions through extensive inferential analysis and thus create a data-cage for the user where every digital trace is constraining and cueing for future digital interactions. Such developments raise important questions concerning fairness, lawfulness, proportionality, informational self-determination, and the limits of behavioural prediction under Convention 108+. The lawfulness and fairness principles established under Article 5 require that such processing operations remain proportionate, transparent, and compatible with individuals' reasonable expectations. Inferences generated through multimodal aggregation and behavioural prediction may significantly affect individuals even where direct identification is absent.

3.1.3 Data Minimisation and Data Subjects' Rights in Personalised and Intention-predictive Systems

Emerging intention-predictive systems differ fundamentally from traditional profiling and recommender architectures. Whereas traditional profiling practices may often be experienced as stereotyping or categorisation by the end-user, increasingly adaptive LLM-based systems may generate dynamic behavioural representations capable of continuously and seamlessly modelling users' intentions, preferences, emotional states, or likely future actions.

The adaptive and personalised nature of such systems may create the perception of assistance, understanding, or companionship while simultaneously reducing individuals' visibility into how their behaviour is being analysed, predicted, or influenced. Such developments raise important concerns regarding the principles of data minimisation, proportionality, and fairness, as well as the effective exercise of data subjects' rights under Articles 8 and 9 of Convention 108+.

---

[7] Convergent research results in the last few years show how LLM-based interactive agents can simulate individual complex behaviour and survey responses (e.g., General Social Survey, Big Five Personality Inventory, Economic Games or Behavioural Experiments) from basic demographic information. A recent study by Stanford University showed that AI agents can simulate human behaviour with up to 80% accuracy based on just two hours of interview data, and 60% accuracy using only basic demographic information. Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C.J., Morris, M.R., Willer, R., Liang, P., & Bernstein, M.S. (2024). Generative Agent Simulations of 1,000 People. ArXiv, abs/2411.10109. See also Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.

[8] Further reading on the concept of behavioural prediction and predictive data caging in the field of Human Rights, see M. Popa-Fabre in the Inaugural CDNET Conference report, Council of Europe, May 2026. As a nota bene, it is important to set a distance between scientific and colloquial references to intentionality in the design of digital services and models described here. In the LLM-based recommender systems that are described here (see above list items 1 and 3), we discuss how to computationally infer elements that are cues for human intent by multimodal capture of users' behaviour on e-commerce platforms that embed LLMs and Agentic simulations.

From the attention economy to the intention economy[9]These recent technological developments suggest a gradual shift from business models centred on capturing human attention towards increasingly predictive systems focused on forecasting, shaping, or operationalising human intentions and behaviours.

The combination of multimodal data aggregation, LLMs' pattern-finding capabilities, reinforcement learning techniques, adaptive personalisation systems, and advanced inferential capabilities may enable highly granular behavioural prediction and highly personalised interaction mechanisms. Such developments raise broader questions regarding autonomy, behavioural manipulation, and informational asymmetries between individuals and organisations. They also reinforce the importance of ensuring that individuals retain meaningful control over their personal data and remain capable of understanding and contesting automated processing practices that significantly affect them. Inferencing users' intention, without implying direct re-identification, is fundamentally questioning the human right to autonomy and private life.

User control, consent, and continuous optimisation

The increasing use of continuous optimisation practices, memory systems, silent feedback loops, behavioural analytics, and persistent personalisation mechanisms weaken users' control over consent, retention periods, and downstream uses of personal data.

In some contexts, users may be unaware that their interactions are retained, reused for optimisation purposes, or integrated into broader behavioural and interactional modelling systems enhancing user-experience. The removal or weakening of opt-out mechanisms, limitations on access to deletion tools, or insufficiently transparent interface design may further restrict the practical exercise of data subjects' rights. These developments may become particularly significant where Agentic and LLM-based systems increasingly function as primary conversational interfaces for human-machine interaction across multiple domains of daily life.

3.1.4 Purpose Limitation in Multimodal and Interconnected Data Ecosystems

The principle of purpose limitation becomes increasingly difficult to operationalise in environments characterised by multimodal aggregation, interconnected systems, continuous adaptation, and large-scale behavioural inferencing, enabling the creation of detailed and continuously evolving representations of individuals' private lives using data originating from multiple devices, services, and operational contexts. In such environments, data initially collected for limited and context-specific purposes may subsequently be repurposed, aggregated, or reused for optimisation, behavioural prediction, personalisation, or inferential analysis in ways that exceed the original purposes of collection or the reasonable expectations of data subjects.

