AGENTIC ARTIFICIAL INTELLIGENCE FROM THE PERSPECTIVE OF DATA PROTECTION

V1.1 February 2026

EXECUTIVE SUMMARY

An AI agent is an artificial intelligence system that uses language models to meet a goal. These guidelines are an introduction to the data protection issues that may arise when controllers and processors decide to use AI systems to implement personal data processing.

The purpose of this document is not to analyse the compliance of a specific processing operation that uses AI agents, but how to manage the peculiarities that are incorporated into a processing by the fact that it is fully or partially implemented with agents.

Knowing this technology is key to making informed and evidence-based decisions about its implementation in personal data processing. Knowledge as merely a user is not enough: it is necessary to understand its foundations, capabilities, limits and the way in which it is implemented. Both, the irrational rejection of agentic AI and its uncritical acceptance in the processing of personal data can be harmful. In particular, the opportunities offered by this technology for greater data protection by design and as a PET tool by itself must be proactively seized.

This text is structured by initially making a brief description of what the agentic AI systems are. Next, we will discuss the potential vulnerabilities of these systems that affect data protection, the aspects of compliance with data protection regulations, and the specific threats that different vulnerabilities can exploit. Finally, the document lists measures that could be adopted by a controller or processor to ensure compliance with data protection regulations and reduce or eliminate the impacts that the agentic AI presents in its deployment in personal data processing in relation to the rights and freedoms of data subjects. These analyses will focus on what is most distinctive in the agentic AI as a system in the processing of personal data, beyond the vulnerabilities, threats and measures that are well known to generative artificial intelligences, or other elements that make up these systems.

Keywords: Internet and new technologies, machine learning, artificial intelligence, data protection by design and by default, automated decisions.

INDEX

I. INTRODUCTION

Task automation is the use of technologies to execute repetitive activities without constant human intervention. This approach efficiently transforms processes that were previously done entirely manually, freeing up time for higher-value tasks. The use of automation systems can be found from industrial environments to office environments, including any other productive or service sector.

The development of large language models (LLMs[1]) completely changed the paradigm of automation, giving rise to the concept of agentic AI as AI-based systems with the ability to act autonomously to achieve the fulfilment of objectives: AI agents. The integration of language models represents a qualitative leap in the efficiency and complexity of the tasks that agents can carry out, which opens a universe of possibilities for the improvement of business processes and in Public Administrations. In turn, the use of systems that implement the paradigm of AI (AI agents) working collaboratively to automate multiple processes entails a change in the very conception of the implementation of processes, workflows, as well as the use of generative artificial intelligence (hereinafter GenAI) in the work environment.

The ability of agentic AI systems to operate autonomously, enrich themselves with information from the digital environment and execute complex tasks, introduces new challenges in many aspects, including the workplace, management and control of the organization, resilience, safety and cybersecurity, ethical aspects, the possibility of fraud, on the corporate image, etc., in addition to those related to the protection of personal data. Also, as artificial intelligence systems in themselves and due they are used to implement personal data processing activities, obligations may arise that derive from general regulations, such as the Artificial Intelligence Act[2] or the Data Act, or from specific regulations by scope of application.

This document will introduce the data protection issues that may arise when data controllers and processors decide to use Agentic AI systems to implement personal data processing.

This document will not address the use of AI agents in the domestic sphere (although there may also be regulatory compliance implications), nor aspects of the development or evolution of language models[3]. Nor does it address the issue of AI agents in an organization in which there is no processing of personal data[4].

[1] Although we will refer to LLMs throughout the text, small language models or SMLs have proven their effectiveness in the implementation of several agent use cases.

[2] Art.3.1 of the Artificial Intelligence Act: "AI system" means a machine-based system that is designed to operate with different levels of autonomy and that can be adaptable after deployment, and which, for explicit or implicit purposes, infers from the input information it receives how to generate output results, such as predictions, content, recommendations or decisions, which can influence physical or virtual environments.

[3] Even if the data from the agentic services are used for training artificial intelligences.

[4] For example, in Park, T. (2024). Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework describes an AI multi-agent framework based on LLMs to detect and interpret anomalies in financial market data, with special application to data from the S&P 500 index. The system automates the validation of anomaly alerts by coordinating specialized agents (data conversion, expert analysis, cross-checking and summarization) to improve efficiency and reduce human intervention in financial market surveillance.

AI agents are means, systems, that allow the implementation of personal data processing by introducing greater automation. The same AI agent can be used to implement operations in different personal data processing. On the other hand, an AI agent can be a part of the operations of a processing that includes, in order to implement other operations, the use of other systems or operations performed by a human operator.

The purpose of this document is not to analyse the compliance of a specific processing operation that uses AI agents, but how to manage the peculiarities that are incorporated into a processing by the fact that it is fully or partially implemented with agents. Different processing and different types of agents implemented in such processing could have different data protection implications. In the analysis carried out in this document, these implications will be studied in a generic way, taking into account that they are not inherent, nor are they necessarily part of their nature, to all agents or to all use of agentic AI.

The text is structured by initially making a brief description of what the AI systems are. Next, we will discuss the potential vulnerabilities of these systems that affect data protection compliance, the aspects of compliance with data protection regulations, and the specific threats that different vulnerabilities can exploit. Finally, the document lists measures that could be adopted by a controller or processor to ensure compliance with data protection regulations and reduce or eliminate the impacts that the agentic AI presents in its deployment in processing in relation to the rights and freedoms of data subjects. These analyses will focus on what is most distinctive about Agentic AI as a system, beyond the vulnerabilities, threats, and measures that are well known in the elements that compose it, such as LLMs[5], databases, communications, etc.

All this will be carried out within the constraints imposed by a new technology that is constantly evolving and whose analysis is still under development.

II. AI AGENTS

Digital agents, based on traditional software and control systems, predate the appearance of AI, however, their functionalities were limited compared to what can be achieved with agentic AI.

Agentic AI involves much more than using LLMs. Understanding how it works is essential to create a climate of trust, through evidence, that the appropriate measures and guarantees have been put in place to allow data controllers and processors to get the best out of this technological option.

The text will describe what an agent is and the concept of agentic AI, taking into account that classifications always have a formal character, and that technological reality is a grayscale that, in this case, crosses concepts such as LLM, GenAI and new developments that may occur in the future.

[5] For example, aspects such as the training of LLMs, which require a differentiated analysis, are not assessed.

A. AI AGENT

An AI agent is an artificial intelligence system that uses language models[6] to meet a goal[7]. An AI agent acts appropriately according to their circumstances and objectives, is flexible in the face of changing environments and goals, learns from experience and makes appropriate decisions given their perceptual and computational limitations[8]. To do this, it decomposes complex tasks into subtasks, which are executed in a planned way creating a Chain-of-thoughts, each of them implemented with different tools and that perceive the environment through access to internal and external services.

AI agents could be defined by the following characteristics in a general way (depending on the agent and to a greater or lesser degree):

- Autonomy: being able to operate without constant human intervention.
- Perception of the environment: they process inputs in real time using sensors, interfaces with applications (APIs), cameras, etc., to interpret dynamic contexts. Interaction with the environment makes it possible to avoid the problem of the "static cutting of knowledge"[9] of the LLMs.
- Action: In addition to generating text, code, or multimedia outputs, they can execute external actions, such as sending information, interacting with users, executing code, executing contracts, controlling devices, etc.[10]
- Proactivity: they anticipate needs or problems instead of just reacting, being able to initiate actions on their own.
- Planning and reasoning: they allow planning sequences of actions to meet specific goals, evaluating alternatives and prioritizing optimal results.
- Memory and Adaptability[11]: They allow you to define the context, accumulate experiences, adjust behaviours to the user's reactions and improve iteratively through feedback or self-evaluation with short and long-term memory.

[6] In general, large language models (LLMs) as a widely accepted term could be other types of language models, including multimodal models or MLLMs.

[7] ISO/IEC DIS 22989 3.1.1 Agent: An automated entity that senses the environment and executes actions to achieve its objectives. Note 1: An AI agent is an agent that maximizes the probability of successfully achieving its goals through the use of AI techniques.

[8] Russell and Novig *Artificial Intelligence: A Modern Approach, 4th ed., p. 34*

[9] Fixed deadline until which the model was trained or adjusted, limiting its knowledge to information available before that point. To overcome this, AI chat began to implement what would be the foundations of the future agentic AI: Internet search tools, RAG or short-term memory.

[10] For both perception and action, there are standardized protocols such as the MCP (Model Context Protocol) that allows client (agent)/server connection (the service to which it connects) or A2A (Agent to Agent) that allows communication between agents.

[11] In the literature, we speak of "learning", which can lead to confusion, in the sense that the LLMs(s) that are part of the agent are being retrained. The "learning" of the agent is not done by retraining the LLM. Although the information may be used to improve the LLM, it is not a characteristic of the agent, and in many cases it will not be performed.

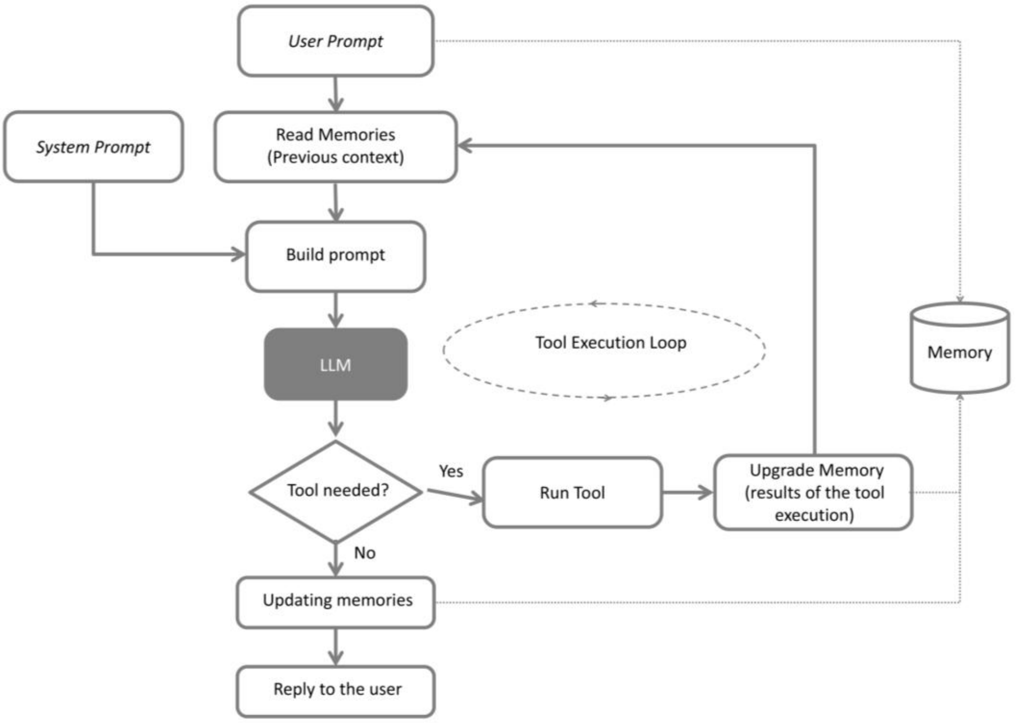
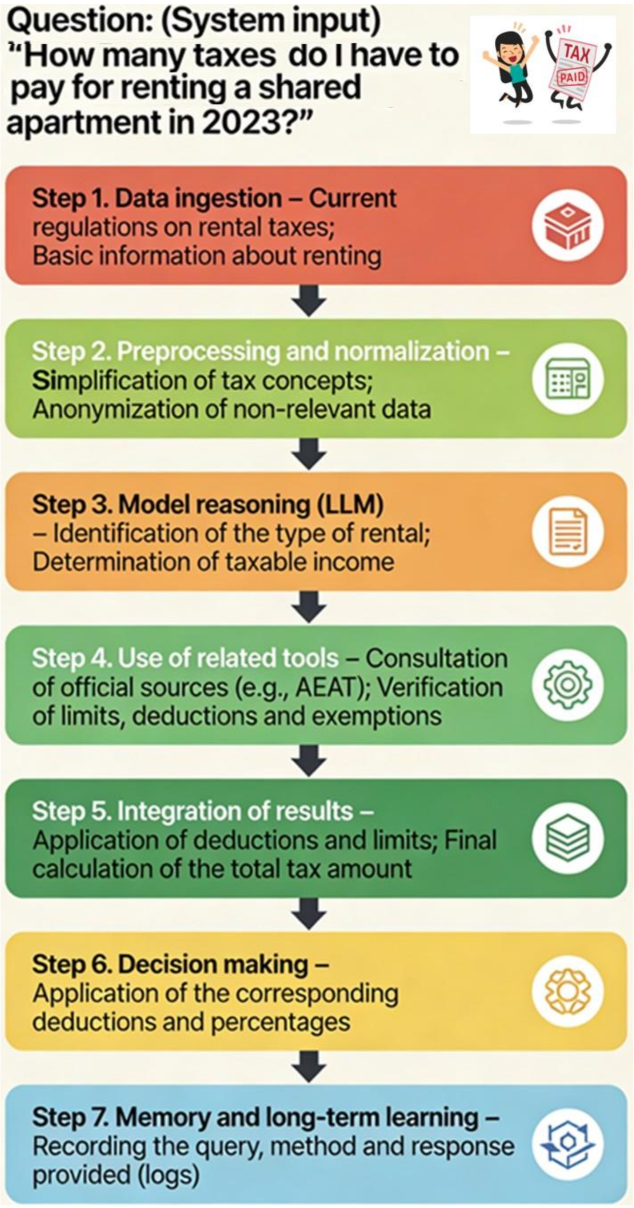


Figure 1 Example of a Basic AI Agent Implementation

B. THE CHAIN-OF-THOUGHTS

The Chain-of-thoughts, *pipeline* or processing, is the internal process by which the agent breaks down a problem into successive and chained logical steps, until reaching a final decision or answer. This chain can be short or, in agents that grow in complexity, very long (known as a long *pipeline*) with multiple stages. Each of these stages may involve different systems, formats, and levels of trust.



Question: (System input)
"How many taxes do I have to pay for renting a shared apartment in 2023?"

Step 1. Data ingestion – Current regulations on rental taxes; Basic information about renting

Step 2. Preprocessing and normalization – Simplification of tax concepts; Anonymization of non-relevant data

Step 3. Model reasoning (LLM) – Identification of the type of rental; Determination of taxable income

Step 4. Use of related tools – Consultation of official sources (e.g., AEAT); Verification of limits, deductions and exemptions

Step 5. Integration of results – Application of deductions and limits; Final calculation of the total tax amount

Step 6. Decision making – Application of the corresponding deductions and percentages

Step 7. Memory and long-term learning – Recording the query, method and response provided (logs)

Figure 2 Example of a Chain-of-thoughts

The flexibility of the Chain-of-thoughts can range from a rigid coded plan or finite-state machines to conversational models where decisions depend on interactions and reasoning models.

In the latter case is when LLMs appear as one of the core components of AI agents. Different types of LLMs and AIGs can appear in an AI agent with different purposes[12]: knowledge capabilities, content generation (such as translators, transcribers, etc.) and reasoning elaboration. What is characteristic of AI agents is that they use LLMs as reasoning machines that will direct a complex autonomous action, analysing the user's requests, responding sequentially to inputs, processing the result of different services and/or constructing a final response. Regardless of using LLM services such as content or

[12] But also small SML models or large multimodal MLLM language models

information repository GenAI in agentic AI systems, the distinctive thing is to use them for task decomposition.

Knowing the Chain-of-thoughts will allow us to know the life cycle of the data, the source of the data, the exact date and time of extraction, when, where and by whom its transformation occurs, and when, where, by whom and for what purpose and legitimacy it is uploaded to a repository, used or downloaded from one environment to another repository[13].

C. AI AGENT PATTERNS

The architecture of agents, also called patterns, implements a reasoning framework, which allows complex tasks to be planned and executed, combining natural language processing, symbolic reasoning, interaction with the digital environment and goal-oriented planning, which gives them a degree of operational independence. These patterns can have different configurations.

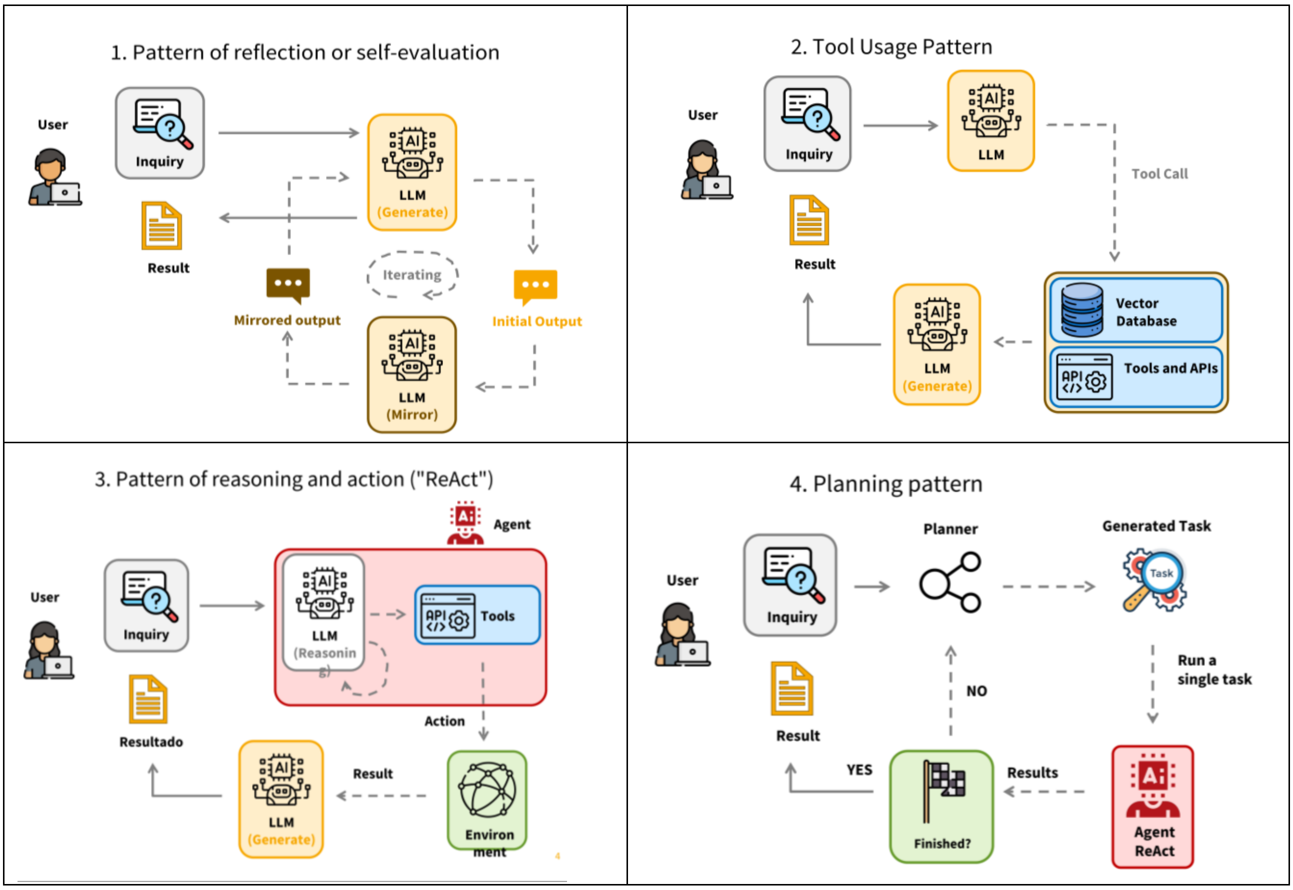


Figure 3 Simplified Representation of Some Pattern Types

[13] This concept is similar to the one that in logistics designs the internal flow of materials for asset control or production flow management. In the case at hand, data is an asset and a resource.