Smartphones, ambient systems, and behavioural data ecosystems

[9] For a detailed discussion of the shift from attention economy to intention detection and its economic impact and harms, see Chaudhary, Y., & Penn, J. (2024). Beware the Intention Economy: Collection and Commodification of Intent via Large Language Models. Harvard Data Science Review, (Special Issue 5). https://doi.org/10.1162/99608f92.21e6bbaa. Specifically, see Section 4 on the implication of projecting false intentionality onto users of LLM based systems users, who is here defined as 'predictive data-caging'.

Emerging privacy risks may become particularly pervasive in the ecosystem of digital traces generated through smartphones, wearable devices, smart-home sensors systems, and ambient connected environments.

These systems increasingly function as central hubs for collecting and aggregating behavioural, contextual, physiological, and environmental data across multiple modalities and interconnected services.

When combined with Agentic and LLM-based inferential and predictive capabilities, such ecosystems may significantly increase the capacity to infer intimate aspects of individuals' private lives, routines, relationships, emotional states, habits, vulnerabilities, or behavioural tendencies.

Data proxies and multimodal inferencing

Modern connected environments increasingly rely on a broad ecosystem of sensors and digital traces capable of generating proxy indicators concerning individuals' behaviour, health, cognition, emotional states, or social relationships. For example:

- GPS, accelerometers, gyroscopes, and mobility sensors may reveal behavioural routines, movement patterns, physical activity, or social habits;
- Wi-Fi spots proximity, and Bluetooth signals may facilitate the inferencing of social relationships, routines, or behavioural rhythms;
- Cameras and eye-tracking systems may enable the analysis of pupil dilatation, eye-movements, emotional expressions, attention patterns, as cognitive indicators[10];
- Touchscreen interaction data may reveal behavioural and cognitive characteristics, including reaction times, motor patterns, or engagement behaviours;
- Light sensors, sleep-tracking systems, and physiological monitoring tools may provide indirect indicators concerning mental health, stress, fatigue, or emotional conditions;
- Voice recordings and conversational systems may reveal emotional states, social interactions, behavioural tendencies, or environmental characteristics.

When aggregated and processed within interconnected LLM-based ecosystems, these heterogeneous data sources may enable increasingly advanced forms of inferential profiling and behavioural prediction.

Hyper-profiling and manipulation risks

The convergence of multimodal data aggregation, advanced inferential and predictive capabilities, reinforcement learning and adaptive optimisation techniques, and persistent agentic architectures, may significantly expand the capacity of organisations to build comprehensive and continuously evolving behavioural profiles of individuals, like in virtual AI assistants and personal intelligence products. These developments may introduce risks extending beyond traditional understandings of

[10] See one among the first use of smartphones for psychiatric patients' tracking: Torous J, Kiang MV, Lorme J, Onnela JP New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research JMIR Ment Health 2016;3(2):e16 doi: 10.2196/mental.5165,https://spectrum.ieee.org/a-software-shrink-apps-and-wearables-could-usher-in-an-era-of-digital-psychiatry

data protection and require renewed attention to broader privacy concerns relating to autonomy, manipulation, behavioural influence, dignity, and private life.

In this context, the principles of purpose limitation and data minimisation should be interpreted 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.

3.1.5 Balancing Principles and Trade-offs in LLM-based Systems

Trade-offs between Privacy Principles and systems' functionalities

The application of Convention 108+ principles in the context of LLM-based systems and Agentic architectures may involve complex trade-offs between competing objectives, systems' functionalities, and legal requirements.

In practice, the design, development, and deployment of Agentic and LLM-based systems often require balancing privacy and data protection considerations with other technical, economic, or operational objectives, such as performance, accuracy, usability, personalisation, robustness, and security. For example, the shift from attention to intention economy and the interactional seamlessness introduced by behavioural prediction introduces a trade-off between privacy and personalisation of the service interactivity. In other words, the more personal data the better customisation is, and this represents *de facto* an incentive to disclose information. Likewise, safety monitoring in AI companions or other emotion AI services requires conversation access which is in turn undermining privacy expectations.

Measures aimed at strengthening data minimisation or limiting data retention may affect system performance or functionality, while increased transparency or explainability may introduce security risks or expose system vulnerabilities.

Trade-offs between data protection principles

Tensions may also arise between data protection principles themselves. For example, the use of large and diverse datasets may support fairness and reduce bias, while raising concerns under the principles of data minimisation and purpose limitation. Similarly, strict minimisation may limit the ability to detect or mitigate discriminatory outcomes, and highly personalised systems may enhance user experience while raising concerns regarding autonomy, proportionality, and excessive profiling. LLM-based and Agentic systems further amplify these tensions through their reliance on inferential processing, multimodal data aggregation, and adaptive learning mechanisms. The capacity to generate predictions, simulate behaviour, and anticipate user intentions may increase system effectiveness while simultaneously creating risks of manipulation, behavioural influence, or loss of individual autonomy.

Analysis of trade-offs in the risk assessment process

In this context, the application of Convention 108+ requires a structured and explicit assessment of trade-offs and balancing of competing considerations. Such assessment should form an integral part of the risk assessment process and be carried out throughout the lifecycle of LLM-based systems, in particular in accordance with the obligations set out under Article 10.

This balancing exercise should be guided by the principles of proportionality, necessity, fairness, and respect for fundamental rights, including the right to private life under Article 8 of the European Convention on Human Rights. It should take into account the nature, scope, and potential impact of the processing on individuals' rights and freedoms and should not be determined solely by considerations of technical feasibility, system performance, or economic efficiency.

HUDERIA impact analysis Such trade-offs require documented, transparent, and accountable decision-making processes, supported by impact assessments, risk evaluations, and appropriate safeguards throughout the lifecycle of LLM-based systems that provide an initial mapping of the scale, scope, probability and reversibility. Such assessments can leverage the Context Based Risk Analysis (COBRA) and the Stakeholder Engagement Process (SEP), and all relevant elements embedded in the HUDERIA methodology developed to map impacts based on Human Rights, Democracy and the Rule of Law and to support the Framework Convention on AI.

Where such assessment identifies that trade-offs may result in a significant risk to individuals' rights and freedoms, priority should be given to ensuring effective protection of those rights (see Mitigation Plan in the HUDERIA Methodology). This may require adapting system design, limiting data processing operations, restricting certain functionalities, introducing additional safeguards, or, where necessary, refraining from specific processing activities or deployment contexts.

3.2 Understanding the Articles of Convention 108+ in the Context of LLM-based and Agentic Systems

While Article 1 affirms the individual's right to privacy and Article 2 defines key concepts such as personal data and data processing, the scale, opacity, and complexity of LLM training, post-training adaptation, deployment pipelines and agentic orchestration architectures may complicate the effective application of Convention 108+ safeguards in practice.

At model level, it may be difficult to determine whether personal data are present within training datasets, how such data are processed, or whether they remain memorised within model parameters. At system level, dynamic integrations, retrieval mechanisms, memory functions, and continuous optimisation practices may further complicate accountability, transparency, and the practical exercise of data subjects' rights (see Figure 2 *Data Flows across the lifecycle stages of LLM-based systems*).

These challenges become particularly significant in relation to purpose limitation, data minimisation, fairness, transparency, data security, and the protection of special categories of data. Informing data subjects, identifying accountable actors, and implementing effective safeguards may become increasingly complex in layered and distributed AI ecosystems.

The probabilistic and adaptive nature of Agentic and LLM-based systems may also complicate the practical implementation of rights relating to rectification, objection, explanation, or meaningful human intervention. The absence of observable data traces, explainability mechanisms, or accessible

user interfaces may further reduce individuals' ability to understand when and how their personal data have been processed.

These developments raise important questions concerning the operationalisation of Convention 108+ obligations in increasingly dynamic, adaptive, and interconnected AI environments.

3.2.1 Article 10 – Additional Obligations: Risk Assessment, Privacy by Design, and Risk Prevention

Article 10 of Convention 108+ constitutes a foundational provision for the governance of privacy and data protection risks in the context of LLM-based systems and Agentic architectures. Article 10.2 requires controllers and, where applicable, processors to examine the likely impact of intended processing on the rights and fundamental freedoms of individuals prior to processing and to design processing operations in a manner that prevents or minimises risks.