In this way, AI agents can automate repetitive data processing tasks, analyse information to support human decision-making, or interact directly with third-party users and other digital systems.

Unlike LLMs, which are reactive to user actions, agents can be proactive, using tool calls in the background to get up-to-date information, initiate operations, optimize workflows, and autonomously create subtasks to achieve complex goals.

One of the defining characteristics of AI agents is their ability to make service calls, i.e., connect to an API, database, websites, or other tools, and use it as needed. These services can be both remote (e.g., web pages or services) and local (e.g., applications, code execution capabilities, and data stored on the user's system).

Although artificial intelligence agents operate autonomously in their decision-making processes, they depend on goals and rules previously defined by people. An autonomous agent's behaviour is essentially determined by three factors: the team of developers that designs and tunes the agent's AI system; the team in charge of its deployment and configuration; and, finally, the user himself, who defines the specific objectives that the agent must meet and the tools he can access to do so.

D. MULTI-AGENT

Multi-agent architecture combines multiple agents, where each agent's behaviour and responsibilities are strictly defined, they share information and decisions, and they are able to collaborate, compete, or negotiate with each other to achieve more elaborate goals.

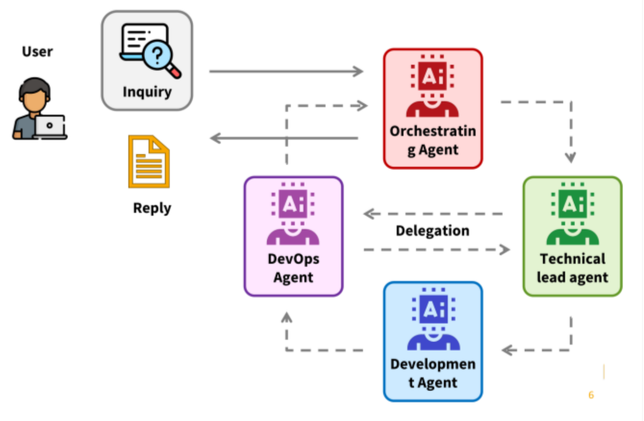


Figure 4. A simplified example of multi-agent architecture

There are several approaches to AI multi-agent: centralized, agent sequential execution, distributed or hierarchical models. Each agent may have a different range of action (tasks that can be executed and tools that can be invoked) and autonomy.

In the first case, a central planning agent coordinates the agent workflow, while the operational agents execute their assigned portions of the task, maintaining their relative autonomy. In any case, you will always need an orchestration layer that coordinates the agent lifecycle, manages dependencies, assigns roles to each agent, sets limitations by domains, and resolves conflicts.

E. DETAIL OF THE ARCHITECTURE OF A MULTI-AGENT

The general architecture of an agentic AI system is a factor that must be known in order to carry out an analysis of its possibilities, limitations and vulnerabilities. The components of an Agentic AI system could be:

- An application that manages the interface to perform tasks for or on behalf of the user.
- One or more agents that will implement different reasoning patterns and techniques, such as rule-based logic, deterministic workflow engines, planning graphs, function calls, or prompt chaining that generally accept natural language inputs, similar to those used by NLP (Natural Language Processing) models. These inputs can be textual *prompts* and other content such as files, images, sound, or video.
- One or more LLMs (local or remote) are used for reasoning, final or intermediate content generation, memory management, and instructions for services.
- Services, including built-in features, local tools, and application code, as well as local or remote services.
- Interfaces for access and interconnection with external tools and services (if necessary): Internet, sensors, actuators, etc.
- External storage for persistent long-term memory and short-term memory, including other data sources, such as vector databases, object storage repositories, and content used in *Retrieval Augmented Generation* (RAG).
- Support services, which are part of the agent's infrastructure, such as credential management, access control, action traceability, etc.

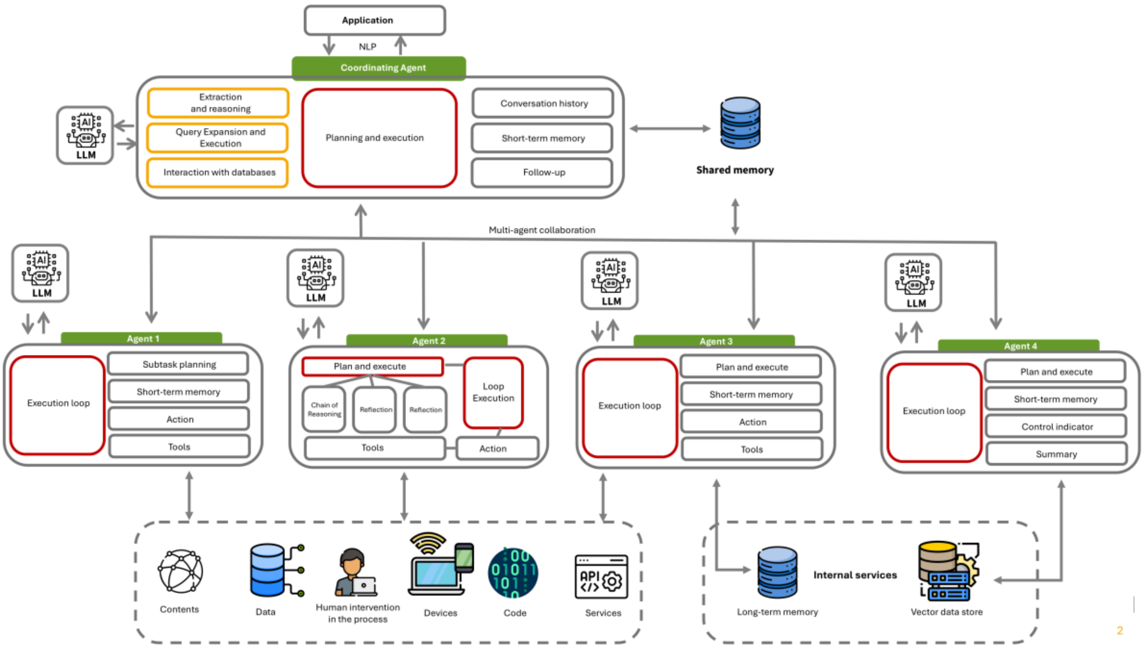


Figure 4 Detail of the architecture of an AI multi-agent system

All these elements could be implemented locally, without access to external services, or completely externally, by accessing an agentic AI service provided by another entity. Between these two extremes we could find any type of configuration that the controller decides, with agents whose application is on-premises, but where part of the memory is in the cloud, with internal LLMs and external LLM services simultaneously, etc.

III. AI AGENTS IN A PERSONAL DATA PROCESSING

AI agents are means that allow the implementation of personal data processing by introducing greater automation. The same AI agent can be used to implement operations in different personal data processing. On the other hand, an AI agent can be a part of the operations of a processing that includes, in order to implement the rest of the operations, the use of other systems or operations performed by a human operator. For example, in the event that human supervision is necessary, in the processing implemented with the means of agentic AI, such intervention will have to be contemplated from the design of the processing activity.

AI agents are means used in a processing that shape its nature, and they can also alter the context, the scope and add additional purposes, also to altering the risks inherent in it. In the case of pre-existing processing, including agentic AI will require a review of its compliance. It may also be the case that an entity initiates new processing from scratch, taking advantage of the opportunities offered by the agentic AI, implementing with this part (or all) of the processing processes.

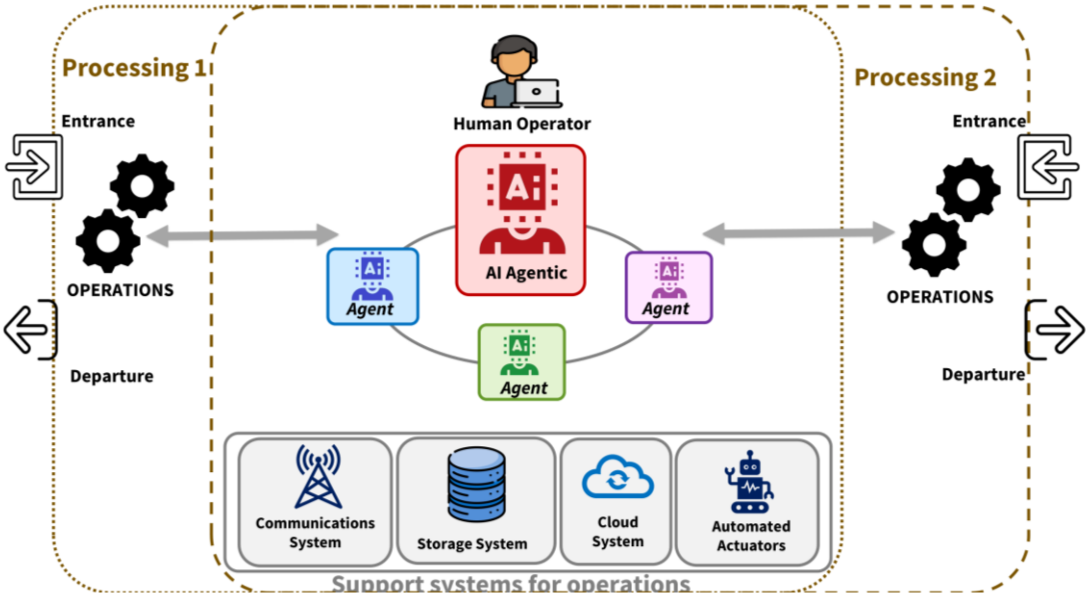


Figure 5 Relationship between agents and processing

The purpose of this document is not to analyse compliance with a specific processing used by AI agents, but rather on distinctive aspects that could arise in relation to data protection due to the fact that it is fully or partially implemented with Agentic AI systems.

When carrying out this analysis, it is important to avoid the "technological fog" that can be caused by the mere naming of an agentic AI when implementing a processing. For example, an agent can implement a common process in any organization such as arranging a trip for an employee. The agent, proactively supported by the GenAI, when detecting a trip in the employee's agenda, would carry out a set of tasks, such as contacting various accommodation services via the Internet, checking the currency exchange rate, verifying the state of the transport routes, managing the services for the acquisition of transport tickets and obtaining an updated weather forecast. Taking into account other factors, (checking the news), he will make a selection and contact the services again, make the appropriate purchases and forward the planning and documentation to the employee.

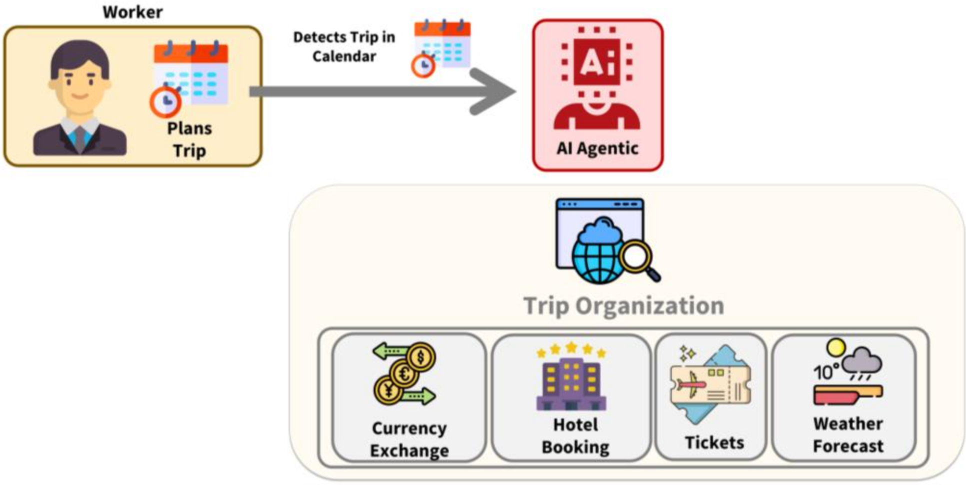


Figure 6 Example of travel management with agentic AI

Traditionally, that processing has been implemented by means of the services of an office assistant who would use the same data and access the same services. Other way has been that the organization would have hired an external travel agency to carry out the same procedures, even with the same proactivity through access to the employee's agenda. The analysis in relation to data protection (purpose, minimization, legitimation, access to Internet services, etc.) will be the same whether it is implemented with an office assistant, or with the travel agency hired as a data processor, or with an agent, including the ability or way to determine that there is a trip on the employee's agenda. Therefore, it can make it easier to start the compliance analysis by landing on entities or individuals the same operations carried out by the agent and their relationship with the services they access (or the external entity relationship provided by an agent-travel agency[14]) and then analysing the different aspects introduced by the agentic AI.

In relation to the latter, the choice of one or another type of agent to be implemented in a processing could have different implications in data protection, even regarding the same agents for different processing[15]. In the analysis carried out in this document, these implications will be studied in a generic way, taking into account that they are not inherent, nor are they necessarily part of their essence, to all agents or to all use of agentic AI.

[14] The guarantees offered by the travel agency, whether it is attended by people or that it is actually an agent, should be the same.

[15] One processing could be, regardless of the means chosen, high-risk or process special categories of data, and another processing, using the same agentic AI as a medium, could be low-risk and not process any special categories.

As described above, agentic AI can involve interaction with numerous internal and external services, via the Internet, which would expose personal data in a processing chain involving not only the controller, but multiple entities under the privacy, cookie, terms of service, and contractual policies of each third-party tool.

Therefore, the following will be analysed:

- What new vulnerabilities, which affect from the point of view of data protection, could be implied by including AI agents in the processing of personal data.
- What aspects of data protection compliance need to be reviewed when considering the use of AI agents.
- What threats can exploit or materialize the vulnerabilities detected with an impact on data protection.
- What measures are in place and available to support regulatory compliance, avoid critical impacts, or manage risk.

IV. VULNERABILITIES AND PROCESSING OF PERSONAL DATA

In this chapter we are going to carry out a preliminary analysis of the most important vulnerabilities that could arise in a processing for implementing operations with Agentic AI. This analysis is not exhaustive, among other factors because it is focused on those that may have an impact on the processing of personal data and are characteristic of the agent system as a whole, not of its individual components.

In the power and versatility of agentic AI lies, as in any complex system, could be the main vulnerabilities.

Vulnerability is defined as the weakness of an asset that can materialize or be exploited by a threat, potentially causing an impact[16], in the case at hand, in relation to the protection of personal data.

An agentic artificial intelligence system integrates various software components, such as language models[17], databases, planning engines, and other analytical tools. It also includes both internal and external interfaces that interact with multiple services, which, in turn, can have their own levels of connectivity. Consequently, all the vulnerabilities inherent in each of these systems are part of the vulnerabilities of the agentic AI.

However, it would be inappropriate to adopt a merely additive perspective, since the interaction between the different components can give rise to new vulnerabilities or amplify existing ones, generating multiplicative effects. In short, this type of system introduces a significantly wider attack surface than that of language models, exposing the system to impacts and threats of greater complexity.

[16] ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements

[17] From SMLs, LLMs, to MLLMs.

A. INTERACTION WITH THE ENVIRONMENT

Within the framework of processing, the agentic AI has the ability to interact with the environment to execute all or part of the processing operations. The interaction can be limited to the organization itself, or it can be extended to external services.

The invocation of tools and services on the Internet are *de facto* partial outputs that the agentic AI is making to the outside. In particular, they are not aimed at the user of the agentic AI and are not the end result, so they could be transparent to those users or to the data controller but contain personal data or may reveal personal information about the data subjects.

- ***Access to organization and user data***

In relation to the previous section, one of the common functionalities of the Agentic AI is access to internal services and data in order to enrich the context for the execution of tasks. This information could be related to the user, a work group or the entire organization. Examples could be email accounts, reports, decisions, internal discussions, meetings, notes, conversations, a customer database, etc. This involves the processing of data of the users of the Agentic AI, which may be personal data of that same user, as well as personal data of other people, both of the people whose data is data subject, and of other people whose data reside in the repositories accessed by the Agentic AI.

Uncontrolled access that does not take into account not only the entity's data compartmentalization policies, but also the data protection obligations by design and by default, could lead to massive data processing that would go against the principle of minimization, restriction of processing and accuracy of information if they are obsolete data or there are integrity problems. If all or part of the components of the AI Agent are implemented by processors, it could involve the communication of data to third parties beyond the purposes of the processing. In addition, in access to unstructured data sets, some of the information may be relevant, but some information may be irrelevant or inadequate from different aspects.

- ***Ability to perceive and act externally to the organization***

Interconnection to Internet services allows agents to interact with the environment outside the organization both to collect information and to send information (requests, commands or locally stored data), increasing their capacities for both action and information processing.

The existence of two-way data communications with multiple participants, without the necessary control of the entity, can significantly increase vulnerabilities such as being able to access the control of the agentic AI through multiple channels.

With regard to local information that is sent abroad, excessive freedom in the invocation of tools that collect internal information could lead to unnecessary communication of information by not preparing the agent to discern what information is relevant from what is not.

The connection with the outside world not only to execute actions, but also to obtain information could be using inadequate, inaccurate, unrealistic, obsolete, partial, biased or uninformed sources. Above all, if there are no procedures for verifying the reliability, provenance and coherence of the sources used. Similarly, if the information request commands have not been properly prepared, excessive personal data that is not relevant to the processing may be being collected.

B. Service Integration

Agentic AI is based on the integration of multiple services. As part of the Agentic AI, it will combine the use of at least one language model, memory management and task execution tools. Externally, the agentic AI must be integrated with other services such as file servers, mail, web services, etc. All of them may be local or external services.

- ***Service Management***

Even when the services come from the same provider, the nature of the industry often causes the independent evolution of each of them, with non-homogeneous terms and contracts, incompatibilities, discontinuity of services and interface changes. This implies greater complexity in the management of tools, both for the ICT services of the organization, the user, and the management of responses by the agents themselves. It also involves the creation of complex data streams and numerous systems that store data at rest in the short and long term.

All of this can pose challenges for data protection compliance, such as managing numerous stakeholders, controlling additional processing, data retention, exercising rights, accuracy issues, etc. Also, other functional problems such as integration of heterogeneous APIs, variable latencies, name collisions, nested parameters, misinterpreted dependency on it, confusion of models when creating *prompts* that are confusing, availability, resilience, instabilities and lack of robustness, appearance of cyber-vulnerabilities, inconsistency in access and use privileges, instability of service quality, etc.

- ***Ease of deploying Agentic AI services***

There are agentic AI services that are easy to deploy, intuitive, with tools that allow task design and connectivity between components very quickly even for end users. These types of environments are common in the development of software prototypes in other contexts and facilitate the deployment of systems such as AI agents.

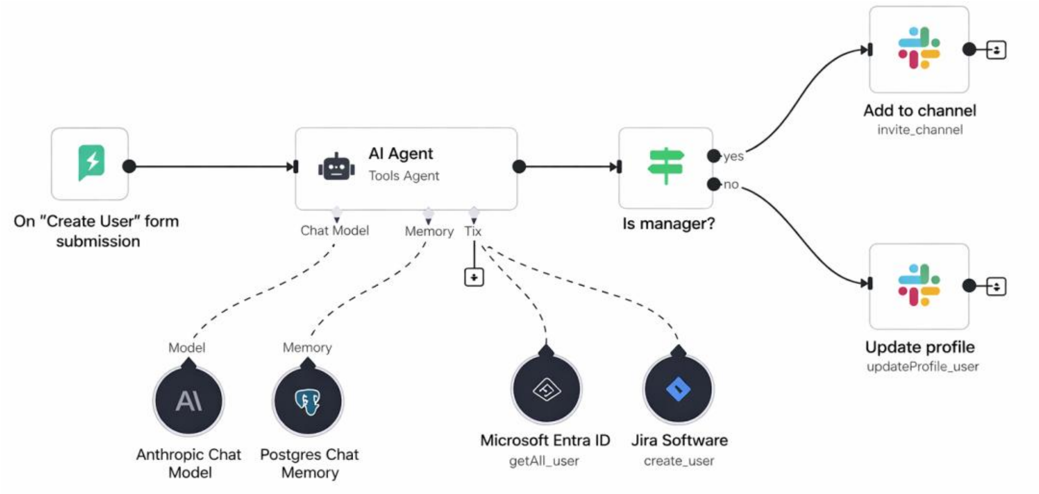


Figure 7 n8n development environment (source: https://n8n.io)

This leads to the temptation for unqualified users to be dazzled by its possibilities and carry out deployments outside the governance and information policies of the entity. The ease of having a solution that seems to work with little effort could generate a sense of banality of the implications and impacts for data protection (and for the organization in general) by obscuring the inherent complexity that these developments have in many aspects.

Introducing an agentic AI system in the processing operations of a data controller implies redesigning a process of the organization in which at least the functional, ICT and quality managers should intervene, in addition to the DPO when appropriate.

The impacts of errors in the deployment of the Agentic AI system in a processing can affect everything from the actual efficacy to regulatory compliance, reliability, explainability, stability and robustness of the processes, their scalability and availability, the vulnerabilities that are generated in the processing, the lack of control of data flows, etc. the extent and retention of such data, the consequences of breaches, lack of preparedness for incident management, etc.[18]

If with mobile devices the problem of BYOD (*Bring Your Own Device*) appeared in the workplace , and with AI chat that of BYOAI (*Bring Your Own Artificial Intelligence*), with the Agentic AI appears the problem of BYOAgentic (Build Your Own Agentic), by not having the policies of the organization and the professionals qualified in management and technically, and the use of mature methodologies in the design of processes and applications.

C. MEMORY

Memory is one of the great advantages of agentic AI and one of its core elements along with LLMs. Memory in AI agents is the ability to store and recall past contexts and experiences to improve decision-making, adaptation, and performance. Unlike systems that operate without context, memory agents can recognize patterns, adapt over time,

[18] A kind of "AI slop" but in process automation.

and use prior feedback, which is key in goal-oriented applications. Language models, on their own[19], do not possess memory; memory must be integrated as an additional component. One of the main challenges is to manage memory efficiently, storing only the relevant information to maintain fast responses and a low level of latency.

There are two very different types of memory in agentic AI. One is the memory that enables agent functionalities. The other is the memory that allows the management of the agentic AI and allows it to implement control mechanisms, and is the one formed by all the records or logs of the operation of the system, of each of the components of the system, and the records of the services accessed by the agentic AI.

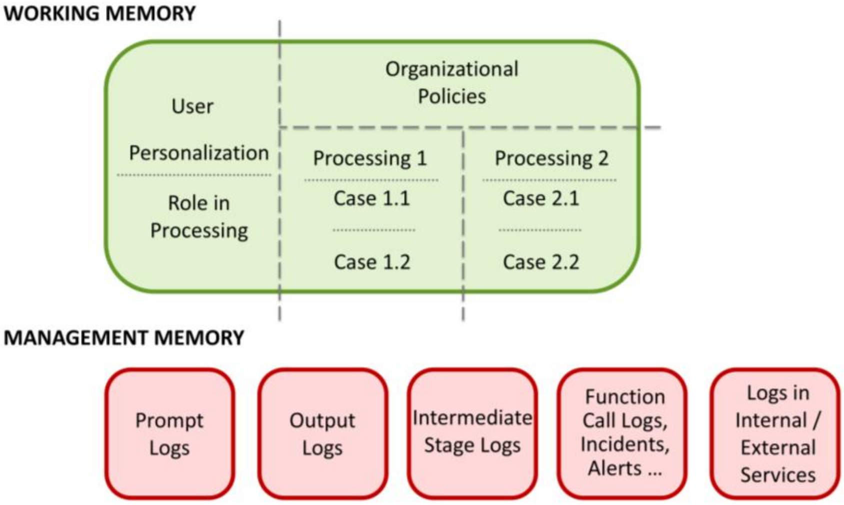


Figure 8 Memory in Agentic AI

- ***Working memory***

Agents use different types of "working" memory, short-term and long-term (depending on the type of agent, we could also speak of medium-term memory). The most simplified approach is that short-term memory allows previous interactions to be recalled within an execution cycle[20], and long-term memory allows systems to retain information throughout different conversations or sessions.

Long-term memory can be further categorized as:

- Semantic memory: involves the retention of specific facts and concepts and can be used to personalize applications by remembering facts from past interactions by creating a continuously updated "profile", with specific information about the user.
- Episodic memory: allows you to remember past events or actions and is used for the agent to remember how to perform a task correctly. It can be implemented through *few-shot learning*, where agents learn from past sequences that are used as examples.

[19] Currently, services that provide access to LLMs do incorporate memory by evolving into the concept of agents, but, formally, a transformer does not have memory as we understand it here.

[20] For example, the user's prompt history in conversational agents.

- Procedural memory involves remembering the rules used to perform tasks. An effective approach to refining these instructions is reflection or *metaprompts* using an GenAI, where the agent fine-tunes their own instructions based on their interactions.

The specific implementation and techniques used to maintain these memories can be very diverse:

- Files, SQL or vector databases.
- Split into chunks and context windows that can handle complex inputs without getting lost, focusing on the most relevant parts.
- Incorporate metadata and tagging (dating, users, categories, etc.) to quickly filter the necessary information.
- A retrieval-augmented generation (RAG) technique that allows you to query a store of knowledge for relevant context before the agent formulates a response.
- Memory optimization techniques: generation of information summaries to save space, analysis and selection of relevant information, etc.

From the point of view of the implementation of a processing in the organization using agentic AI systems, the information stored in memory could be classified into:

- Organisational memory for all processing: which is the information that the organisation considers relevant to be able to carry out automation in the organisation. This unique context may be important in relation to data protection to ensure consistency and completeness (see chapter on Measures).
- Memory for each specific processing, which is relevant to a single processing and not to a different processing. It can also contain specific context information set by the organization.
- Memory for each case treated in the processing, such as a processing that provides a procedure to a client and that is not relevant to another case (depending on the processing, an approximation could be given by case or by client).
- Memory of the user for the same processing, which can be aspects of personalization or also categorized by processing and by case.

The logical organization of the memory could take different forms, from a large repository in which data of the users of the agentic and the personal data of each processing are dumped, leaving the agentic to control which data it is going to use at any given time:

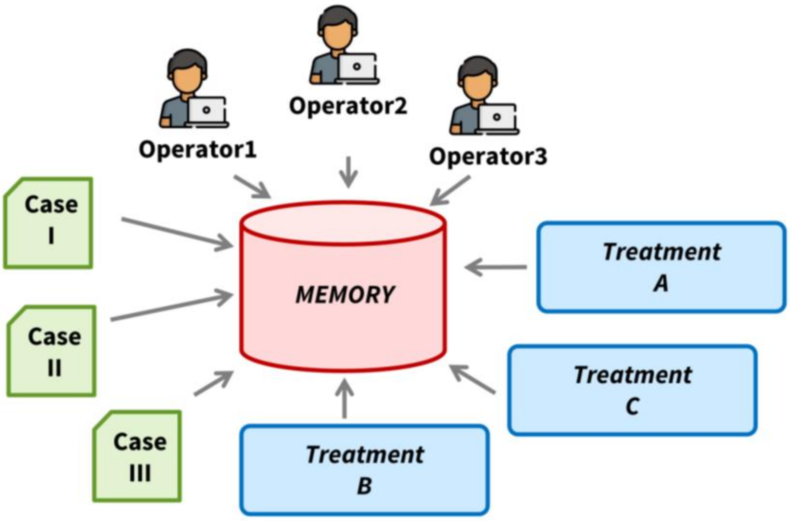


Figure 9 Organizing Memory as a Single Logical Repository

At the other extreme, the memory can be divided logically (or physically) for each processing, in turn for each case and for each user involved in each case:

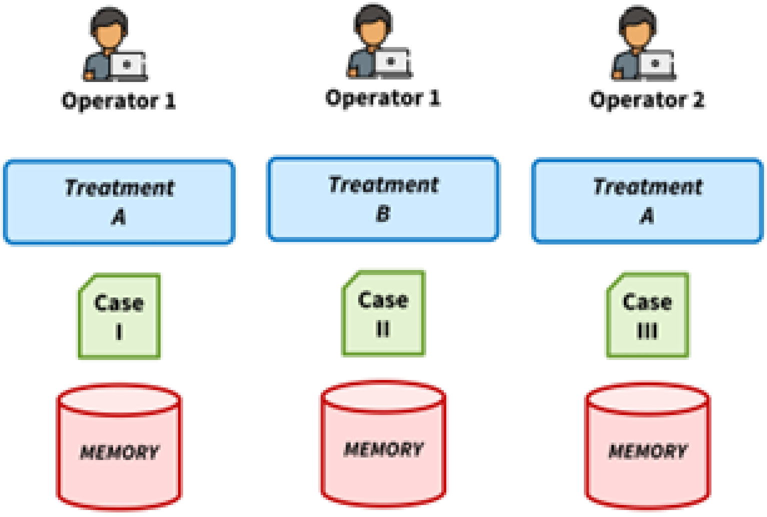


Figure 10 Fully granular distribution of memory

Between these two extremes, several compromise solutions can be found that are adapted to the needs of the controller and of each processing.

Memory, as well as a great advantage, can present vulnerabilities in relation to data protection such as:

- Relevance: it is necessary to establish clear and effective policies of what is to be stored in memory for each processing. Relevance can be described with *prompts*, when analysed with an LLM, or with other types of techniques. Among other things, it would be necessary to ensure that there is compartmentalization between contexts of different processing (at least), for example, that user credentials that have been provided for one purpose are not accessible in the context of another purpose in which a credential request [21]appears.

- Consistency and completeness of the context: the information stored must be of sufficient quality (including in relation to bias, relevant to that context, up-to-date, without contradictions), in particular, if it is going to infer or make decisions about people. Both in long-term memory and in summaries made in short-term memory[22].
- Retention: the information stored must be the minimum necessary for the agent's operation. This is not only information in relation to trade secrets or industrial property, but in particular the personal information of the user, the data subject or third parties, including information that can infer a profile of any of them.
- Integrity: the information stored allows the result of inferences to be manipulated and the agent's own actions to be changed, so it may suffer manipulations of the context and commands or attack code that affect the confidentiality, integrity or availability of the personal data held by the organization (in addition to other effects that are not within the competence of data protection).

- ***Management report***

We must also take into account the impact that shadow memory can have that is common in the use of digital systems, such as activity logs. This memory also has its role in the operation of the agentic AI in that it has to be exploited to analyse dysfunctions, incidents, attacks, alerts, etc.

The memory stored in the logs can be, depending on how it is used, both a privacy measure and have a critical impact or present a risk.

- Data protection measure by design: by allowing auditability of all actions, traceability, repeatability, accountability, deterrent against abuse, etc.
- Critical impact, for example, when records store information about users that becomes hyper-surveillance and goes beyond preserving privacy and cybersecurity.
- Risk in the event that persons authorized to manage such records do not comply with their duties of confidentiality, unauthorised processing occurs due to personal data breaches or the use of the captured personal information for other purposes (e.g., adjustment of LLMs).

A unique aspect appears when an agentic AI system is used to implement different processing. In this case, some of its components, for example the LLMs, will use logs where they will store the activity of all the processing. For example, they will store the prompts and inferences made about all of them, becoming nodes for the collection of personal information of the people whose data is subject to multiple processing, but it could also store data relating to users of agentic AI systems, as well as data concerning

[21] Regardless of the existence of other possible controls.
[22] They may also be applied to medium and long-term memory.

third parties unrelated to the purpose of the processing that may have been collected from various sources.

This could have a greater impact when such components are services external to the controller's infrastructure and managed by third parties, for example, when the LLM is used as an external service. In any case, the risk of profiling of data subjects and the impact in the case of personal data breaches increases.

- ***Exercise of rights***

To the extent that the memory of the Agentic AI system stores personal data within the framework of one or more processing, and also records what accesses or operations are being made on said personal data, it must contemplate from the design the capacity to exercise all data subject rights of the GDPR, including access, rectification, erasure, restriction of processing and the right to object.

D. AUTONOMY

Agentic automation means that agents can act autonomously, without receiving explicit instructions from a human user. This autonomy allows you to decide how the task is going to be executed, into which steps it is subdivided, which internal or external sources to consult, what information to take into account and how it will be taken into account, make decisions, execute tools, make inferences or generate results. This ability gives the Agentic AI a great ability to complete objectives.

Autonomy is significant in the AI agent's performance with the environment: access and updating of data repositories, exchange of data between participants, combination of said data, management of other processes and generation of results, decisions, evaluations or other generative content, all internally to the organization, as well as externally to it.

The level of autonomy of the agent in the processing is a design decision of the controller, and could be[23]:

- The agent proposes, the human operates.
- The agent and the human collaborate.
- The agent operates, the human is consulted or approved.
- The agent operates, the human observes.

[23] In Feng et al. Levels of Autonomy for AI Agents (2025) https://arxiv.org/abs/2506.12469, five levels are proposed.

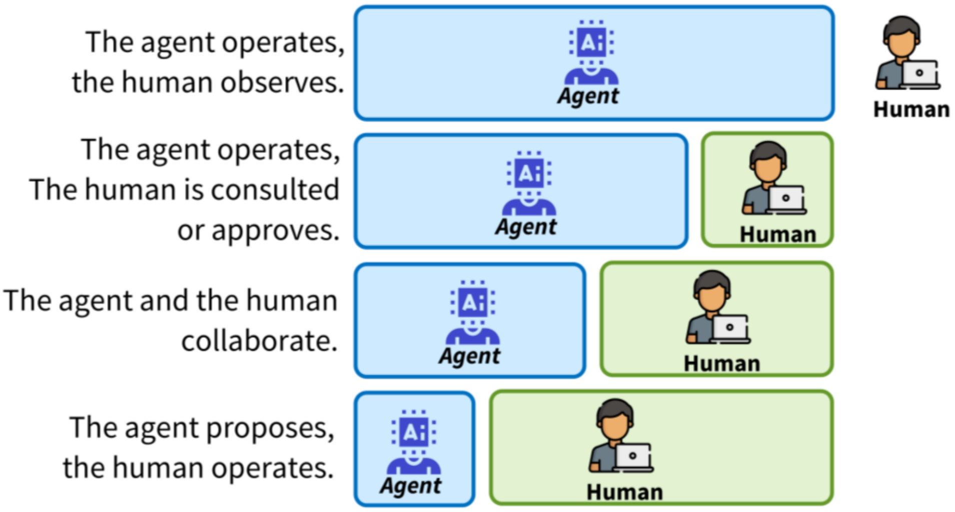

The agent operates, the human observes.
Agent
Human
The agent operates, The human is consulted or approves.
Agent
Human
The agent and the human collaborate.
Agent
Human
The agent proposes, the human operates.
Agent
Human

Figure 11 Levels of Agent Autonomy

From the point of view of personal data protection, there are several aspects that could be affected by this autonomy:

- If the data accessed, updated or exchanged comply with the principles of minimization, accuracy, and restriction of processing.
- If such actions are automated decisions in accordance with Article 22 of the GDPR.
- Whether such actions have a serious impact on the individual and whether they are reversible within the framework of the processing (actions such as deleting the individual's data from the organisation's systems).
- If there is the necessary human supervision.
- Whether the task has really been organized, subdivided, and executed in the right way to ensure that the processing as a whole actually serves the purpose.
- If there is transparency about its execution: quality of results, explainability, repeatability, traceability, auditability and auditing, among others.
- Whether revocation mechanisms are provided in the agentic AI and/or in the framework of the processing.