In the context of LLM-based systems, this obligation requires lifecycle-oriented assessments capable of addressing both model-level and system-level risks, including risks associated with training data collection, inferential processing, post-training adaptation, retrieval mechanisms, deployment environments, agentic orchestration, downstream integrations, and post-deployment monitoring.

Continuous and iterative risk governance and measures

Given the dynamic, adaptive, and interconnected nature of LLM-based systems, Article 10 should be understood as requiring continuous and iterative risk governance throughout the lifecycle of the system rather than a one-time assessment conducted prior to deployment. Privacy by design and by default should therefore be implemented throughout all stages of the lifecycle, including model development, system integration, deployment, optimisation, and post-deployment adaptation.

Technical and organisational measures should remain proportionate to the nature, scope, sensitivity, and potential impact of the processing operations involved. Relevant safeguards should include data filtering and minimisation mechanisms, privacy-enhancing technologies such as differential privacy, secure integration and access control measures, monitoring and auditing mechanisms, human oversight procedures, transparency and documentation practices, and safeguards addressing inferential profiling, manipulative processing, or excessive behavioural prediction.

Impact assessment

Importantly, impact assessment under Article 10 should not be understood as a purely formal compliance exercise or as a mechanism that automatically legitimises processing operations. Where identified risks cannot be adequately prevented or mitigated, certain processing operations, functionalities, deployment environments, or system architectures may need to be modified, restricted, suspended, or abandoned in order to ensure compliance with Convention 108+ and the effective protection of individuals' rights and freedoms.

3.2.2 Article 5 – Legitimacy of Processing and Data Quality

Article 5 requires that processing operations be lawful, fair, proportionate, transparent, and compatible with specified legitimate purposes.

In the context of LLM-based systems, significant challenges may arise concerning the lawfulness of large-scale data collection practices, secondary uses of personal data, inferential processing, and the reuse of interaction data for optimisation, behavioural analysis, or retraining purposes. These challenges may become particularly significant where processing operations involve large-scale web scraping, multimodal aggregation, interconnected services, or adaptive agentic systems capable of continuously collecting and processing interaction data.

Users' awareness

In complex Agentic and LLM-based ecosystems, the opacity of training pipelines, post-training adaptation processes, retrieval mechanisms, and downstream integrations may further complicate the practical implementation of fairness, transparency, and accountability obligations. Individuals may not always be aware of which datasets were used to train or adapt systems, whether their interactions are retained or reused, how personal data are processed across interconnected systems, or how inferential processing, behavioural prediction, or profiling practices may affect them.

A Lifecycle approach

The principles of legitimacy, fairness, proportionality, and purpose limitation therefore require particular attention throughout the lifecycle of LLM-based systems, especially where personal data are processed in ways that may exceed the reasonable expectations of data subjects or involve extensive inferential analysis capable of affecting individuals' rights, freedoms, autonomy, or private life.

Data quality, memorisation, and extraction risks

The principle of data quality under Article 5 becomes particularly significant in relation to memorisation, hallucination, and data extraction risks. Research has demonstrated that LLMs may memorise[11] personal information contained within training datasets, including names, email addresses, phone numbers, signatures, or other identifiable information. Memorised sequences may subsequently be reproduced or extracted[12] through prompting techniques, adversarial strategies, or training data extraction attacks. The likelihood of memorisation may increase where datasets contain repeated sequences, sensitive information appears frequently, models are scaled without appropriate safeguards, filtering and deduplication mechanisms remain insufficient.

---

[11] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In USENIX Security Symposium, volume 267, 2019. Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In 2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020, pages 1314-1331. IEEE, 2020.Huseyin A. Inan, Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. Privacy analysis in language models via training data leakage report. CoRR, abs/ 2101.05405, 2021.

[12] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In the 30th USENIX security symposium (USENIX Security 21) (pp. 2633-2650).

Prompt-based extraction attacks[13] may exploit model response patterns to induce the regurgitation of memorised training data. Such risks raise important questions concerning data minimisation, storage limitation, confidentiality, proportionality, and the lawfulness of processing.

Accuracy and inferential processing of generated outputs
LLM-generated outputs may also contain plausible but inaccurate personal information. Such outputs may affect individuals' reputation, opportunities, or rights even where no direct training data extraction occurs.