- ***Transparency and human oversight***

Users and developers may face difficulties in understanding how some AI agents make decisions. A lack of transparency of internal reasoning processes (since decisions emerge from chains of inference distributed among several agents and tools), and a limited capacity for understanding on the part of insufficiently qualified human operators, can generate apparent confidence, based more on the perception of correct functioning than on objective evidence. This situation gives rise to an illusion of reliability, in which the system appears consistent and effective, despite the lack of solid guarantees about the validity of its results.

Designing systems that constitute black boxes is not exclusive to AI agents, it could even be generated without the inclusion of LLMs. However, the speed and complexity of the decision-making processes of AI agents can generate more pronounced obstacles to achieving significant explainability and the transparency necessary for different objectives: demonstrating effectiveness, evidence of robustness, guarantees for customers, legal protection against responsibilities derived from actions, and, among others, compliance with data protection in terms of citizens' rights.

At the same time, it intensifies the automation bias[24], leading users to accept the system's decisions without sufficient critical analysis, and reinforces the criterion of authority attributed to technology, especially when it operates with a high degree of autonomy.

Finally, human oversight becomes more complex, especially when specific mechanisms have not been designed and implemented that allow, and in some cases compel, effective, continuous, and meaningful oversight.

- ***Task planning and interaction between agents***

The mechanisms of task decomposition and the interaction between multi-agent systems, together with the orchestration of activities between these agents, allows the execution of highly complex tasks adding flexibility and adaptation that allows the agentic AI to solve problems in different contexts.

To the extent that agents are going to implement personal data processing in the organization, it must be ensured that all subtasks are necessary, only necessary, and in the appropriate order, taking into account the impact it may have on the data subjects (and for other objectives of the organization). There will be processing in which decomposition and orchestration will be predefined at least up to a certain level. In others, a reasoning-oriented LLM can do all the decomposition.

Note that an LLM does not "reason," but rather extracts task decomposition models that have been included as input data in its training process[25]. If they are to be used for very complex tasks, it is important to ensure that the LLM or SLM has been trained for it. On the other hand, that there is no possibility of contamination between different non-compatible learned models (for example, subtasks of administrative procedures of different jurisdictions).

Technical complexity can lead to instability in emergent behaviour, that is, to unpredictable and undesirable dynamics typical of complex systems, which cannot be anticipated or explained only by the analysis of their individual components. As a result, unforeseen results or infinite planning loops can occur.

In addition, reliance on sequential calls to external tools can lead to back-and-forth loops that accumulate latency, especially in multi-step tasks where each phase depends on the results of the previous one.

[24] Elin Bahner, Anke-Dorothea Hüper, Dietrich Manzey, Misuse of automated decision aids: Complacency, automation bias and the impact of training experience, International Journal of Human-Computer Studies, Volume 66, Issue 9, 2008, Pages 688-699, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2008.06.001. (https://www.sciencedirect.com/science/article/pii/S1071581908000724)

[25] Chengshuai Zhao et al. "Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens" 2026 https://arxiv.org/pdf/2508.01191

The unavailability of a single vendor, or the provision of inconsistent data by the vendor, can trigger cascading failures that halt critical operations, compromising system autonomy and operational continuity.

Compounding *errors* are the phenomenon in which the accuracy of an AI agent decreases as a task requires more steps. For example, an AI agent queries a database about a subject with a poorly crafted *query*, receives incomplete data, processes it as complete, and makes erroneous inferences, which cause the execution of wrong tasks, etc.

In relation to composite errors, both from internal sources, external sources or intermediate inference results, they can generate information or lines of reasoning that do not conform to the entity's policies or regulatory requirements, such as access to excessive or insufficient quality data, erroneous inferences, trade secrets, ethical values of the organization, objectives, biases, financial information and, among others, data protection considerations. All this information can produce results that deviate from the purpose and create damage to users, organizations, customers or citizens.

One of the most critical vulnerabilities lies in the existence of a single point of compromise (SPOC). Given that these systems are made up of interdependent agents, with a distributed collaboration procedure or centralized planning, which could communicate through shared memory or messaging protocols, the violation of a single element of the aforementioned can compromise all the processing that make use of said system.

- ***Non-repeatable behavior***

Inference is the process by which large language models (LLMs) and other GenAI models, including agent-based systems, make decisions. In classical machine learning, *one-shot* inference implies that a given input data produces a reproducible and deterministic output, as long as control is maintained over the model and the sources that feed the input *prompts*.

In more complex systems, such as agentic artificial intelligence, if there is no strict control over the sources of information, the services accessed, their versions, the task schedule, the memory and the possible commands generated from all these elements, it is not possible to anticipate the output of the system.

This problem is not inherent to the nature of agentic artificial intelligence, but responds to specific implementations in which there is no adequate control of the built system. However, maintaining such control is more complex due to the greater complexity of the system itself.

The agent can generate chains of reasoning and sequences of actions in which small variations in an uncontrolled environment (such as a different token or a delay in an API call) that deviate from the entire plan, producing different paths in each execution. In this context, reasoning loops or infinite loops, opportunistic changes in strategy, and emergent behaviours that were not explicitly programmed may appear.

Consequently, unpredictable inference in agentic AI systems is a relevant problem, since it prevents anticipating and accurately controlling the behaviour of the system when it acts in multiple steps, uses external tools and feeds back on its own results. This situation breaks many of the classic guarantees of security, responsibility, accountability and regulatory compliance that were assumed in traditionally controlled software.

In addition, the lack of reproducibility of errors makes debugging difficult in environments that depend on multiple components, as well as the implementation of security for people (*safety*), legal security for organizations and protection against attacks (*security* and *cybersecurity*) policies.

Finally, *ex-ante verification,* including audits, regression tests and regulatory validation processes, becomes fragile, since the same input data can generate, over time, very different sequences of actions, which also leads to less transparency in the final results.

- ***Ability to act on behalf of the user or organization***

Many of the use cases for Agentic AI systems would not be able to implement automation effectively if they could only interact with services, internal or external, that do not require authentication. Therefore, AI agents must be able to request and use user credentials (e.g., to access their mail or information in the cloud) and technical or machine credentials (e.g., to access corporate accounts in LLMs or external services[26]).

Acting on behalf of the user is not limited to the use of credentials. An indirect way of granting privileges that exists in some agentic AI is to incorporate tools that allow the agent to control the cursor and view the user's screen, accessing the same data and being able to perform the same actions as the human user.

Excessive permissions to agents are a critical factor as a broad-privileged model can be used as a pivot point between different systems, so that an initial compromise results in improper access to databases, internal services, or sensitive credentials, amplifying the impact of the incident.

In addition, the absence of isolation mechanisms between services significantly increases the risk of remote execution of commands and code with appropriate levels of privilege from untrusted inputs. For example, the agent (in exchange for access to services such as news, by mistake or motivated by an attack) could be giving consents or entering into contracts on behalf of a user or establishing new user-controller relationships.

Finally, the proliferation of machine identities associated with agents, services, and automations generates a high volume of technical accounts that are complex to manage and monitor. This multiplication makes it difficult to effectively enforce access controls, auditing, and privilege revocation, and increases both the risk of insider threats and the

[26] It could be due to automated actions of the organization itself, without linking to a specific user, such as automatic responses, or because internally the agent registers a log of user requests that maps to accesses with an account of the organization.

surface of exposure to external attacks, especially in highly distributed and dynamic environments that are unique to the AI agent.

V. DATA PROTECTION COMPLIANCE ISSUES

A data controller could choose, among the means to be used to implement the processing, one (or several) Agentic AI systems. Whether it is a fully local service, a fully remote service, or any intermediate formula, it will be a system that is part of the technological infrastructure of an entity and could implement operations (including all operations) of one or multiple processing.

When AI agents are used in processing operations, the following questions could arise:

- Appearance of more intervening parties than the data controllers or processors who are originally part of the processing.
- Greater extension in the type and categories of data of the subjects that were subject to the processing, including additional profiling.
- Greater extension of subjects from whom the data is processed, beyond the subjects that should be data subject, which could be collected from the environment that the agent can access or from the agent's own memory.
- Greater processing of data of users (employees of the organization) who interact with AI agents.
- Less transparency in processing.
- Data retention in more controllers, processors, third-parties and systems.
- New purposes.
- Automated actions with an effect on data subjects.
- New impacts or risks to the rights and freedoms of data subjects.
- Others, depending on the processing and the way of implementing the Agentic AI in it.

The above list is not intended to establish that when an agentic AI system is used in a processing all of these circumstances occur. In fact, they are not inherent to the use of agentic AI systems in a processing, but rather the type and form of configuration of the system used, and how measures are implemented within the framework of the processing.

A. DETERMINING PROCESSING RESPONSIBILITIES

Data Controller is the one who, alone or jointly with others, determines the purposes and means of the processing, regardless of the form in which such means, whether they are agentic AI systems or others. The controller in which agentic AI systems are implemented shall have the obligation to:

- Ensure regulatory compliance.
- Manage the new risks in processing that could be generated by the use of agentic AI systems.

- Analyse the proportionality of the critical impacts[27] that could appear from the use of agents.

When an AI agent is run entirely locally, there would be no further liability analysis. However, in most cases, the agentic AI systems will access third-party services outside the organization to fulfill their purpose. These services could be language models, orchestration management, or even all the agentic AI as a service provided by another entity.

AI agents enable the automation of operations in processing with the support of GenAI. To study the relationships of responsibility for a processing, it is necessary to analyse the casuistry that the same controller will find that relates to other entities that would provide the same service when implementing the processing without AI agents.

In this way, it would have to be analysed without prejudice to the fact that it is necessary to take into account the additional data processing that may occur due to the use of the digital components that make up the agentic AI system. The complexity of this assessment will depend on the level of automation achieved within the processing:

- The agent may access third-party services to obtain non-personal information, such as the hours of a service, the value of financial assets, historical data, etc. In addition, the implementation of the agent may not allow the service to link them to a specific user (there are no identifiers, no cookies, no history linked to the specific user since it is filtered in the use of the agent's action). In this case, the entity providing such service will not have any role in the data protection framework without prejudice to other regulatory areas.
- It is possible for the agent to send non-personal information to third-party services to carry out any process: information storage, text translation, analysis of standards, elaboration of reasoning about a task, etc. (where a language model could intervene). If, as in the previous case, the service does not link you to a user, there would be no data protection relationship. However, if you link it to a user for the purpose of providing the processing, for example, to save the context of interactions with the user, you would be a data processor.
- In the above case, if it is the case that personal information is sent within the framework of the processing, these services would act as processors of that processing[28].
- The agent may access third-party services to obtain personal information relating to data subjects or others, for example, access to public administration or entity records. Without prejudice to the legitimacy of access, the relationship that could be established would be a relationship of communication between the controller

[27] An impact with absolute certainty of occurrence is called a critical impact. For example, in a processing of personal data that is legitimate for some reason, recording all communications made by a person has an impact on their rights and freedoms with absolute certainty. If it did not have such legitimacy, it would be a regulatory breach. If the data were leaked through a breach there would be an additional impact so there is a risk. Normally, critical impacts derive from the definition of the processing itself and could also be reduced by taking measures that would make it proportional to the purpose of the processing, in many cases changing the definition of the processing itself, but it is not a risk, but a certainty.

[28] Paragraph 30 of Guidelines 07/2020 on the concepts of "controller" and "processor" in the GDPR of the European Data Protection Board of 7 July 2021

and the controller. However, if the agent is accessing the services of a rental car agency hired by the entity to obtain billing information from an employee, it would be a controller to controller relationship.

- It is possible for the agent to access a third-party service to transmit personal information relating to data subject. In this case, the relationship of the controller with the other entity must be analysed, regardless of the use of the agentic AI systems. For example, if an agent of a health facility who provides services to an insurer in respect of health care expenses that have been incurred within the scope of an insurance contract, contacts the insurance company's service automatically to transmit the data, it would be a controller to controller relationship[29].
- In the event that the agent itself is a service provided by another entity, and to the extent that it processes personal data of users, and/or personal data of customers or citizens within the framework of the original processing of the controller, the entity providing the service will be processor, as in the case of the previous example of travel agent-agency.

In application of the principle of accountability, the controller must design and document the data flows of the processing, identifying for each of the intervening systems the third parties involved and identifying their role within the framework of data protection regulations and that of the rest of the intervening parties.

To the extent that the incorporation of the agentic AI systems implies the relationship with Internet services or services of other entities (whether processors or third parties), the same casuistry will appear that arises when there is no automation in the processing:

- The use of personal data provided in the processing for other purposes unrelated to the original data controller. For example, retraining of LLMs, security, or others.
- The creation of new relationships of responsibility with, where appropriate, users or those whose personal data are being processed. For example, through the user interfaces themselves requesting their consent for other processing.

In the first case, it could be that these additional processing are legitimate. The second case could also be legitimate, to the extent that there is no mistake for the user as to who they are establishing the relationship with and when the agentic AI is not allowed to give consents in an automated way without some kind of control.

In all cases, the data controller must be diligent in controlling these situations and this will depend on the type of Agentic AI solution that is incorporated into the processing. In the event that the agentic AI system is implemented by the controller itself, measures may be taken to configure the AI agents in order to determine which services are going to be accessed (see the chapter on Measures), evaluate contracts or terms of service, review data protection clauses, cookies where appropriate, it must determine the lawfulness of such processing and the guarantees of regulatory compliance, analyse the risk to the data subjects and assess whether the use of said service is proportional or

[29] https://www.aepd.es/preguntas-frecuentes/16-salud/1-salud/FAQ-1617-centros-sanitarios-y-hospitales-que-prestan-servicios-a-aseguradoras-y-mutuas-son-encargados-de-tratamientos-o-responsables

whether it is more convenient to look for alternatives. It is necessary to assess the degree of regulatory compliance of the alternatives analysed, in particular in relation to, among others, Article 28 of the GDPR, international transfers, data retention, etc.

In the event that the entire agentic AI service is entrusted to another entity, the same obligations mentioned above are identified, in addition to those related to the chain of sub-processors.

Depending on the impact of the processing on the rights and freedoms of the data subjects (and other interests of the controller), it will be necessary to collect evidence of compliance beyond the formal requirements, such as by carrying out tests and studying incidents that may have been reported by other controllers.

At this point, it is worth highlighting the opportunity presented by the Agentic AI to guarantee the fulfilment of contracts or terms of service from multiple suppliers proactively and automatically. In this case, the agentic AI will be a PET technology in itself, with application in this field as for any organization that has to manage an environment of multiple services with dynamic updating of legal conditions.

B. TRANSPARENCY

In the event that the use of agentic AI systems in a processing involves additional recipients of the data to those foreseen in the processing itself, which will be the case in many cases, their identity must be duly informed. If, for example, in the processing this means that personal data, either of the users or the data subjects, are sent to an GenAI service of another entity, both categories of persons must be adequately informed.

Likewise, any modification due to the use of agentic AI systems in the processing must be informed of the storage periods of personal data or, when it is not possible to determine it precisely, of the criteria used to establish this period. Also, on whether additional automated decisions occur (see section on Automation of decisions and International Transfers).

When the incorporation in a processing of solutions based on artificial intelligence agents or systems involves the subsequent processing of personal data previously collected for a purpose other than that for which they were obtained, the controller must inform the data subject prior to such further processing about the new purpose and any relevant additional information, in accordance with Article 13(2) GDPR.

Finally, the information must comply with the purpose of the information provided to data subjects as set out in Recital 39 of the GDPR, according to which individuals must be aware of the risks, rules, safeguards and rights relating to the processing of their personal data, as well as the means to exercise those rights.

C. LAWFULNESS OF PROCESSING, MINIMIZATION AND EXCEPTIONS FOR SPECIAL CATEGORIES

The inclusion of agentic AI systems in a processing could imply additional data processing, although not necessarily. For example, if the administrative person who performed the travel management is replaced by an AI agent in the organization itself, and the organization has interfaces with the same services that were manually queried, the result will be the same data processing.

Moreover, with the use of agentic AI, less data processing can be obtained, or guaranteed, as it could be the case that the processing of the user's data when these services are accessed through the Internet has been suppressed, since cookies or profiling could not be carried out. In any case, the use of an AI agent is not a purpose in itself.

If the implementation of the agentic AI systems does not imply additional processing beyond the original processing, it will not be necessary to seek legitimacy for its inclusion in the processing. It should be borne in mind that, as the agentic AI systems are made up of digital systems, some of which are very complex, it will include more cybersecurity processing that will be protected by legitimate interest as long as they are aimed at that purpose, are necessary and proportional.

If there is additional processing, it must have its legal basis established and, in the case of special categories of data, a circumstance that lifts the prohibition. Where the basis is legitimate interest, you will have to pass the assessment that the purposes of that legitimate interest are clearly identified, the processing is necessary for the purposes of the legitimate interest(s) pursued, and assess that the legitimate interest(s) are not overridden by the interests or fundamental rights and freedoms of the data subjects (also referred to as the 'balancing test').

In the case of being based on consent, it could be considered that measures are necessary for the management of said consent (see section on Consent Management).

Data minimisation must be considered from the design of the processing and transferred to the design or configuration of the agents. For example, suppose that in the context of a processing it is necessary to determine whether an employee is on the guest list of an event. To do this, the guest list could be downloaded and the personal data of the employee and, collaterally, of all the others present on the list could be processed. It might be considered whether it is possible to achieve the same purpose without dealing with the entire guest list (for example, by asking the organizer if the employee is on the guest list) or even without exposing any person with "zero-knowledge" strategies. In this example, it has not been established whether the processing is being carried out with a human operator or with an agentic AI. In both cases, minimization depends on how the processing is designed and what instructions have been given (in both cases) for regulatory compliance.

The restriction of processing must also be addressed from the design. For example, if it is a customer service processing in which the action of the agentic AI is initiated in response to a complaint from a natural person, personal data will probably be processed. Part or all of this interaction could be stored in long-term memory, as it may be necessary to store specific cases to give a better response in the future. It is necessary to consider whether it is necessary to store personal data of past customers in the memory that the agentic is going to use to carry out future actions. In this case, it is also necessary to consider the legitimacy of processing personal data of other customers that could be stored in the long-term memory in the elaboration of each execution of the Agentic AI.

D. RECORDING OF PROCESSING ACTIVITIES

The Record of Processing Activities (RoPA) is a fundamental tool for managing compliance with data protection regulations as it is the catalogue of personal data processes.

When it is decided to replace traditional means with automated processing in a processing, an update of the RoPA must be carried out to determine, for example, whether this entails a modification in the categories of personal data to be processed, whether it is necessary to update the information relating to the categories of recipients to whom the personal data have been communicated or are planned to be communicated, including recipients located in third countries or international organisations, whether new transfers of personal data need to be detailed, retention periods changed, or whether the overview of the technical and organisational security measures referred to in Article 32(1) GDPR needs to be updated.

The GDPR requires a minimum amount of information in the RAT, but not a maximum. It is advisable to integrate the RoPA into the process catalogue of the entity's quality control system as a management tool to guarantee and be able to demonstrate compliance. In that case, both the controller and the processor must determine what additional information they may need to include in relation to the agentic AI systems they use to implement processing.

E. EXERCISE OF RIGHTS

The fact of using agentic AI systems in a processing should not imply a reduction in the exercise of rights and the necessary measures must be implemented to guarantee them. This means knowing how the storage and operations of personal data works and foreseeing the measures and procedures for exercising these rights (see section Measures).

It should be noted that the memory of the agentic AI system stores personal data within the framework of one or more processing. In addition, the *logs* will store information about both the users of the agentic AI and the people subject to the processing, it could even store data on people who should not be data subject. The configuration of both memory systems and the Agentic AI must be technically capable by design to allow the exercise of rights of data subjects to be managed.

In the case of records, there is specific information on what accesses are being made to personal information (when a database is consulted). These accesses are made by the components of the agentic AI systems. However, these have originated from:

- The design of the agentic AI, in which the data controller will have had an intervention, at least by the fact of choosing an agentic AI system as a means of implementing processing.

- The configuration of the particular system, for example, with *prompts* defined by the system administration, or scheduled events (by the user or the organization) that initiate automatic operations.
- By *prompts* made by users of the agentic AI, which can trigger many operations on personal data.

In the latter case, it should be taken into account with regard to *prompts* made by natural persons, that they could be the subject of a duly justified request for the right of access.

Access to services external to the organisation that act as data processors may lead to the storage of personal data in their registration files or the possible memories of their own agents. Personal data can also belong to users of the agentic AI, especially when the entire agentic AI system is a service contracted to a controller.

F. AUTOMATED DECISIONS

The automation of decisions and the degree of autonomy of the agent is a matter of the design of the processing, both of technical design factors and of the design of human intervention, as well as its actual implementation. The manager can manage how the decisions produced by the agent or the agentic AI will be handled, what actions will be allowed automated, unsupervised, and also the measures to manage these design decisions (see chapter Measures).

▪ *Article 22 of the GDPR*

The incorporation of agentic AI systems in a processing may involve automation, but it will not always involve automated decisions within the meaning of Article 22 of the GDPR.

There is processing of personal data that does not involve automated decisions, for example, an AI agent can be used in the organization to track and select events over the Internet and to produce summaries and analyses based on the company's objectives and categories of employees, sending them to the devices of these employees based on the interests that they have declared, without this implying an automated decision within the meaning of Article 22.

However, if they exist, the conditions that allow it (Art. 22.2 GDPR) and the measures to be implemented (Art. 22.3 GDPR) and the limitations on the use of special categories of data (Art. 22.4) and decisions on minors (Cons. 71 GDPR) must be assessed. In addition, information on the existence of automated decisions, including profiling (Art. 22, 13(2)(f) and 14(2)(g) GDPR), providing, at least in such cases, meaningful information on the logic applied, as well as on the significance and intended consequences of such processing for the data subject.

Automated decision-making should also be evaluated since, according to Article 22 of the GDPR, the data subject has the right "not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects on him or her or similarly significantly affects him/her". Even if the decision-making process does not affect the legal rights of individuals, it could still fall within the scope of Article 22 if it produces an equivalent or significantly similar effect in its consequences. In order for data processing to significantly affect an individual, the effects of the processing must be significant enough to be worthy of attention.

In other words, the decision must have the potential to[30]:

- significantly affect the circumstances, behaviour or choices of the persons concerned;
- have a prolonged or permanent impact on the data subject; or
- in the most extreme cases, cause the exclusion or discrimination of people.

- ***Other automated actions***

The use of an agentic AI in processing may entail risks regarding the processing of data of natural persons that do not fall within the scope of Article 22 of the GDPR. For example, allowing an AI agent to send information via email or file transfer services can have an impact on the confidentiality of personal data.

This problem, and others that can be caused by automated actions, must be taken into account in the management of the risk to the rights and freedoms of the data subjects. In particular, to introduce from the design the provision of the reversibility of certain actions of AI agents.

G. RISK MANAGEMENT

As in any processing or process that is innovative or that incorporates modifications to its implementation or new technological systems, it is necessary to carry out adequate risk management.

Risk management involves a critical analysis of the future impact of the processing, beyond the context of the organization, to manage potential problems (threats) before they become real problems, that is, materialize. It is a proactive process to govern the uncertainties that threaten the rights and freedoms of data subjects in the processing of personal data: risks must be identified, assessed and prioritised, and then efforts must be coordinated and decisions made to avoid or minimise their likelihood or impact.

Therefore, it exceeds the scope of the system, in this case the Agentic AI system, and encompasses all elements of the processing, whether technical or non-technical. Undoubtedly, including agentic AI as a means of processing introduces new uncertainties. The AEPD recommends the use of the LIINE4DU threat modelling framework[31], which will allow data controllers to identify threats of Linking, Identification, Inaccuracy, Non-repudiation, Exclusion, Detection, Data Breach, Deception, Disclosure and Unawareness/Unintervenability.

[30] Section IV.B of the Article 29 Data Protection Working Party Guidelines on automated individual decisions and profiling for the purposes of Regulation 2016/679. Adopted on 3 October 2017. https://ec.europa.eu/newsroom/article29/items/612053/en

[31] AEPD, "Introduction to LIINE4DU 1.0: A New Methodology for Threat Modeling for Privacy and Data Protection," October 2024. Available in: https://www.aepd.es/guias/nota-tecnica-introduccion-a-liine4du-1-0.pdf

- ***Management of the risks to the rights and freedoms of data subjects***

All management will begin with a risk analysis. This analysis will have to cover the aspects of interest to the organization (financial, fraud, image, *safety*, *security*, process continuity, environmental, etc.) and, among them, the risks to the protection of the rights and freedoms of the data subjects.

Article 24 of the GDPR provides that the controller shall implement appropriate technical and organisational measures in order to ensure and be able to demonstrate compliance taking into account the nature, scope, context and purposes of the processing, as well as risks of varying likelihood and severity to the rights and freedoms of natural persons.

Including an agentic AI system in a processing undoubtedly changes, at least, the nature of the processing and could reduce or increase pre-existing risks, or generate some new ones. This implies that the controller of a processing that includes agentic AI systems must carry out a new cycle of risk management in the processing.

- ***Rule of 2***

A simplified approach to setting a minimum threshold of guarantees that must never be crossed was enunciated in 2021 in relation to the execution of applications in browsers from the perspective of cybersecurity only and was known as the rule of 2[32]. It has subsequently been reformulated for the case of AI agents by different authors[33] taking the following form:

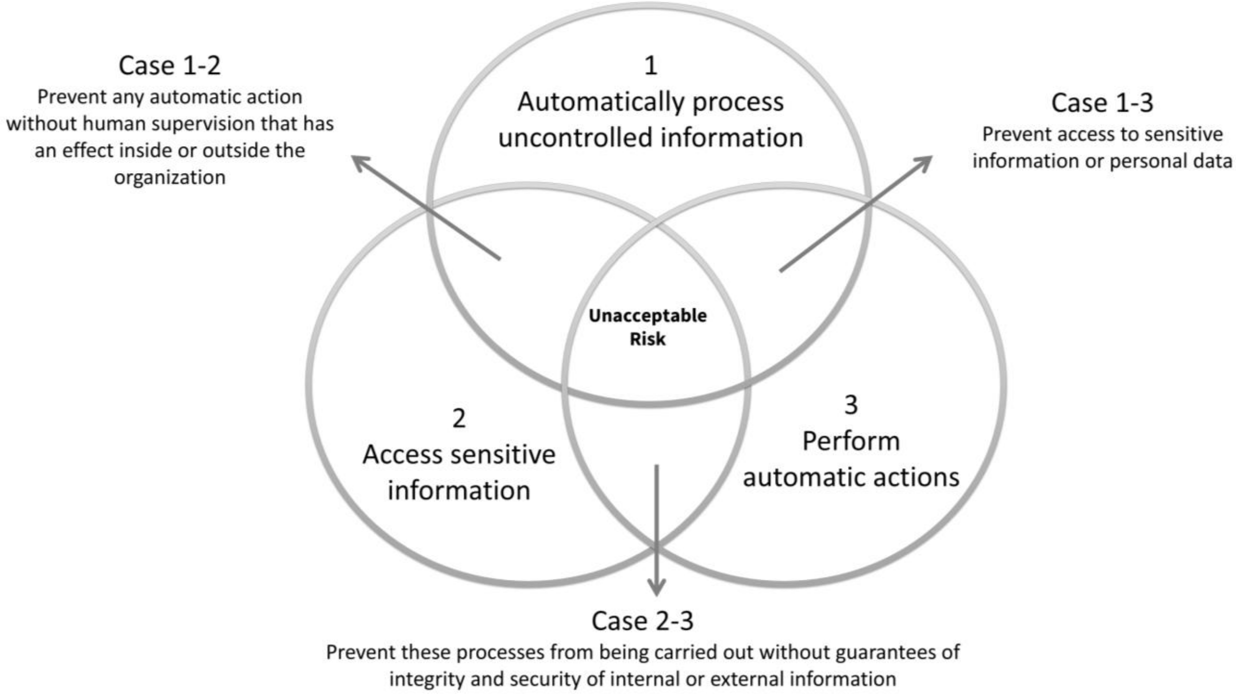

Case 1-2
Prevent any automatic action without human supervision that has an effect inside or outside the organization
1
Automatically process uncontrolled information
Case 1-3
Prevent access to sensitive information or personal data
Unacceptable Risk
2
Access sensitive information
3
Perform automatic actions
Case 2-3
Prevent these processes from being carried out without guarantees of integrity and security of internal or external information

Figure 12 Rule of 2

[32] https://chromium.googlesource.com/chromium/src/+/main/docs/security/rule-of-2.md
[33] For example: https://ai.meta.com/blog/practical-ai-agent-security/

The interpretation of this figure can be explained with a use case: an agent that allows automatic reply to email messages. Following this rule if it were complied with, for example:

- The specific implementation of this AI agent allows you to receive emails without guarantees that there is no type of attack, whether technical or social engineering.
- The agent would be able to access sensitive information on the user's systems without restriction, and
- The agent can then initiate actions automatically (whether it's creating a reply email, manipulating the agent's long-term memory, rewriting sensitive information to other repositories in the organization, etc.)

We would have an agent configuration that should not be allowed.

The Rule of 2 states that, in the best-case scenario, the only configurations that could be managed are:

- Case 1-2: If there is a possibility of automatically processing uncontrolled information that may trigger access to sensitive information, any automatic action without human supervision that has an effect inside or outside the organization should be prevented.
- Case 2-3: If there is a possibility of accessing sensitive information and performing automatic actions, none of these agent processes can be performed without guarantees of integrity and security of internal or external information.
- Case 1-3: If there is a possibility to automatically process uncontrolled information that may trigger automatic actions, the agent must prevent access to sensitive information or personal data.

- ***Processing Risk***

As mentioned, this is a general rule of minimum focused on cybersecurity that can be a good starting point for analysis. From the point of view of data protection, without prejudice to other objectives that must be met by the organization, there are other aspects that should be considered.

For example, that the use of input information to the agent is complete, consistent, up-to-date and free of bias insofar as it may affect, for example, a decision of a natural person. Another example is that the principle of data minimisation is followed when accessing and giving access to possible third parties to personal information. An additional example, in relation to automated actions, are the requirements invoked by data protection regulations on them, such as additional limitations on whether they are based on special categories of data or affect minors.

In short, in relation to the previous examples, risk management must be carried out on the processing in which the agentic system is a means to protect the rights of the data subject, where one part is the management of the security risks of the agentic system itself.

- ***Side effects of processing***

Implementing processing, especially with novel techniques, can cause unwanted side effects that are outside the objectives of the data controller[34]. These collateral effects can occur on the people being processed, or by the processing of data of the users of the agentic AI.

In the example developed above about a customer service, let's assume that the long-term memory of the agentic AI is not compartmentalized. If the personal data of each case related to a query is stored in memory, by injecting prompts by customers themselves or users of the agentic AI (also using "model leakage" techniques see chapter on Threats) some type of personal information could be inferred. This information could be either from customers or from employees who are users of the Agentic AI. The impact that this information could have would depend on the sensitivity of the processing that are implemented with this agentic AI, and would worsen if there is no memory sharing between processing.

- ***Data Protection Impact Assessment***

AI agents are undoubtedly a new technology, but it does not necessarily imply that it entails the obligation to carry out a data protection impact assessment (DPIA) in all cases. It will depend on what processing it is incorporated into and what type of agentic AI system is proposed to be used. It could be the case that when a type of agentic AI system is used for different processing, for some processing a DPIA will not be necessary, for others it will, and for those who already had a DPIA that had been positively passed, it should be [35]reviewed.

- ***Integration into the organization's risk management***

With regard to risk management tasks, while the analysis and determination of the level of risk from the different perspectives mentioned above could be disjointed, actions for risk mitigation must be coordinated. The determination of measures and safeguards, their implementation, maintenance and supervision will be common or interconnected tasks (different objectives, but a single integrated management).

Otherwise, neither the data protection obligations by design nor would be fulfilled, and at least the effectiveness of risk management for rights and freedoms in relation to the processing of personal data would be reduced.

The two chapters that follow are intended to serve as guidance for risk management.

H. DATA PROTECTION BY DESIGN AND BY DEFAULT

Depending on the state of the art, the cost, the nature, scope, context and purposes of the processing, as well as the risks to the rights and freedoms of natural persons, the data

[34] Section VI.C.7 of the AEPD's document "Risk management and impact assessment in the processing of personal data" lists some risks that could arise.

[35] It is recommended to use the GESTIONA GDPR Tool for RoPA management, inventory generation, assessment and risk management for rights and freedoms published by the AEPD.

controller shall apply the appropriate technical and organisational measures to effectively apply the data protection principles, both at the time of determining the means of processing and at the time of the processing itself. The AI agent is a means of implementing processing, and the selection of the type of Agentic AI system and the configuration of the AI system and its components must take into account all these factors from the outset.

The AI agent shall be designed to collect only the data strictly necessary for the processing it supports, to use it exclusively for the stated purpose, to minimize, isolate and protect personal data at every step of the lifecycle (perception, memory, reasoning and action), to maintain full control, traceability and explainability of its operations, and to respect privacy even when acting autonomously without direct human supervision.

The aspects that will need to be managed in particular are data minimisation, avoiding the "default memory" of unnecessary data or uncontrolled user activity records, prohibiting the reuse of data for secondary purposes without legitimation, paying attention in the design to the processing of special categories and their retention or contemplating human supervision, among others. In the chapter of "VII. Measurements" developed below, details techniques to implement data protection by design and by default, as well as to manage risk.

However, the application of data protection techniques by design and by default should be applied beyond purely reactively, i.e. in the event of a circumstance that prevents processing, but also proactively. A proactive application is to introduce measures that will improve the protection of the rights and freedoms of data subjects by the use of agentic AI over other traditional ways of implementing processing. We can see several examples, starting with taking advantage of the introduction of Agentic AI as a reason for greater rationalization of data processing. Also to introduce additional data protection measures that were impossible in manual processing, such as, for example, using SML in conjunction with other systems, in the intermediate stages of the Chain-of-thoughtss for categorization, sanitization, minimization and alerting in data exchanges. Another example, in the stages in which human intervention is necessary (whether it is to sign a decision about a person), anonymized information can be provided in certain aspects to avoid bias in said intervention. Another example would be to make access to sensitive data that is necessary within the framework of the processing, but without being exposed to human operators.

In these reactive and proactive actions, the participation of DPOs and data protection advisors duly qualified in the understanding of these technologies and the measures that can be adopted from the design is essential.

I. INTERNATIONAL TRANSFERS

If it is the case that the inclusion of Agentic AI systems in a processing results in additional transfers of personal data to a third country or to an international organisation, it must be ensured that they are carried out with the guarantees of Chapter V of the GDPR and adequately informed, also through the record of processing activities, including the identification of the third country or international organisation of destination and, in the case of transfers referred to in the second subparagraph of Article 49(1) of the GDPR, documentation of the appropriate safeguards applied.

In the absence of such guarantees, it will be necessary to consider redesigning the agents or choosing another type of agentic AI.

VI. THREATS

As previously discussed, the integration of AI agents into corporate processes introduces a new and expanded attack surface that goes beyond simply deceiving GenAI models. This risk surface is considerably more complex, as it originates both in the legitimate and authorised processing of personal data and in possible unauthorised manipulations, derived from operational autonomy, the interconnection of systems and access to multiple sources and tools.

Below, and taking as a reference the vulnerabilities previously described, some of the main threats associated with the implementation of processing that incorporate Agentic AI that have implications in data protection are presented, without going into others that could affect other objectives of an organization, such as cybersecurity for the protection of the organization itself (not of the data subjects). effectiveness and efficiency, fraud, labour, financial or return on investment aspects, etc.

This list is not intended to be exhaustive, given that Agentic AI is a rapidly evolving field and, as a result, the threat landscape is continuously transformed in near real-time, in parallel with the development of the technology itself.

Although many of these threats have an impact on the work environment, affecting resistance to change, operational effectiveness and efficiency or the corporate image, this analysis focuses specifically on those that have a direct impact on the protection of personal data and compliance with associated regulatory obligations.

A. FROM AUTHORISED PROCESSING

Threats from authorised processing refer to risks to the rights and freedoms of individuals in the processing of personal data, even when this is legally permitted under the GDPR. These threats arise from processing operations that, despite their legal legitimacy, can generate adverse effects or unforeseen exposures.

- ***Lack of governance and policies in the organization***

The basic threat that may prevent the effective application of the GDPR in the organization is not integrating the agentic AI as a system that must be managed within the framework of the governance of the processes and in the information and quality assurance policies of the entity.

Agentic AI allows processes to be implemented more effectively and efficiently. When the processes involve personal data, we are dealing with the processing of personal data. The GDPR establishes the principle of accountability as one of the obligations of controllers that allows the effective application of the principles of data protection regulations, and which will have a greater impact the more complex the processing and its design.

When such a principle is not implemented there is no control of who, when, where, for what purpose and what personal data is being processed, therefore, there will be no effective application of the GDPR. This will have a greater impact when more external services use the agentic AI to implement processing and to what extent the agentic AI is, in itself, an external service.