The principle of accuracy under Article 5 therefore applies not only to datasets themselves, but also to generated outputs, inferential processing practices, and downstream uses of probabilistic information concerning identifiable individuals.

3.2.3 Article 6 – Special Categories of Data

Article 6 provides that special categories of data shall only be processed where appropriate safeguards are established in law.

Agentic and LLM-based systems may process, generate, access, or infer sensitive information relating to health, biometric characteristics, political opinions, religious beliefs, sexual orientation, emotional states, behavioural vulnerabilities, or other highly sensitive aspects of individuals' private lives. Such processing may occur directly through training datasets, indirectly through inferential processing and behavioural analysis, through multimodal aggregation, via retrieval systems and contextual integrations, and through user interaction and conversational disclosures.

Systems' autonomy related risks
Additional risks may arise where Agentic and LLM-based systems are capable of autonomously accessing, retrieving, combining, or acting upon sensitive information across multiple applications, databases, services, or connected environments. In such contexts, Agentic architectures may significantly expand the scope, persistence, and sensitivity of personal data processing by enabling continuous interaction with external systems, memory functions, cloud services, wearable devices, or other data-rich environments.

Proxy-data for inferencing sensitive information
In certain contexts, seemingly non-sensitive data may function as proxies capable of revealing sensitive characteristics when aggregated and analysed at scale. The combination of multimodal data sources, behavioural analytics, and inferential processing may therefore result in the indirect processing of special categories of data even where such information was not intentionally collected.

Sensitive deployment contexts

[13] A memorized sequence being found in the output of a Chatbot is called in technical terms a regurgitation.

Particular caution should therefore be exercised where Agentic and LLM-based systems are deployed in contexts involving healthcare, education, employment, public administration, law enforcement, financial services, or minors and individuals belonging to vulnerable groups.

Appropriate safeguards may include:

- strict purpose limitation;
- enhanced transparency;
- minimisation of sensitive data processing;
- restrictions on access, retrieval, and downstream sharing of sensitive information;
- additional technical and organisational security measures;
- human oversight and intervention mechanisms;
- restrictions on inferential profiling and behavioural prediction involving multimodal and sensitive data;
- monitoring and auditing of agentic interactions with external systems and databases.

3.2.4 Article 7 – Data Security

Article 7 requires controllers and, where applicable, processors to implement appropriate security measures against risks such as accidental or unauthorised access, destruction, loss, modification, disclosure, or misuse of personal data.

The architecture of LLM-based systems introduces distinctive security challenges arising from training data memorisation, prompt injection attacks, adversarial manipulation, insecure APIs and integrations, retrieval system vulnerabilities, cross-context data leakage, insecure plugin or orchestration mechanisms, and the exposure of sensitive interaction logs or memory systems. These risks may affect both the confidentiality and integrity of personal data processed throughout the lifecycle of LLM-based systems.

Data extraction risks

Research has demonstrated that personal data may in certain circumstances be extracted[14] LLMs through carefully designed prompting strategies or adversarial attacks capable of triggering memorised outputs or unintended disclosures. Security risks may become particularly significant in interconnected agentic systems capable of autonomously accessing external services, executing actions, retrieving information, or interacting with multiple operational environments and data sources.

Technical and organisational protections

In such contexts, appropriate security safeguards should extend beyond traditional cybersecurity measures and include technical and organisational protections adapted to the specific operational characteristics of LLM-based systems. These may include secure integration and authentication mechanisms, robust access controls, compartmentalisation of sensitive environments, monitoring and

[14] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A.F., Ippolito, D., Choquette-Choo, C.A., Wallace, E., Tramèr, F., & Lee, K. (2023). Scalable Extraction of Training Data from (Production) Language Models. ArXiv, abs/2311.17035.

incident response procedures, adversarial testing and red-teaming practices, safeguards against prompt injection and manipulation attacks, as well as encryption and secure storage measures.

Lifecycle and continuous evaluation
Given the adaptive and interconnected nature of LLM-based and Agentic systems, security measures should also remain subject to continuous evaluation and updating throughout the lifecycle of the system in order to address evolving vulnerabilities, emerging attack from LLMs through carefully designed prompting strategies or adversarial attacks capable of triggering memorised outputs.