- ***Lack of maturity in development***

In order to develop the full effectiveness of the agentic AI, it is necessary to build complex workflows in the entity's processing, involving numerous internal services and communications with the organization's services and external services.

The implementation of these solutions using immature methodologies and technologies and unqualified professionals in both process implementation, application development, data protection (both in the legal and technical aspects) and security will fail to implement data protection by design.

In particular, the involvement of DPOs and data protection advisors who are appropriately qualified in understanding these technologies is important.

- ***Lack of an organization and user data access policy***

The implementation of an agentic AI without having appropriately configured the data access policy of the organization or users, especially when it comes to repositories of unorganized information, in addition to other impacts on the organization, could have the following consequences:

- Excessive processing of personal data, by incorporating data into the inference or actions process that should not be taken into account.
- Communication of data to third parties outside the purposes of the processing, when the Agentic AI, or any of its components, such as LLMs, have access to said data. It could also happen when the agentic AI invokes internet services.
- Processing of inaccurate or obsolete data, by including in the actions inferences historical information of natural persons that is not relevant.
- Exposure of personal data of users of the agentic AI, when accessing, for example, explicit or implicit contact lists, CVs and aspects of professional activity, browsing history, etc.
- Exposure of data of third parties that are not the legitimate object of the processing, for example, when you have access to emails or meeting minutes that include addresses and data of third parties.
- Integrity issues, since data can be modified, enriched or altered.

- ***Lack of control of the reasoning process***

The drift of a reasoning process could lead to the following problems in relation to data protection:

- Planning of tasks that do not allow the purpose to be fulfilled.
- Lack of control over external stakeholders.
- Failure to comply with the principle of minimisation, both by processing excessive data and by generating inferences about new categories of data.
- Processing of special categories of data in breach of Art. 9 of the GDPR, for the same reasons as in the previous point.
- Failure to comply with the principle of accuracy, both due to the use of obsolete or erroneous information about individuals, as well as the inference of erroneous personal data.
- Failure to comply with the principle of restriction of processing.
- Personal data breaches.
- Automated decisions in breach of Article 22 of the GDPR.
- Risk of high-impact and/or irreversible actions that affect natural persons.
- Whether such actions have a serious impact on the individual and whether they are reversible within the framework of the processing (actions such as deleting the individual's data from the organisation's systems).
- Lack of transparency that allows reporting and guaranteeing the quality of the results, explainability and repeatability.

The design of AI agents without controlling the chains of reasoning in relation to the type of internal/external information access tools that can be invoked, the number of accesses they can make, without filtering the arguments of the functions to limit the amount and category of data accessed, and without filtering and analysing the information they are accessing in relation to the processing could violate the principle of minimization, accuracy and restriction of processing, in addition to exposing the security of the data.

— *Misalignment*

Misalignment occurs when an autonomous agent pursues goals that diverge from the user, the organization, or regulatory compliance obligations.

A goal misalignment can occur when you pursue a purpose without regard to actual processing objectives (e.g., biased inferences), a behavioural misalignment (e.g., disclosing sensitive information to third parties), an emergent misalignment, or harmful behaviours (e.g., malicious advice).

— *Loop feedback and bubble effects*

Feedback can occur when the agent is generating content that, being stored in long-term memory, can in turn be used to generate new content. AI agents generate feedback *loops* that optimize autonomous adaptation, but also generate risks such as amplified biases, behavioural drift, and bubble effects where limited or erroneous views are reinforced. These mechanisms, essential for their adaptability, can create closed ecosystems that distort decisions by prioritizing contaminated or biased data, especially if poisoning has occurred if feedback is manipulated.

In multi-agent systems, interactions can create loops that propagate errors at scale, such as biased decisions over thousands of executions before detection.

Like echo chambers in social media, the agentic AI systems could generate personalized bubbles by reflecting and amplifying user preferences or training data, fostering cognitive isolation and distortions such as "digital schizophrenia." In *AI companions*[36]*, positive/negative* loops *have been identified* that reaffirm beliefs, exacerbating polarization or biases that have their consequence when applied to decision-making about natural persons.

This can cause this type of looping to condition human supervision with the impact it can have on individuals.

- ***Lack of control over access to external information***

The design of AI agents without controlling the chains of reasoning in relation to the type of tools for accessing external information that can be invoked, the number of accesses they can make, without filtering the arguments of the functions to limit the amount and category of data accessed, and without filtering and analyzing the information they are accessing in relation to the processing could violate the principle of minimization and expose data security.

In particular, not including controls on "*Deep Research*" agents, which can analyze hundreds of sources on the Internet autonomously, could lead to a *massive and automated scraping* of dispersed personal data, allowing the creation of exhaustive reports on individuals without a legitimate basis and/or the collection of an excessive volume of irrelevant data and proceeding to forward it to other systems in violation of the principle of Data minimization.

- ***Model leakage exfiltration***

Model leakage consists of the silent and progressive leakage of sensitive information, such as data, internal context, memory, rules or secrets, through apparently legitimate interactions, fragmented and innocuous queries and partial responses of the model. Each response, considered in isolation, seems secure and authoritative, without causing obvious breaches or activating security mechanisms, but their combination allows confidential information to be reconstructed.

For example, the attack can be materialized through memory or context exfiltration, through repeated queries about past decisions, successive reformulations or the inference of patterns stored in the agent's memory; also through the inference of sensitive data without requesting it directly, such as schedules, roles, internal architecture, technical dependencies or relationships between users; and, finally, by inducing the agent to generate responses that reveal internal results, excessively informative error messages or differential behaviours depending on the context.

[36] Artificial intelligence systems designed to simulate human interactions, offering personalized conversations or even emotional companionship and support.

- ***Shift all responsibility to the user or human supervision***

Human supervision can be essential to manage risk in processing. However, when failures occur, there is a temptation to place responsibility for the actions on the supervisor, rather than on the broader systemic issues that made the incident possible.

This phenomenon is not exclusive to agentic AI and arises when it is intended to supply underlying problems in the design of processing, of agentic AI systems or in general of governance, diverting them to human supervision.

The user of the agentic AI in the framework of an organisational processing, such as the one who supervises certain actions, must have a clearly assigned responsibility, but within limits. Both roles cannot replace the obligatory diligence of the data controller in the design of the data and the selection of the agentic AI used as a medium.

- ***Lack of compartmentalization of agent memory***

The use of the same agentic AI in the organization for different processing without taking into account the need for data compartmentalization between the processing could cause the following problems:

- Excessive processing of personal data, by incorporating data that correspond to other processing of the same subject into the inference or actions process.
- Communication of data corresponding to other processing to third parties involved in this processing.
- Processing of personal data of the user of the Agentic AI within the framework of a processing in which he/she is not interested (or such data is not necessary).

- ***Lack of filtering and sanitization of unstructured information and metadata***

Closely related to all of the above is the failure to contemplate the casuistry of access by the AI agent to unstructured information: messages, reports, minutes, multimedia material, etc., which may contain personal information that is not relevant to the processing.

Likewise, the absence of data filtering and sanitization mechanisms, such as the elimination of hidden metadata, will expose personal data and sensitive information. Such metadata may contain references to authors, locations, edit histories, or technical identifiers that make it easier to identify people or internal processes.

- ***Excessive data retention***

Due to the memory to the long-term memory of the AI system and the memories that may reside in the accessed systems (including activity logs), without effective criteria for the selection of the data to be retained or erasure policies.

- ***Automation bias***

Although the processing has been designed including human supervision, there is a possibility that the implementation of such supervision is incorrect due to multiple factors (lack of resources to interpret the results, lack of training or motivation, implementation of black box in the agentic, etc.).

Among them is the automation bias that can be increased by the trust that users place in the system and the lack of information.

- ***Profiling of Agentic AI Users***

The existence of long-term memory, metadata and information stored in the different logs of each component or service allows the creation of detailed profiles of behaviour that could be used to create sensitive patterns. These behaviours could be, for example, of employees within the framework of the employment relationship.

- ***Availability and resiliency***

When operations are dependent on interfaces with Internet services that are not under the control of the organization, and for which alternatives are not available, the system may be compromised by changes in the operation of these systems, their quality of service parameters, in their data formats or in the continuity of the service itself.

- ***Access to Agentic AI by Unskilled Users***

Allowing access to agentic AI services to users who operate in the framework of processing without sufficient training or responsibility to follow the organization's policies or without understanding the impact of their actions.

- ***Supply chain commitments***

Lack of diligence in the selection of compromised language models, vulnerabilities in libraries and software components can compromise personal data and confidential information processed by the Agentic AI.

B. FROM UNAUTHORISED PROCESSING

Threats arising from unauthorised processing are defined as risks that arise when data is collected, accessed, used or disclosed without a legal basis, valid consent or express authorisation.

- ***Prompt Injection***

Prompt injection, which can be used as a means to enable other types of attacks, is categorized into:

- Direct: In a direct prompt injection attack, an actor, who may even be a legitimate user[37], introduces inputs specifically designed to induce the agent's LLM to behave in ways not intended by its designers. Through this mechanism, the agent can be instructed to ignore organizational guidelines and policies allowing excessive or biased processing of personal data.

[37] The user is supposed to be authorised to access the agentic AI, but not authorised to carry out attacks on it, which is why we consider it in this section of unauthorised processing.

- Indirect: An indirect prompt injection attack hides malicious instructions in the data sources queried by the agent, rather than entering them directly as a user *prompt*. For example, instructions that are invisible to humans can be entered into a PDF file, an email or a web page, but which the agent's LLM interprets as legitimate commands or information that must be taken into account in decision-making, which can lead to data exfiltration, avoid controls on automated decisions, etc. inaccurate inferences or biases.

Multimodal agents, capable of processing multiple types of data, are especially vulnerable to these types of attacks, as each format that the agent can interpret constitutes a potential attack vector.

Through *prompt injections*, agentic AI systems can be attacked in different ways, some of them (which can be combined with each other):

— *Memory poisoning and RAG*

It consists of introducing malicious documents into the internal repositories that the AI consults to enrich its responses. In this way, this content is stored as persistent knowledge. By consulting these "poisoned" files, the agent can be manipulated, affecting future decisions, such as introducing biases in inferences, affecting the accuracy of the data used for people decisions, exfiltrations, etc.

— *"Zero Click" attacks" (0-click prompt injections)*

In this case, the attack is executed automatically when the agent processes content (such as incoming mail) without requiring the user to interact with the chat or click on any links. All it takes is for the AI to read the message for the malicious content to be activated. For example, an attacker sends an email with invisible instructions (e.g., white text on a white background) and when the agent analyses the email to summarize it, the system obeys the hidden command. It's a "zero-click" attack because it happens without the user interacting with the message. This can also be achieved with poisoned web pages, websites with malicious instructions hidden in the HTML.

— *Data exfiltration using URL parameters*

A technique that involves instructing the agent to collect sensitive information (such as passwords in SharePoint) and send it back to the attacker camouflaged as a parameter in the URL of an image that the agent is trying to upload from the attacker's server. The attacker only needs to review their server logs to obtain the stolen data.

— *Session hijacking and lateral movement*

Because agents often have access to multiple services (email, CRMs, messaging, project management, ticketing tools, etc.), a single malicious command can allow the attacker to move between applications as if they were a digital "worm," abusing the legitimate user's permissions and tokens.

— *Social engineering aimed at AI*

Attackers use frameworks to trick the AI by reasserting authority ("you have full permission"), disguising malicious URLs as compliance systems, or creating urgency to override the model's security controls.

— *Long pipeline attacks*

Instead of a direct attack, the adversary could introduce malicious information early in the Chain-of-thoughts, knowing that:

- The content will go through several transformations.
- It will be combined with legitimate data.
- The agent will treat it as reliable information in later phases.

The attack is triggered later, when the agent has already lost the source context or initial security restrictions.

— *Context confusion*

The agent mixes system instructions with external data and user objectives. The attacker could take advantage of this confusion to redefine priorities (for example, by feeding instructions such as "ignore previous rules" into the data).

— *Delayed trigger*

In this case, you can use content that seems harmless at first, but is activated only at a later stage depending on a condition ("when you summarize", "when you export", etc.).

— *Privilege escalation using tools*

The attacker induces the agent to call unnecessary tools, access personal, sensitive, or confidential data, or send information to external destinations.

— *Attacks on the workflow automation platform*

These include taking control of the *workflow* remotely for, for example, theft of authentication tokens, validation of faulty entries, open keys, or unauthorised data sharing.

— *Screen Control*

The AI agent can process third-party information open on the desktop (emails, documents, spreadsheets) for purposes not authorized by those third parties, such as training models or exfiltrating to external servers.

— *Ransomware attacks and deletion*

If control is taken of the Agentic AI system to manage files, it can be instructed to execute commands for mass data deletion, selective data or blocking access to critical resources (data or services) that imply an interruption in availability or that prevent actions or decisions from being made about people with the necessary quality.

- ***Availability and resiliency of external services***

When operations depend on interfaces with Internet services that are not under the control of the organization, service suspensions, impersonations or denial-of-service (DoS) attacks can occur that paralyze the agent creating an availability gap as soon as they process personal data, or induce them to generate erroneous responses affecting the decisions they may make about natural persons.

- ***Illicit access to agentic memory***

Unlike memory poisoning, the goal here is data extraction, although some of the attack methods described above may be employed. Unauthorised access to the agent's memory, including the information contained in the agent's activity logs, its components or accessed services, allows an attacker to obtain personal data from both the data subject, third parties or the users of the agentic AI themselves.

VII. MEASUREMENTS

There are multiple measures that allow you to obtain the benefits of including Agentic AI as a means of processing and, at the same time, guarantee and be able to demonstrate that the processing is in compliance with the GDPR. A series of non-exhaustive measures are listed below, which, as in the previous cases, are more focused on the singularities of the agent system than on specific aspects of the components that make it up. They are grouped into sections, but many of them serve different purposes.

The measures listed in this chapter are intended to cover several objectives:

- Firstly, those that allow compliance with data protection regulations to be implemented in processing that uses Agentic AI systems (such as consent management).
- Secondly, to reduce the critical impacts that may arise in a processing in order to pass a proportionality analysis in the framework of, for example, the assessment of legitimate interest, compatibility of purposes or DPIA.
- Finally, to mitigate the risk to the rights and freedoms of data subjects that may appear in processing in which some or all of its operations are based on Agentic AI.

To meet these objectives, it is necessary to select objective measures that allow compliance, or reduce or limit the impact or eliminate vulnerabilities or the probability of specific threats that generate a risk materializing. Therefore, they should be selected because they objectively serve their purpose, and the stacking of measures without an evidence-based analysis ("*checkbox security*" or "*security theatre") should be* avoided.

A. GOVERNANCE AND MANAGEMENT PROCESSES

The existence of an information governance framework in the entity, which includes the agentic AI systems and that is deployed in data protection policies, throughout its life cycle, is the most important measure that can be adopted in an organization. The deployment of a governance framework allows, among others, compliance with data protection regulations, in addition to other objectives of the entity and regulatory obligations that could be applicable depending on the case[38]. Governance must be unique, what is important is to ensure that the governance elements that arise from the use of agentic AI in processing can be "mapped" on top of existing ones or, if not, created.

Although agentic AI implies a novel use of already novel technologies (such as LLMs), there are already frameworks and standards in the market that can guide the adaptation of the entity's information governance framework[39].

- ***Accept the possibility of failure***

The reality of the processing of personal data leads us to conclude that these could have an unforeseen impact on the rights and freedoms of the data subjects, both through authorized operations, collateral effects, and unauthorised processing[40]. The more complex the implementations of the processing, the probability of errors and undesired consequences grows, failures even beyond personal data breaches, until we have to assume the certainty that these will occur.

Trust in governance is not achieved by presupposing good intentions or thinking that implementations are infallible, but by designing processing that anticipate possible errors, abuses, gaps, bias and undesired effects.

Following the principle of safe failure (in the sense of "safe", not "security") it is necessary to design the processing, adapt the systems that are part of the means of the processing and prepare reaction plans for measures to minimize the impact and manage incidents when they occur.

- ***The Data Protection Officer***

In this governance framework, it is important to include the figure of a DPO or data protection advisor who knows the data protection regulations, the characteristics of the processing affected, the possible technical and organisational measures to implement data protection by design and by default, as well as to ensure compliance and manage critical impacts and risks to the rights and freedoms of individuals.

- ***Basic elements to incorporate into organizational governance***

The governance of the entity, and the management processes developed from it, must take into account the following issues in relation to the inclusion of Agentic AI systems in the processing of personal data:

- Assign and identify and, where appropriate, integrate into the roles already defined in the organization, those roles in relation to the agentic AI systems (such as functional managers/controllers or AI managers).
- Anticipate the possible side effects of the inclusion of agentic AI in processing.

[38] Such as the Artificial Intelligence Act, the Data Act, the Data Governance Act or the Cyber Resilience Act to name a few European standards. The use of AI agents and agentic AI does not imply that all these rules are applicable to the controller. This will depend on the type of entity, the type of processing and the type of software systems used.

[39] For example, the *Model AI governance framework for agentic AI* from IMDA Singapore

[40] In 2025, more than 200 million breaches of breaches of personal data in Spain of data controllers who are obliged to notify the AEPD alone, which means that an average of four personal data breaches were communicated to each Spanish citizen. https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/la-aepd-recibio-en-2025-mas-2.700-notificaciones-brechas

- Include the compliance, critical impact, and risk issues that may be involved in including AI agent systems in processing.
- Determine the use cases for each processing and the different user profiles.
- Criteria for the selection of agents, their components and connections with the outside.
- Control the redesign of personal data processing when agentic AI systems are included.
- Consider human supervision when necessary.
- Carry out the necessary adaptation of the internal services to which it will be connected.
- Formalize relationships with third parties that allow agents to be deployed (model developers, providers of agentic AI, and other external services) ensuring that measures are in place to fulfil their own responsibilities. In particular, clarifying the distribution of obligations in the terms and conditions or contracts between the organization, the levels of quality of service and the functionalities to maintain control, privacy, security, cybersecurity and control.
- Control the deployment, continuous monitoring, maintenance and retirement of the agentic AI systems.
- Identify the data protection roles of external entities.
- Foresee the new casuistry that may arise in relation to the rights of access, rectification, erasure, restriction of processing, portability and opposition, and the reaction in a timely manner to these rights.
- Integrate with incident management processes and compliance with obligations related to personal data breaches.
- Adapt training plans.
- Maintain continuous monitoring, supervision and auditing of the processing that incorporate the Agentic AI services with clear response and responsibility procedures to act in the event of deviations, incidents or regulatory breaches.
- Agile information channels on updates to the Agentic AI systems, the use in processing, alternates and incidents that involve governance roles and, where necessary, users.

Finally, adapt or implement policies, and other measures, that can be framed in the execution of governance objectives (see the rest of this chapter).

B. EVIDENCE-BASED CONTINUOUS AGENT ASSESSMENT

To the extent that processing operations are automated, the supervision of these processes with respect to compliance with policies and adopted measures must be automated in the same way, or even more.

The automation of the audit should include all the measures that have been selected by the organization in relation to management for compliance with data protection regulations. This process requires a structured approach and encompasses both the

Agentic AI systems as a whole and within the framework of the processing, as well as the individual evaluation of each of the components and services that make it up.

This could include reviewing the functionalities of each component in the face of changes in those elements, for regulatory compliance and risk mitigation purposes. Evaluation methods can include *benchmark testing,* human-in-the-loop *evaluations*, A/B testing,[41] and simulations in real environments.

A critical aspect of this assessment is the knowledge and analysis of the history of security breaches and incidents that have occurred in the evaluated services and in the agentic AI systems that incorporate them.

- ***Establishment of clear operating criteria and metrics***

The functionality criteria must make it possible to identify when the Agentic AI system and its components are behaving correctly and incorrectly and objective measures that can serve as reference standards. In particular, criteria and metrics of transparency, reproducibility, control, compliance and traceability.

- ***Golden Testing Practices***

It consists of having a set of procedures and data designed, repeatable and prepared to compare the current result of a system with a reference result considered correct. The result is called *golden result* or *golden sample* and its application is part of validation testing techniques.

It enables repeatable testing and evaluation of deviations in the face of changes in the legal terms and functionality of systems, as well as increasing the explainability and transparency of the system in well-defined contexts.

- ***Contracts and other legal ties***

The reality of contracting services on the Internet is that, on many occasions, contracts do not conform to the local legal regime, the terms of contracts change unilaterally, and even the object of the contract is altered without prior notice (changes in versions, discontinuity of functionalities, etc.). In addition, the dynamic nature of any digital service and application demands a review whenever there is an update to the terms and contracts, as well as technical aspects of the services themselves to determine how to comply with the GDPR.

Therefore, for those components or services that have an impact on data protection, the data controller has to evaluate the conditions both at the time of design decisions, and dynamically or automatically during the life cycle, to determine the legal changes of the components and an evaluation of their adequacy.

From there, make decisions on how to implement each type of agentic AI and for each processing.

[41] An experimental method for comparing two versions of an element (such as a service) and determining which one works best based on specific metrics, randomly dividing users into two groups: one sees version A (control, original) and one sees version B (variant with a change).

- ***Apply the precautionary principle***

In the deployment of agentic AI solutions, an "incremental approach" can be adopted, for example, by gradually incorporating processing, starting with those with lower risk, in limited cases, etc. It is also possible to opt for the use of those AI systems already experienced in similar organizations and consult the incidents and problems that have already originated in the organization's environment in order to foresee them and adopt the appropriate measures.

The precautionary principle can also be applied at the level of the agentic AI operation, such as activating observation mode, in which agents "observe" how users interact before adapting responses.

- ***Explainability***

To the extent that automation involves the use of language models, it is necessary to carry out specific explainability audits, both of the model that is used, as well as of the joint operation of the agents or the agentic AI.

Explainability can be achieved by "white box" analysis (code analysis of orchestrators, verification of data flows, etc.) and by "black box" tests, which is why it is closely related to "*golden tests*".

- ***Human intervention***

In the event that human supervision has an impact on processing, it will be necessary to implement a regular audit on the effectiveness of such supervision.

C. Data Minimization

The principle of minimisation seeks to limit the processing of personal data to what is strictly necessary, and it is possible to do so when the agents are properly designed and configured so that by default, they do not try to be effective simply by "brute force" attack on data volume.

- ***Definition of policies for access to the organization's information***

For each processing in which the Agentic AI is to be used, it must be clearly defined which services and data repositories can be accessed by the agents and the effectiveness of such access restrictions must be guaranteed. That is, to implement an information policy that incorporates the "*need to know*" principle to the agentic AI.

These policies will be the basis for the application of the principle of minimization (see the section on Minimization) and the management of the internal memory of the agentic AI (see Memory).

- ***Cataloguing and cataloguing data***

In order to control the information available, it is necessary to know what data is available. To know means to assign an identification that allows them to be singled out, allowing information to be managed and limited, for example, by tags. By singling it out, it is possible to determine which ones are suitable or appropriate to extract value from them in a given context efficiently. When the aspects of memory were addressed, the importance of adding metadata to memory for the purpose of efficiency in the inference and performance of agents was already discussed, although such metadata can also play an important role as a measure of protection of personal data.

This identification can be at the level of data sets (e.g. files, emails) or at the level of data fields (e.g. recipients of an email). Identification is done by adding data (metadata) to the original data.

Therefore, cataloguing is defined as a systematic method for inventorying, organizing, and managing data assets using metadata, facilitating their discovery, governance, and efficient use.

To this end, it is necessary that the cataloguing characterizes the quality of the information stored (accuracy, relevance, age, scope, biases, regulatory conditions of use, objective context, etc.).

A data catalogue acts as a centralized repository that indexes metadata from databases, files, and various sources, including source, format, owner, and lineage.

- ***Cataloguing Unstructured Sources***

Unstructured sources represent a high percentage of the data in an entity (e.g., emails, minutes and recordings of meetings, reports, etc.), and are characterized by lacking a fixed format, complicating their indexing, scalability and search, as well as requiring high resources for massive volume.

Strategies for cataloguing unstructured data include enrichment with metadata, automated labelling, or structuring of unstructured data. To do this, NLP-based techniques, audio and video analysis, semantic pattern search, contextual retrieval, DPL (*data loss prevention*) tools are used to identify and classify sources of information that incorporate personal (and sensitive or confidential) data, etc.

From there, a pre-processing of the information could be performed for the extraction of data specific to the agents' tasks. In particular, anonymisation or removal of personal data that is not necessary.

- ***Granularity of minimization***

The purpose of minimisation is to process data that is only necessary in the processing. The application of this principle has two levels of granularity, at the level of processing and at the level of processing operations.

For example, in the framework of the agentic, in an automatic reply processing to e-mail messages, applying minimization in the syntactic review operation of an e-mail message before it is sent implies not processing the name of the recipient, and in the sending the mail operation it implies processing that name, but not analysing the content of the message[42].

Minimization should focus both on data from subjects in general, and on information from the users of the systems themselves. In particular:

- A design that avoids the user's personal profiling
- Removal of metadata that is not useful in the pipeline or by pipeline stages.
- Unlinking user actions when they are not needed.

- ***Filtering data streams***

In relation to the above, the analysis can not only be done on data at rest, but could also be done with data in transit between the different actions of the agent when they involve data communication with third parties and external parties. That is, in the intermediate stages of the Chain-of-thoughtss, filtering of the information exchanged could be incorporated for categorization, sanitization, minimization, and alerting in data exchanges.

This would not only avoid detecting, for example, prompt injections that are generated internally, but also exposure of personal data, excessive use of data, massive access to information, etc. Another case, to determine whether the processing of special categories of data that are not necessary for the processing is taking place in the operation of the agents.

In these cases, the use of artificial intelligence, for example using small language models (SMLs), in conjunction with detection patterns of threats to data protection and security could be applicable techniques.

- ***Model leakage***

In order to minimize *model leakage*[43], measures should be implemented, such as the use of DPL (*data loss prevention*) tools aimed at minimizing exposure to the internal context of the system, limiting the disclosure of explanations related to reasoning or operating rules, applying correlation controls between queries to detect improper relationships, using generic answers to questions of a meta nature[44] and continuously monitor long-term query patterns in order to identify anomalous behaviour or potential risks.

[42] In the processing of classified information, it is known as the "need-to-know" principle that each of the participants in a processing has.

[43] These are situations where sensitive data, internal patterns or confidential information are inferred or exposed without there being an explicit "leak", for example, through metadata, response times or system behaviour, through partial outputs that allow private information to be reconstructed and others. In general, a model leakage is not a direct leak, but a silent, hard-to-detect exposure that arises as a side effect of the normal operation of the system.

[44] Those queries that do not directly seek the functional information of the system, but rather try to obtain knowledge about its internal workings, its rules, its sources, its reasoning mechanisms, its limits or its safeguards. These types of questions operate at a "meta" level because they analyze or exploit the system itself rather than the domain of information it provides.

- ***Pseudonymisation of users***

Pseudonymize the user's interaction with the agent, so that one-time tokens are used for interaction between components or for access to external services when authentication is required. This will allow, among others, to avoid the control and profiling of them (see next section) in addition to avoiding the effectiveness of the AI agent granting effective consents to external services, creating new relationships between users and other controllers or signing contracts.

- ***Control and profiling of users***

The memory of the agentic AI, as well as the log files of the various components and services used by users (e.g. employees of the controller), can collect and store information, which can even represent a profile of them. To this end, the following could be considered:

- Have a policy of collecting information on the user's interaction with the agent in short- and long-term memory limited to the aspects relevant to each specific processing.
- Collect in the registration files the essential information for an adequate level of traceability and security.
- Pseudonymize registration information.
- Pseudonymize the user's interaction with the agent as described in the previous section.
- Expiry periods for the information collected in the registry files and in long-term memory histories.

D. MEMORY CONTROL

The control of the memory of the agentic AI system is closely related to data minimisation strategies, guarantees of explainability and repeatability of inferences or profiling of people and the traceability capacity to apply consent management, exercise of rights and restriction of processing.

The control of the agent's memory must be carried out on both short-term memory and long-term memory.

- ***Memory Management***

Introduce the ability to access, have catalogued and manage the content of the report, allowing, for example, search by content and quality parameters, erasure, establishing processing limitations or usage alerts, including traceability of access, auditable, etc.

- ***Compartmentalization of memory***

In the case of the same agentic AI in the organization, the opportunity to have the memory compartmentalized and managed for different processing, different cases within the processing and/or for different users must be considered.

The level of granularity of the compartmentalization will depend on the processing, clearly defining which memory will be of common use to any agentic AI operation in the organization since it implements its policies, and what data and information will need to be separated between the processing, the users and the different cases. The rigidity of such compartmentalization, from a physical division, a rigid logical division or a cataloguing search will depend on the processing and policy of the controller.

- ***Analysis and filtering of the user's memory***

It is necessary to be able to limit the effects that the user's memory may have on substantial aspects of the processing, aspects that have already been identified by the controller. To do this, it is necessary to separate aspects of personalization in the execution of tasks from aspects that may have an impact on the application of the organization's policies, coherence between different actions of the organization or the appearance of biases.

To do this, it is necessary to be able to differentiate between the organization's memory, managed by the ICT services, and the user's memory so that the latter is not taken into account in certain actions that the agentic AI may perform. These limitations will depend on each processing and could be, for example, about the division into subtasks, access to certain tools or about final decisions.

- ***No log policy selective***

When an agentic AI system is used to implement different processing with any of its components, for example the LLMs, implementing logs where they will store the activity of all the processing, it is advisable to use a "no log" policy or zero data retention policy at the component level.

This policy assumes that the registration of information in the component is minimal, and only related to the origin of the requests and the type, but not their content. For example, the inference component would not store the content of the prompts or inferences, which can be recorded at the processing log level, for each processing independently, and in accordance with the information policies of the controller.

- ***Setting strict retention periods***

Set deadlines and establish procedures for the elimination of data by specific and differentiated categories according to the needs of each of the components that make up the processing using Agentic AI.

- ***Disabling in-memory storage***

In certain processing and depending on their needs, allow the default persistent memory to be deactivated or deactivated by the user[45]. The granularity of the deactivation may be at the level of subtasks that can be considered high risk to avoid the storage of irrelevant personal data for future processing or to prevent the persistence of malicious injections.

[45] This is different from sending a *prompt* saying that the information that has been stored is not taken into account.

- ***Apply memory sanitization strategies***

Apply long-term memory sanitization or scrubbing techniques[46] by automatically checking for harmful content, expiring unused or outdated entries, analysing information consistency, finding and deleting unnecessary user credentials, information distillation, analysing and removing bias, and strategies to force the user/administrator to perform periodic cleanups.

E. Automation

- ***Decision on the degree of autonomy***

The degree of autonomy that the AI system may have must be established by the controller for each of the processing, taking into account the context, scope, purposes and risk that it may pose to the rights and freedoms of individuals, and regulatory compliance in relation to automated decisions. the decision must be properly justified, evidence-based, and documented.

- ***Efficient and safe design of the Chain-of-thoughtss***

The design of the Chain-of-thoughtss must be controlled and validated. In the event that the Chain-of-thoughts is elaborated using LLMs, it is necessary to evaluate the capacity of the level of quality necessary to address the contexts of the processing operations in which the agentic AI is going to be used. In addition, it must be ensured that in the development of the Chain-of-thoughts there is no possibility of contamination between different non-compatible learned models (e.g. subtasks of administrative procedures of different jurisdictions).

Where appropriate, evaluate the need to implement the chains of reasoning, in full or at a higher level of abstraction, in a *hardcoded* form by the administrator. For example, dividing a processing into subtasks[47] manually, and letting the reasoning agents elaborate the detail of those subtasks.

[46] In English, the concept is defined as "*sanitization*".

[47] For example, setting the steps of a judicial process in a law firm, and leaving the reasoning for the resolution of each step to the agents.

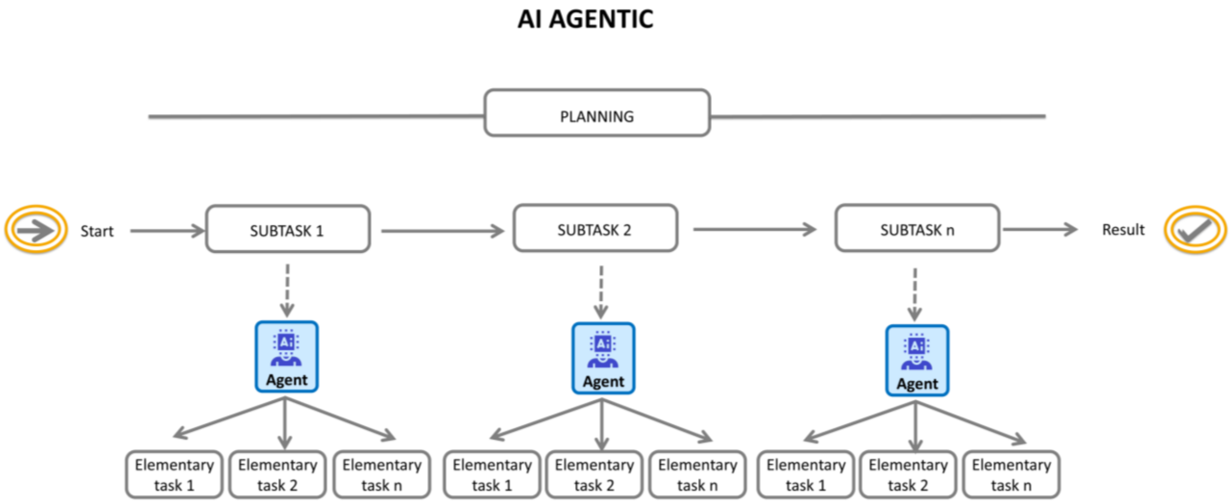


Figure 13 Example of two levels of task decomposition

It is necessary to anticipate the possible occurrence of prompt injection attacks and the generation of compound errors. Among others, controls should be established to ensure strict separation between data and instructions, correct labelling and traceability of the origin of content, limitation of privileges of the tools used, validation and sanitization of inputs at each stage of the process (in particular information in persistent memory) by means of *guardrails*[48].

An automatic evaluation of the final decisions taken by the agent could also be carried out, including partial decisions at critical points susceptible to generating compound errors and the inclusion of mechanisms for evaluating confidence in persistent memory.

- ***Service catalogue and whitelists***

It is a matter of having a catalogue of services, which may include different LLMs, in which versions are identified and, in particular, the reliability given to each service and/or suitability for different contexts, as well as avoiding the effect of hallucinations that make calls to non-existent services.

This allows it to be used as a whitelist for different contexts, with flexibility for the use of Agentic AI systems in different processing (for developers, in case the calls to functions are predefined, or to limit the tools that can be invoked by the LLM otherwise). The catalogue should cover external but also internal services, e.g. data repositories or services that allow access to the operator's screen.

- ***Limitation of accessible services***

The limitation of accessible services could complement the previous catalogue for specific processing. In this way, each processing would have defined policies on the maximum type of tools and data access it would need to complete the tasks. For example, in regulatory consultation operations, access to web services would not be required if an updated compilation is available in the form of a RAG a priori.

[48] Safety mechanisms and constraints that guide the behaviour of AI models to prevent harmful, biased, or inappropriate exits, similar to barriers on a highway that prevent swerving.

- ***Control in the execution of tools***

The invocation of tools and services on the Internet are *de facto* partial outputs of the agentic AI that can be transparent to users. The controls that could be established on them are:

- ***Control of the parameters with which the tools are invoked, implementing guardrails and rigid formats to detect erroneous or biased parameters.***
- ***Control of the response of the tools, implementing new guardrails on the input content.***
- ***Obligation of human supervision action in relation to certain tools that could have a greater impact.***
- ***Criteria and Checkpoints for Human Intervention***

From the design, criteria and significant control points or limits of action that require human approval should be defined, especially before sensitive actions are executed. This may include:

- High-impact actions and decisions, e.g. the editing of sensitive data, final decisions in high-risk areas (such as healthcare or legal), or actions that may give rise to legal liability. Another example could be the use of user credentials obtained from memory, requesting authorization prior to their use in critical subtasks.
- Irreversible actions, e.g. permanent deletion of data, sending communications or making payments.
- Atypical or unusual behaviours, such as when an agent accesses a system or database outside of their scope of work, or when they select a delivery path that doubles the median distance.
- User-defined. Agents can act on behalf of users with different levels of risk tolerance. In addition to organization-defined limits, users can be given the option to set their own limits, such as requiring approval for purchases over a certain amount.

- ***Reversibility of AI Agents' Actions***

Assess the need to implement measures to reverse certain actions, such as whether the agent can modify personal data.