Agentic systems' Autonomy
Security risks may become particularly significant in interconnected Agentic systems capable of autonomously accessing external services, executing actions, retrieving information, or interacting with multiple operational environments.

3.2.5 Article 8 – Transparency of Processing

Article 8 requires that individuals be informed about the processing of their personal data and be able to understand relevant aspects of such processing.

In the context of LLM-based systems, transparency may become difficult to operationalise due to the opacity of training datasets, probabilistic outputs, dynamic retrieval architectures, post-training adaptation processes, inferential processing practices, and interconnected services or third-party integrations. Individuals may therefore face significant difficulties in understanding which data sources are used, how personal data are processed or retained, whether interactions are reused for optimisation purposes, how inferential profiles are generated, or which actors are involved in processing operations within complex AI ecosystems and lifecycles.

Meaningful information and Interface design
Meaningful transparency in such environments requires more than formal privacy notices. It also requires accessible explanations, intelligible design and user interfaces, contextual information, and practical mechanisms enabling individuals to understand how processing operations may affect them and how their personal data are used across the lifecycle of LLM-based systems.

Anthropomorphic or conversational risks
Particular attention should also be paid to anthropomorphic or conversational interface designs that may create misleading perceptions regarding the nature, capabilities, autonomy, or reliability of LLM-based systems, especially where such systems simulate human-like interaction or emotional engagement.

3.2.6 Article 9 – Rights of Data Subjects

Article 9 establishes a range of rights for individuals, including the right not to be subject to decisions significantly affecting them based solely on automated processing without their views being taken into consideration.

The operational characteristics of Agentic and LLM-based systems may significantly complicate the effective exercise of data subjects' rights, including rights relating to access, rectification, objection, erasure, transparency, meaningful human intervention, and the contestation of automated decisions. Individuals may face difficulties in determining whether their personal data were used during training or post-training adaptation, whether personal information remains memorised within systems, how inferential profiles or behavioural predictions are generated, or which actor is responsible for responding to requests within complex and distributed AI ecosystems. Such challenges may become more significant in systems involving continuous learning mechanisms, memory functions, distributed infrastructures, interconnected services, agentic orchestration, or extensive inferential processing practices. In these contexts, the absence of observable data traces, explainability mechanisms, or clear governance structures may reduce individuals' practical ability to understand, challenge, correct, or delete personal data processed throughout the lifecycle of LLM-based systems (see Figure 1).

Technical and organisational safeguards
Proper implementation of Article 9 in the context of LLM-based systems may therefore require additional technical and organisational safeguards, including accessible mechanisms for exercising rights, enhanced traceability and documentation practices, clear allocation of responsibilities between actors, effective human oversight procedures, safeguards against excessive inferential profiling, and effective complaint and redress mechanisms.

3.2.7 Article 14 – Transborder Data Flows

Article 14 provides that transborder flows of personal data may only take place where an appropriate level of protection is ensured, in accordance with the safeguards established under Convention 108+.

The development and deployment of Agentic and LLM-based systems frequently involve cross-border data flows, distributed infrastructures, cloud-based services, remote APIs, and international data-sharing arrangements.

Training datasets, interaction logs, memory systems, retrieval architectures, and optimisation pipelines may involve the transfer of, or remote access to, personal data across multiple jurisdictions and actors throughout the lifecycle of LLM-based systems. In such environments, ensuring an appropriate level of protection for transborder data flows may become particularly complex where processing operations are distributed across jurisdictions, multiple actors participate in deployment and optimisation processes, downstream integrations rely on external services, or cloud-based infrastructures dynamically allocate processing resources.

Opacity of processing and responsibility chains
Additional challenges may arise from the opacity of processing chains, the allocation of responsibilities between controllers and processors, and the difficulty of identifying where personal data, inferential profiles, or interaction histories are stored, accessed, or further processed. Risks may also emerge through onward transfers, secondary processing operations, or the cross-border sharing of sensitive or inferentially derived information within interconnected AI ecosystems.

Agentic systems' autonomy

These considerations become increasingly significant in agentic environments capable of autonomously interacting with multiple services, tools, databases, or external systems across jurisdictions. In such contexts, effective implementation of Article 14 of Convention 108+ may require enhanced transparency regarding international data flows, clear allocation of responsibilities between actors, and safeguards capable of addressing the complexity and evolving nature of distributed LLM-based and agentic systems.