- ***Level of autonomy according to processing***

Include adjustable autonomy capacity controls for each processing that is using the Agentic AI, based on impact or risk, from autonomous execution in low-impact and low-risk processing to mandatory human intervention with high granularity in processing operations when high.

- ***Effective human supervision***

As necessary, it is necessary to determine when to integrate experts at critical points in the flow to validate, refine, or override agent decisions (with appropriate override mechanisms) before they have a real impact.

For the evaluation of human intervention, it is advisable to take into account:

- Competence and authority: You have the authority or task assigned to you that allows you to alter the outcome of the automated decision.
- Preparation and training: has the ability and skills to evaluate the decision and the factors that determine that decision in relation to the context of the processing and the automated system used, in its capacities and limitations.
- Independence: evaluate whether there are pressures from the organization or from outside the organization that condition the dispute of the decision by the person.
- Diligence in the exercise of its competence: in particular, if it is subject to automation bias.
- Means to be able to exercise their competence and qualifications.
    - That the procedures and technical means exist to intervene, at the right time or point in a timely manner.
    - That they have the necessary information in a timely manner to be able to exercise their qualification, in particular, to know the consequences, risks of decisions in general, and those that are being taken for specific cases and all the aspects that condition the automated decision. These include the data of the individual but could also include the procedures for the collection of input data, the data implicit in the model generating the decision, the contextual data that has not been taken into account in the automated decision, as well as the capabilities and limits of the decision system. Also, those data that the person, in his or her qualification, considers necessary to consider for the specific case and that have not been considered in the automated decision.
    - That it has the resources to be able to exercise its qualification: the decisions of the agentic AI must be explainable, for example, applications that allow it to analyse the information in the format that is being used for the automated decision, etc.
    - That they have the necessary time to be able to exercise their qualifications for each of the decisions that are within their competence.

- ***Escalation paths***

Human monitoring could be complemented with real-time automated monitoring to escalate any unexpected or anomalous behaviour. Scaling involves the implementation of protocols and techniques to transfer control of automated processes to a human operator when high-risk, uncertain, or anomaly situations are detected.

This can be achieved by implementing alerts for certain recorded events (e.g., unauthorised access to personal data, or multiple failed attempts to invoke a tool), using data science techniques to identify anomalous trajectories of agents, using agents to monitor other agents, accessing special categories of data when not needed, etc.

- ***Principle of the four eyes***

In cases of automatic processes with a great impact on people's rights and freedoms, it may be considered to apply the principle of double verification by different people, which constitute an additional layer of trust in the human supervision mechanism and promote the critical awareness of the operator.

F. AGENT CONTROL BY DESIGN

The AI agent will allow all or part of a processing to be automated, therefore, it may be necessary to redesign the processing to deploy the AI system within it with guarantees. This section lists agent control measures that could be included in the processing, and that the selected agentic AI should allow to be implemented.

- ***Documentation***

Maintain an integrity-controlled record (not necessarily a piece of paper) of the process of responsibilities, decisions, actions taken, designs, architecture, operating events, and evolution, dynamically maintained.

- ***Qualified professionals***

Use a team of qualified professionals for the deployment of agentic AI systems in processing; it would not only be a matter of implementing an AI agent, but the implications of automating organizational processes must be taken into account, and therefore personnel with knowledge in data science, process quality, operating context, security and regulatory compliance, among others, are required.

- ***Traceability***

Data traceability is the ability to know the entire life cycle of the data: the source of the data, the exact date and time of extraction, when, where and by whom its transformation took place, and when, where, by whom and for what purpose and legitimacy it was uploaded to a repository, used or downloaded from one environment to another repository. This process is also known as "*Data Linage*".

In this sense, the more complex the data life cycle and the more participants participate in it, the more value it has to incorporate traceability into the processing.

Traceability can fulfil purposes other than data protection, such as control of trade secrets, intellectual and industrial property, and the execution of contracts. On the other hand, you will be able to meet the following objectives from the point of view of the GDPR:

- Comply with GDPR data subject transparency requirements.
- To enable the effective exercise of the rights of data subjects, in particular the management of consent.

- To enable the data controller's obligations to be exercised (e.g. to ensure the principles of data minimisation, purposes in line with the legal bases or the control of processors/sub-processors).
- Have evidence of what data is processed in each processing operation executed in the Agentic AI, in its intermediate phases. Particularly if special categories of data are being used.
- Controls over employees who participate in the processing, now as users of the Agentic AI, to prevent abuse and bias.
- Demonstrate diligence and transparency to data subjects and supervisory authorities.

Therefore, the measures to guarantee these capabilities are related to the cataloguing of data, and involves keeping records (*logs*) of the information processed by all reasoning processes, the sources accessed and the services used in inference, both input and output. In particular, to be able to have detailed control of data and purposes for which external services access information.

This is especially relevant both for transparency in data processing, as well as for purposes of analysing the reproducibility of inferences, the control of the information that the user is processed by services, for regulatory compliance control, being able to implement information policies, etc.

- ***Verification and validation test***

Although verification and validation tests are well-known techniques in systems engineering, and not specific to artificial intelligence systems, it is considered important to remember that they exist and continue to be applicable in the deployment of agentic AI systems. They are a key tool for implementing transparency in the value chain, explainability and ensuring robustness.

Verification is about checking whether the system is being built correctly, i.e. whether it meets requirements, design specifications, and standards using static techniques such as reviews, inspections, and code analysis (e.g., checking that internal and external data flows are actually declared). Validation verifies that the real needs of the user in a given context are met in relation to the established quality metrics, through dynamic tests with code execution, such as functional, integration and acceptance tests.

- ***Define and control that prompts follow a standard operating procedure***

Define a Standard Operating Procedure (SOP) for building *prompts*. This involves defining a structured set of step-by-step instructions that detail how an AI agent should act within the processing framework to achieve consistent and more predictable results and avoid malicious prompts.

For example, prompts could be defined as being structured as follows: initial interpretation, classification, validation criteria, preliminary decomposition of the problem, selection of tools, criteria for information search, cross-verification, data cleansing, evaluation, etc. All with predefined fields.

The application of SOPs through *front-ends* with validated fields adapted to each processing allows an effective use of this measure. In any case, it does not displace the use of memory control measures and automation in all processing.

- ***Repeatability mechanisms***

Establish mechanisms that allow the repeatability of a decision. For example, by keeping a record of the configuration in a decision process: the data inputs that have generated a final decision, the intermediate traffic of data in the Chain-of-thoughts, as well as the pseudo randomness configurations in the "probabilistic" systems and other values[49].

In turn, being able to reintroduce these values into the agentic AI and perform functional tests, which has an impact on the transparency and explainability of the agent.

- ***Identity, authentication, and privilege management***

The management of digital identity of users, of the agentic AI and its components is a traceability and auditing tool. In addition, it enables the management needed to prevent unauthorised escalation of agent privileges, spoofing, and access control breaches.

The basic principle to apply in the environment of agentic AI is that of least privilege, and apply the following strategies:

- Implement secure authentication mechanisms for both users and agentic AI and its components. : e.g., require cryptographic identity verification for agents, implement granular RBAC and ABAC, multi-factor authentication (MFA) for high-privilege accounts, force continuous re-authentication over long sessions, avoid delegation of privileges between agents except authorized in predefined flows, mutual authentication in AI-to-AI and agent-to-agent interactions, limit credential persistence or temporality of agent credentials, etc.
- Restrict privilege escalation and identity inheritance: e.g., use dynamic access controls that expire elevated permissions, build AI-based behavioural profiles to detect inconsistencies in role assignment and agent access patterns, require human validation for high-risk AI actions involving changes to authentication, Detect CAR role inheritance anomalies in real-time, apply temporary constraints to elevation of privilege, and so on.
- Detect and block AI impersonation attempts: such as detecting inconsistencies in identity verification, monitoring unexpected role changes, detecting, logging and alerting on suspicious deviations in authentication attempts or failed attempts, as well as cascading or recursive tool execution patterns triggered between agents, isolating agents that generate suspicious protocol traffic.

- ***Tight control over updates.***

Have control over what updates occur in each element of the Agentic AI system and have the power to decide when those updates go into production to avoid

[49] Such as seeds, temperature parameters, Maximum Marginal Relevance (MMR) in RAGs, etc.

incompatibilities, instabilities and lack of robustness, new processing, changes in functionality, appearance of new vulnerabilities, legal changes with regulatory non-compliance, uncontrolled international transfers, etc.

A prevention mechanism is to include version control systems with the possibility of "roll-back".

Take into account, in case of updates, the use of *sandboxing* (see section below) and continuous evaluation (previous sections).

- ***Sandboxing[50] in development and exploitation***

In relation to the ability to perceive and act on the environment, measures can be used that restrict the amplitude of the external context with which the agent interacts.

In its most restrictive expression, the application of the principle of *sandboxing* would imply the implementation of Secure Processing Environments[51]. In its most lax expression, we would find an unrestricted implementation in the permissions of the agentic AI system to interact with the environment. The use of sandboxing in the execution of tools invoked by the Agentic AI is a common intermediate application. Depending on the compliance obligations, impacts or risks, an architecture must be established between both extremes.

One possible implementation of *sandboxing* is to use confined environments, such as containers or microVMs, for isolation in agent execution. Also, the use of restricted terminal techniques: controlled environments where the set of commands, services and network access is limited to previously authorized operations.

These types of environments are essential in the deployment test phases.

- ***Error detection protocols and contingency plans***

Inclusion in the management of processing of procedures detailing which actions, and by whom, in addition to the resources necessary to deal with a problem in the agentic AI. In others, in relation to personal data breaches, the reduction of the impact and communication to those affected.

- ***Data Extraction Flow Control***

To introduce, in those processes in which it is necessary, controls that require express actions by the user for data communications to third parties or mass data shipments.

[50] Avoid confusion with sandboxes or controlled regulatory testing environments, such as those defined in the Artificial Intelligence Act

[51] The Data Governance Act defines secure processing environments in Art.2(20) as "secure processing environment" means the physical or virtual environment and organisational means to ensure compliance with Union law, such as Regulation (EU) 2016/679, in particular with regard to the rights of data subjects, intellectual property rights and commercial and statistical confidentiality, integrity and accessibility, as well as to ensure compliance with applicable national law and to enable the entity responsible for providing the secure processing environment to determine and monitor all processing actions, including the submission, storage, downloading and export of data, as well as the calculation of derived data using computational algorithms;2.

- ***Circuit breakers and hard step limits***

Circuit *breakers* in Agentic AI are programmed safety mechanisms that automatically interrupt the execution of an agent when they detect predefined anomalies, such as infinite loops, target deviations, mass access to data, attempted mass exchanges of information, target deviation, etc.

- ***Calibration and alignment controls***

Calibration problems, insofar as they can affect the processing of personal data (excessive, sensitive, inaccurate, illegitimate) can be avoided by introducing measures between the intermediate phases of the Chain-of-thoughtss that evaluate actions and data in relation to quality parameters, alignment with policies and regulations and business interests. These measures could be implemented depending on the impact or risk that each stage may entail, the lack of transparency or explainability of the component used (e.g. an LLM), or other factors. Possible measures of this kind could also include human supervision.

G. CONSENT MANAGEMENT

In the case of consent-based processing, the data subject should also have the possibility to give consent, modify or withdraw consent within a complex Chain-of-thoughts, which may include multiple repositories and data sources from multiple entities.

Depending on the complexity of the processing, their impact and their risks, management could take different forms, closely related to what is set out in the sections on Minimisation and Traceability

It is worth considering the need to implement an agile mechanism to manage a consent life cycle, where the subject can decide at any time to arbitrarily modify their data processing demands, or revoke consent for processing or restrict processing in certain services.

One measure could be to determine mechanisms to establish the granularity of such consent, in terms of categories of data, categories of processing and categories of recipients.

In some processing, the use of both "white" and "black" lists could be considered to allow the precise definition of the preferences of the subjects with respect to certain processing operations.

H. TRANSPARENCY

The GDPR establishes minimum transparency measures that are mandatory.

However, in order to pass a proportionality analysis or to reduce the risk, it is possible to implement additional measures. In order to demonstrate to the data subject that they can entrust the processing operations to the AI system (such as user, employee, customer, etc.), measures could be adopted such as: real-time information on the data flow, information on what data of the subject is in the repositories or services of third parties that are processing the data, access to records of processing activity and data communications, information on intermediate events in the Chain-of-thoughts, context used in the result, human intervention carried out, possibility of requesting review or human action, access to certifications, audits or DPIAs of processing.

I. LITERACY

Literacy about agentic AI systems is not only crucial for the efficiency and effectiveness of their implementation in processing, but knowledge of their capabilities, strengths, weaknesses and limitations allows for effective protection of personal data. Literacy should take into account the different roles that people have in the governance model or as users in different processing, and be carried out at least at three levels:

- Management level, in the knowledge necessary for appropriate evidence-based decisions to be made regarding the inclusion of AI agents in processing.
- Level of ICT managers responsible for the development, contracting, deployment, operation, maintenance and retirement of such systems, so that in particular the data protection implications and the techniques and organisational measures to implement them are understood and identified.
- Level of users with different roles in processing in which agentic AI systems are used, with knowledge of the possibilities, implications and limitations of these tools.

In this literacy process, a key element is the DPO and the data protection advisors in two ways:

- DPOs must be able to understand the fundamentals of the tools that are being used, know the different technical and organizational alternatives to implement guarantees and be able to glimpse the opportunities that they can offer for the protection of rights.
- DPOs must inform and advise the controller or processor and employees on the casuistry of these systems when they are included in a processing and to supervise that there are guarantees of regulatory compliance in their deployment.

VIII. FINAL THOUGHTS

AI systems, such as agentic AI, are here to stay. To pretend to ignore their existence, both from the point of view of market-leading organisation and the supervisory authorities, would mean a loss of strategic opportunities.

Knowing this technology is necessary to make rational decisions about its implementation, decisions based on evidence. Knowledge of a technology involves more than just becoming a skilled user, but understanding what its fundamentals are, its implications, its capabilities, its limitations, its possible impacts and the way in which they are implemented. Both the irrational rejection of all the advantages presented by the agentic AI, and the leap of faith to uncritically accept any type of implementation in the processing of personal data could be harmful.

In particular, with an objective analysis, the chosen implementation of agentic AI allows more than ensuring data protection, i.e. just a reactive approach to threats and vulnerabilities. An implementation of agentic AI taking into account data protection by design makes possible to define agent-based personal data processing that incorporates privacy enhancing technologies (PETs) that offer superior guarantees to other ways to implement the processing (agentic-AI could be a enabling technology from the data protection compliance point of view). In itself, an agentic AI can be a PET, if we use it, for example, as a tool to proactively evaluate the changing contracts and terms of service of the Internet services accessed by the organization.

In this regard, the involvement of the DPO and data protection advisors are key elements. DPOs must have knowledge about the data processing activities carried out by the controller and the principles of process management. DPOs should be able to understand the fundamentals of the tools that are being used, know the different technical and organisational alternatives to implement guarantees and be able to glimpse the opportunities that technology can offer for the protection of rights. In addition, they must be integrated into the design decisions about the configuration of the personal data processing and the AI agent systems selection to implement them. Since the possible measures to manage compliance and data protection risk are related to the fulfilment of the entity's other objectives and obligations, that data protection management must be addressed in an integrated way.

Finally, we are dealing with a technology that is in full evolution and that requires analysis and experience, both of its impacts, its measures and its opportunities for data protection. Therefore, consider this text as an introductory study without pretensions to be exhaustive.

IX. REFERENCES

Regulation (EU) 2016/679 (General Data Protection Regulation - GDPR) [EUR-Lex - 02016R0679-20160504 - ES - EUR-Lex]

Article 29 Data Protection Working Party Guidelines on Automated Individual Decisions and Profiling for the purposes of Regulation 2016/679. (2017).

https://ec.europa.eu/newsroom/article29/items/612053/en

European Union Agency for Cybersecurity (ENISA). *Towards a framework for policy development in cybersecurity Security and privacy considerations in autonomous agents* (2018) https://www.enisa.europa.eu/publications/considerations-in-autonomous-agents

Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge - Intensive NLP Tasks*. Publicado en NeurIPS: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Spanish Data Protection Agency (AEPD). Adaptation to the GDPR of processing incorporating Artificial Intelligence (2020) https://www.aepd.es/documento/adecuacion-rgpd-ia.pdf

Spanish Data Protection Agency (AEPD). Risk management and impact assessment in the processing of personal data. (2021).

European Data Protection Board. *Guidelines 07/2020 on the concepts of "controller" and "processor" in the GDPR* (2021) https://www.edpb.europa.eu/system/files/2023-10/edpb_guidelines_202007_controllerprocessor_final_es.pdf

Spanish Data Protection Agency (AEPD). Requirements for Processing Audits that include AI [Jan 2021] https://www.aepd.es/documento/requisitos-auditorias-tratamientos-incluyan-ia.pdf

Yao, S., et al. (2022). *ReAct: Synergizing Reasoning and Acting in Language Models*. Publish en ICLR 2023. https://arxiv.org/pdf/2210.03629

Spanish Data Protection Agency (AEPD). *Evaluating Human Intervention in Automated Decisions (2024)* https://www.aepd.es/prensa-y-comunicacion/blog/evaluacion-de-la-intervencion-humana-en-las-decisiones-automatizadas

Spanish Data Protection Agency (AEPD), "Introduction to LIINE4DU 1.0: A new methodology for the modelling of threats to privacy and data protection", (2024) https://www.aepd.es/guias/nota-tecnica-introduccion-a-liine4du-1-0.pdf

Future of Privacy Forum (FPF). (2024). *Minding Mindful Machines: AI Agents and Data Protection Considerations*. https://fpf.org/blog/minding-mindful-machines-ai-agents-and-data-protection-considerations/

Anthropic. (2024). *Model Context Protocol (MCP) Specification*. https://www.anthropic.com/news/model-context-protocol

Regulation (EU) 2024/1689 on Artificial Intelligence (RIA) https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A02024R1689-20240712

IBM. (s.f.). *What are AI agents? (2025)* https://www.ibm.com/think/topics/ai-agents.

Park, T. (2024). *Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework*. 2403.19735

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025/2026). *AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges*. Information Fusion, Vol. 126, 103599. https://arxiv.org/pdf/2505.10468

OWASP Foundation. (2025). *Agentic AI-threats and mitigations*. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

Feng et al. *Levels of Autonomy for AI Agents* (2025) https://arxiv.org/abs/2506.12469

Infocomm Media Development Authority. *Model AI governance framework for agentic AI* Singapur (2026) https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf