aepd
agencia
española
protección
datos

AGENT-BASED ARTIFICIAL INTELLIGENCE FROM A DATA PROTECTION PERSPECTIVE

February 2026

EXECUTIVE SUMMARY

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

The purpose of this document is not to analyze the compliance of a specific processing operation that uses AI agents, but rather to address how to manage the peculiarities that arise in a processing operation when it is implemented wholly or partially with agents.

Understanding this technology is key to making informed, evidence-based decisions about its implementation in personal data processing. User knowledge is not enough: it is necessary to understand its fundamentals, scope, limitations, and how it is applied. Both the irrational rejection of agentic AI and its uncritical acceptance in the processing of personal data can be harmful. In particular, we must proactively take advantage of the opportunities offered by this technology for greater data protection by design and as a PET tool in itself.

The text begins with a brief description of what agentic AI systems are. It then analyzes the potential vulnerabilities of these systems that affect data protection compliance, aspects of data protection regulation compliance, and specific threats that can exploit the various vulnerabilities. Finally, the document lists measures that a controller or processor could take to ensure compliance with data protection regulations and reduce or eliminate the impacts that agentic AI has on the rights and freedoms of data subjects when deployed in processing operations. These analyses will focus on what is most distinctive about agentic AI as a system in the processing of personal data, beyond the vulnerabilities, threats, and measures that are well known in generative artificial intelligence, 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.

TABLE OF CONTENTS

I. INTRODUCTION

Task automation is the use of technologies to perform repetitive activities without constant human intervention. This approach efficiently transforms processes that were previously performed entirely manually, freeing up time for higher-value tasks. Automation systems are used in everything from industrial to office environments, including any other productive or service sector.

The development of large language models ($^{\text{LLMs}}$) completely changes the automation paradigm, giving rise to the concept of agentic AI as AI-based systems with the ability to act autonomously to achieve objectives: AI agents. The integration of language models represents a qualitative leap in the efficiency and complexity of the tasks they can perform, opening up a universe of possibilities for improving business processes and public administration. In turn, the use of systems that implement the agentic AI paradigm (AI agents) working collaboratively to automate multiple processes leads to a change in the very conception of the implementation of processes, workflows, or *workflows* of entities, as well as the use of generative artificial intelligence (hereinafter GAI) in the workplace.

The ability of agentic AI systems to operate autonomously, enrich themselves with information from the digital environment, and perform complex tasks introduces new challenges in many areas, including the workplace, organizational management and control, resilience, *safety* and cybersecurity, ethical issues, the possibility of fraud, corporate image, etc., as well as those related to personal data protection. Also, as artificial intelligence systems themselves and due to their data processing, obligations may arise from general regulations, such as the Artificial Intelligence Regulation[2] or the Data Regulation, or from specific regulations depending on the area of application.

This document provides an introduction to data protection issues that may arise when data controllers and processors decide to use agentic AI systems to process personal data.

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

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

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

[3] Even if data from agent services is used for training artificial intelligence.

of AI agents in an organization where there is no processing of personal data[4] .

AI agents are means, systems, that enable 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 operations. On the other hand, an AI agent can be part of a processing operation 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 analyze compliance with a specific processing operation that uses AI agents, but rather how to manage the peculiarities that are incorporated into a processing operation due to the fact that it is implemented wholly or partially with agents. Different processing operations and different types of agents implemented in such processing operations could have different implications for data protection. The analysis carried out in this document will study these implications in a generic way, taking into account that they are not inherent to, nor necessarily part of the nature of, all agents or all uses of agentic AI.

The text begins with a brief description of what agentic AI systems are. It then analyzes the potential vulnerabilities of these systems that affect data protection compliance, aspects of data protection regulation compliance, and specific threats that can exploit the various vulnerabilities. Finally, the document lists measures that a controller or processor could take to ensure compliance with data protection regulations and reduce or eliminate the impacts that agentic AI has on the rights and freedoms of data subjects when deployed in processing operations. 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 its component elements, such as LLMs, databases, communications, etc.

All of this will be carried out within the limitations 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 emergence of AI, but their functionalities were limited compared to what can be achieved with agentic AI.

Agentic AI involves much more than just using LLMs. Understanding how it works is essential to creating a climate of trust, through evidence, that they have been

[4] For example, Park, T. (2024). Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework describes an LLM-based multi-agent AI framework for detecting and interpreting anomalies in financial market data, with particular application to S&P 500 index data. 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.

[5] For example, aspects such as LLM training, which requires a separate analysis, are not assessed.

put in place the appropriate measures and safeguards to enable data controllers and processors to get the most out of this technological option.

The text will describe what an agent is and the concept of agentic AI, bearing in mind that classifications are always formal in nature and that the technological reality is a gray scale that, in this case, crosses concepts such as LLM, IAG, and new developments that may arise in the future.

A. AI AGENT

An AI agent is an artificial intelligence system that uses language models to fulfill an objective. An AI agent acts appropriately according to its circumstances and objectives, is flexible in the face of changing environments and goals, learns from experience, and makes appropriate decisions given its perceptual and computational limitations. To do this, it breaks down complex tasks into subtasks, which are executed in a planned manner, creating a chain of reasoning, each of which is implemented with different tools and perceives the environment through access to internal and external services.

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

- Autonomy: ability to operate without constant human intervention.
- Perception of the environment: they process inputs in real time using sensors, application interfaces (APIs), cameras, etc., to interpret dynamic contexts. Interaction with the environment avoids the problem of "static knowledge cut-off"[9] in 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 rather than just reacting, and can initiate actions on their own.
- Planning and reasoning: they allow for the planning of sequences of actions to meet specific goals, evaluating alternatives and prioritizing optimal results.

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

[7] ISO/IEC DIS 22989 3.1.1 Agent: an automated entity that perceives 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 objectives through the use of AI techniques.

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

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

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

- Memory and Adaptability[11] : These enable the context to be defined, experiences to be accumulated, behaviors to be adjusted to user reactions, and iterative improvement through feedback or self-assessment with short- and long-term memory.

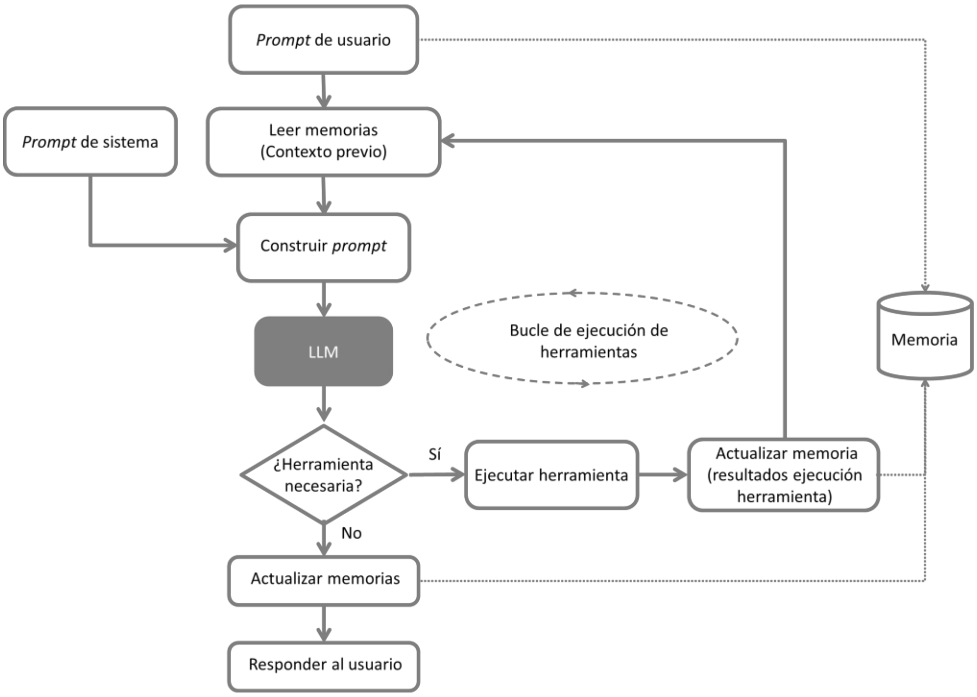
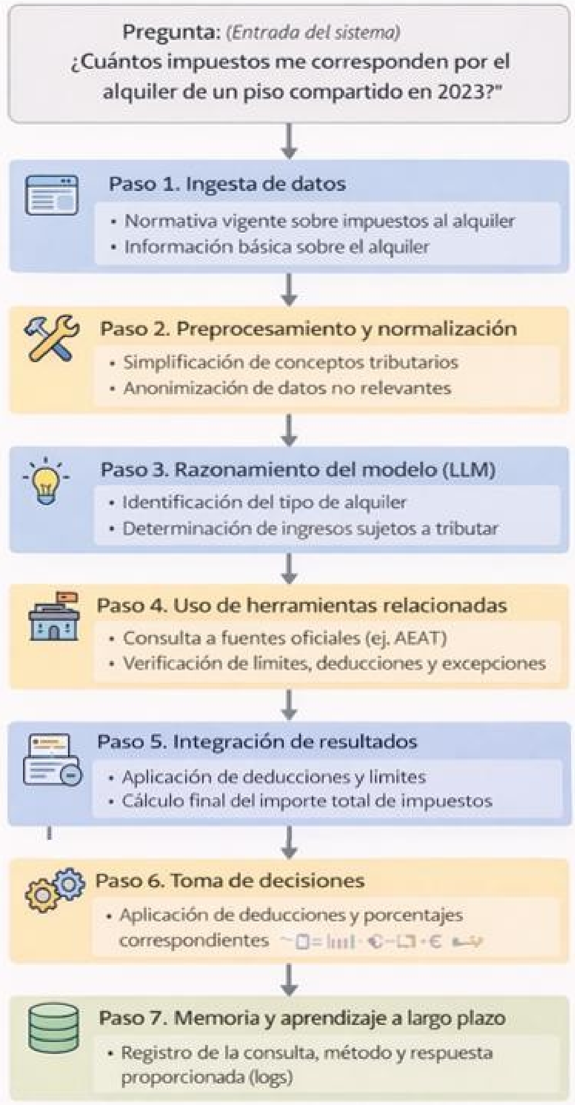


Figure 1 Example of a basic AI agent implementation

B. THE REASONING CHAIN

The reasoning chain, *pipeline*, or processing is the internal process by which the agent breaks down a problem into successive, chained logical steps until it reaches a final decision or response. 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.

[11] The literature refers to "learning," which can lead to confusion in the sense that the LLMs that form part of the agent are being retrained. The agent's "learning" is not achieved by retraining the LLM. Although the information may be used to improve the LLM, this is not a feature of the agent, and in many cases will not be done.



Pregunta: *(Entrada del sistema)*
¿Cuántos impuestos me corresponden por el alquiler de un piso compartido en 2023?"

Paso 1. Ingesta de datos
- Normativa vigente sobre impuestos al alquiler
- Información básica sobre el alquiler

Paso 2. Preprocesamiento y normalización
- Simplificación de conceptos tributarios
- Anonimización de datos no relevantes

Paso 3. Razonamiento del modelo (LLM)
- Identificación del tipo de alquiler
- Determinación de ingresos sujetos a tributar

Paso 4. Uso de herramientas relacionadas
- Consulta a fuentes oficiales (ej. AEAT)
- Verificación de límites, deducciones y excepciones

Paso 5. Integración de resultados
- Aplicación de deducciones y límites
- Cálculo final del importe total de impuestos

Paso 6. Toma de decisiones
- Aplicación de deducciones y porcentajes correspondientes

Paso 7. Memoria y aprendizaje a largo plazo
- Registro de la consulta, método y respuesta proporcionada (logs)

Figure 2 Example of a reasoning chain

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

In the latter case, LLMs appear as one of the core components of AI agents. Different types of LLMs and AIGs can appear in an AI agent for different purposes: knowledge capabilities, content generation (such as translators, transcribers, etc.),and reasoning. What is characteristic of AI agents is the use of LLMs as reasoning machines that will direct complex autonomous actions, analyzing user requests, responding sequentially to inputs, processing the results of different services, and/or constructing a final response. Regardless of whether LLMs such as IAG are used in agentic AI systems

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

content or information repository models, the distinctive feature is to use them for task decomposition.

Knowing the chain of reasoning 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 it is transformed, 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. PATTERNS OF AI AGENTS

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, giving them a degree of operational independence. These patterns may have different configurations.

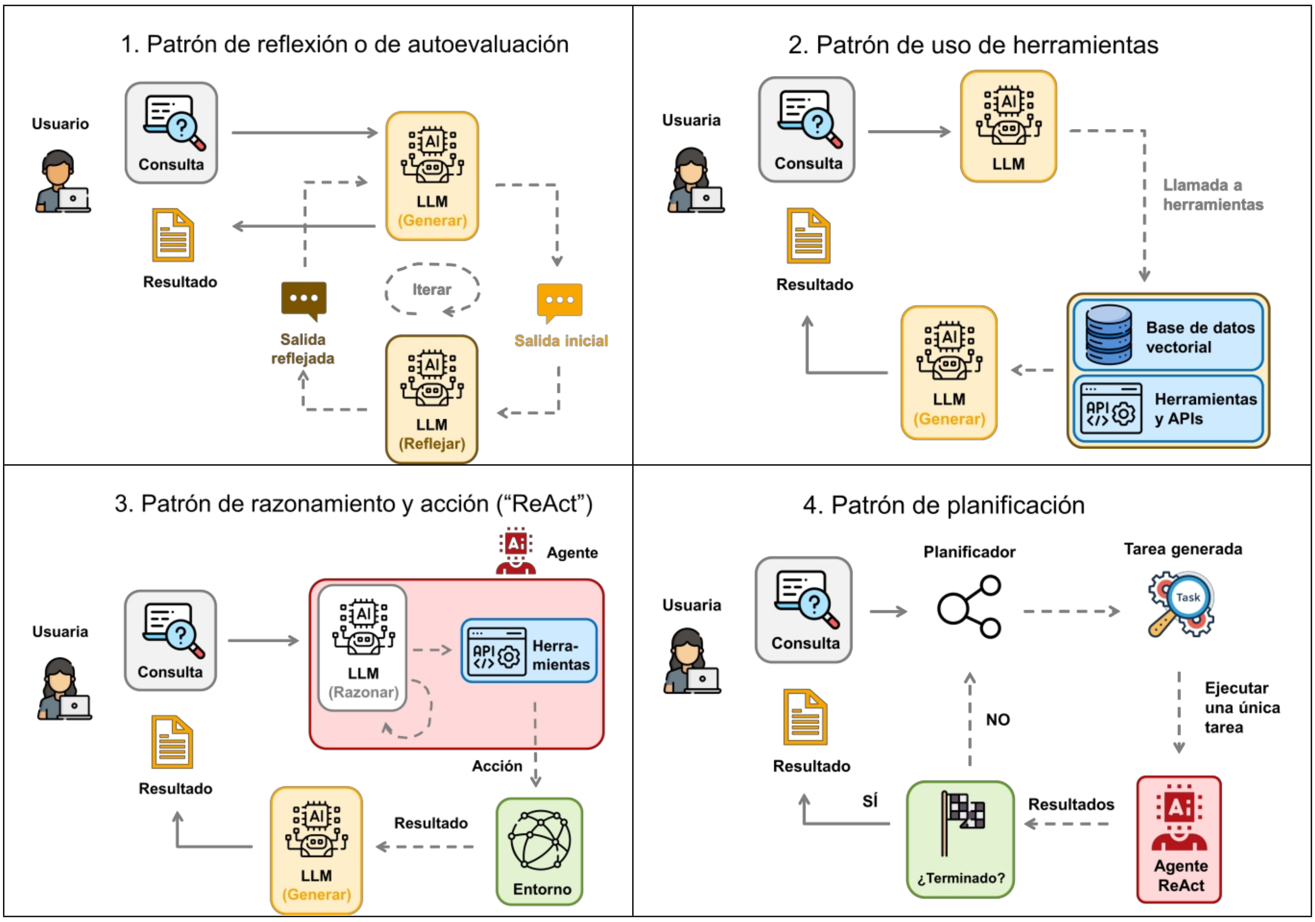


Figure 3 Simplified representation of some types of patterns

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

In this way, AI agents can automate repetitive data processing tasks, analyze 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 calls to/from background tools to obtain up-to-date information, initiate operations, optimize workflows, and autonomously create subtasks for the purpose of achieving complex objectives.

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 them as needed. These services can be either remote (e.g., web pages or services) or 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 humans. The behavior of an autonomous agent is essentially determined by three factors: the team of developers who design and adjust the agent's AI system; the team responsible for its deployment and configuration; and, finally, the user themselves, who defines the specific objectives that the agent must fulfill and the tools it can access to do so.

D. MULTI-AGENT

Multiagent architecture combines several agents, where the behavior and responsibilities of each are strictly defined, they share information and decisions, and they are able to collaborate, compete, or negotiate with each other to achieve more elaborate objectives.

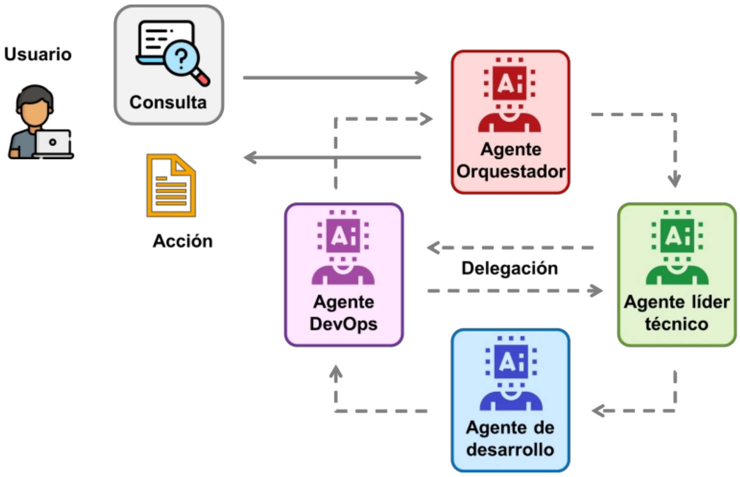


Figure 4 shows a simplified example of multiagent architecture.

There are various approaches to multi-agent AI: centralized models, sequential agent execution, distributed or hierarchical models. Each agent may have a different range of action (tasks it can perform and tools it can invoke) 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, an orchestration layer will always be needed to coordinate the agents' lifecycle, manage dependencies, assign roles to each agent, establish domain limitations, and resolve conflicts.

E. DETAILS OF THE ARCHITECTURE OF A MULTI-AGENT

The general architecture of an agentic AI system is a factor that must be understood in order to analyze its possibilities, limitations, and vulnerabilities. The components of an agentic AI system could be:

- An application that manages the interface to perform tasks for the user 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 service instructions.
- Services, including built-in functions, 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 long-term persistent 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 that form part of the agent's infrastructure, such as credential management, access control, action traceability, etc.

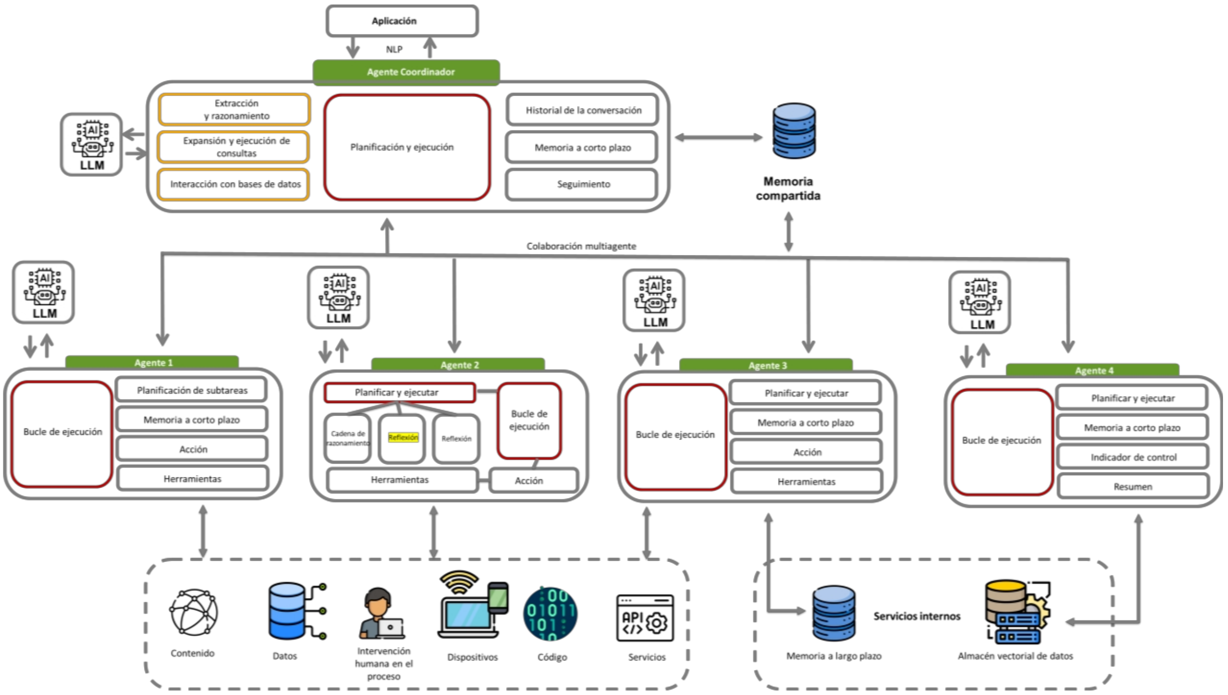


Figure 5 Detail of the architecture of a multi-agent AI system

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

III. AI AGENTS IN PROCESSING

AI agents are means that allow personal data processing to be implemented by introducing greater automation. The same AI agent can be used to implement operations in different personal data processing activities. On the other hand, an AI agent may be part of a processing operation 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, if human supervision is necessary, in processing implemented with agentic AI means, such intervention will have to be considered from the design stage.

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

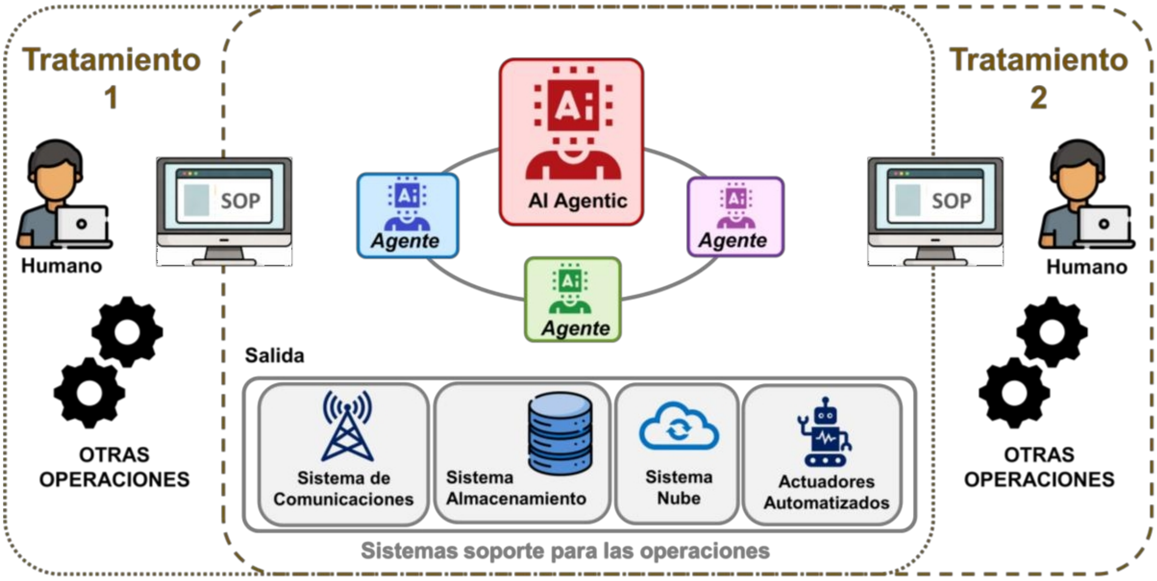


Figure 6 Relationship between agents and processing

The purpose of this document is not to analyze compliance with a specific treatment that uses AI agents, but rather to examine distinctive aspects that could arise in relation to data protection due to the fact that it is implemented wholly or partially with agentic AI systems.

When conducting this analysis, it is important to avoid the "technological fog" that can be caused by simply mentioning agentic AI when implementing a treatment. For example, an agent can implement a common process in any organization, such as organizing a trip for an employee. The agent, proactively supported by AGI, would detect a trip in the employee's calendar and develop a set of tasks, such as contacting various accommodation services via the Internet, checking the currency exchange rate, verifying the status of transport routes, managing services for the purchase of transport tickets, and obtaining an updated weather forecast. Taking other factors into account (by consulting the news), they would make a selection and contact the services again, make the necessary purchases, and forward the planning and documentation to the employee.

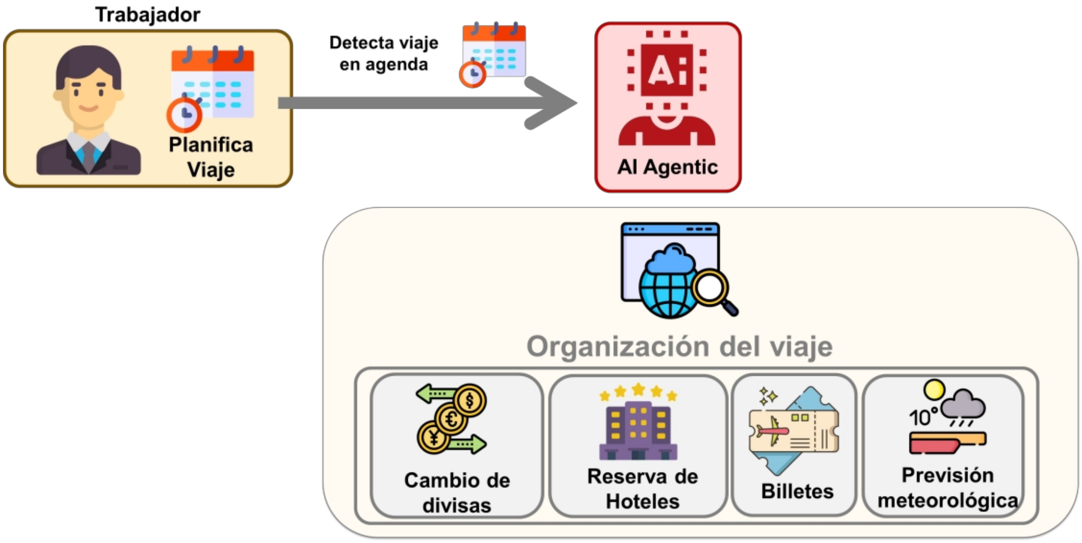


Figure 7 Example of travel management with agentic AI

Traditionally, the services of an administrator who would use the same data and access the same services have been used, or the organization would hire 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, legitimacy, access to Internet services, etc.) will be the same whether it is implemented with an administrative assistant, a travel agency hired as a data processor, or an agent, including the ability or means to determine that there is a trip in the employee's calendar. Therefore, it may facilitate the start of the compliance analysis to identify the same operations carried out by the agent and their relationship with the services they access (or the relationship with the external entity that provides an agent-travel agency[14]) in entities or individuals, and then analyze the different aspects introduced by agentic AI.

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

As described above, agentic AI may 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 also multiple entities under the privacy, cookie, service, and contractual policies of each third-party tool.

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

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

Therefore, the following will be analyzed:

- What new vulnerabilities, from a data protection perspective, could be involved in including AI agents in personal data processing.
- What aspects of data protection compliance need to be reviewed when considering the use of AI agents.
- What threats could exploit or materialize the vulnerabilities detected with an impact on data protection.
- What measures exist and are available both to support regulatory compliance and to avoid critical impacts or manage risk.

IV. VULNERABILITIES AND PERSONAL DATA PROCESSING

In this chapter, we will conduct a preliminary analysis of the most significant vulnerabilities that could arise in a treatment involving the implementation of agentic AI operations. This analysis is not exhaustive, among other factors because it focuses on those vulnerabilities that may have an impact on the processing of personal data and are characteristics of the agentic system as a whole, not of its individual components.

As with any complex system, the power and versatility of agentic AI are also its main vulnerabilities.

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

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

However, it would be inappropriate to adopt a purely additive perspective, as the interaction between the different components can give rise to new vulnerabilities or amplify existing ones, generating multiplicative effects. Ultimately, this type of system introduces a significantly broader attack surface than language models, exposing the system to more complex impacts and threats.

[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, agentic AI has the ability to interact with the environment to execute all or part of the processing operations. The interaction may be limited to the organization itself, or it may extend to external services.

The invocation of tools and services on the Internet are *de facto* partial outputs that agentic AI is performing externally. In particular, they are not oriented towards the user of agentic AI and are not the final result, so they could be transparent to those users or to the data controller, but contain personal data or reveal personal information about the individuals subject to such processing.

- ***Access to organizational and user data***

In relation to the previous section, one of the common features of agentic AI is access to internal services and data for the purpose of enriching the context for task execution. This information could relate to the user, a work group, or the entire organization. Some examples could be email accounts, reports, decisions, internal discussions, meetings, notes, conversations, a customer database, etc. This involves the processing of data belonging to users of agentic AI, which may be personal data belonging to that same user, as well as personal data belonging to other people, both those whose data is being processed and others whose data resides 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 data protection obligations by design and default could lead to massive data processing that would violate the principles of minimization, processing limitation, and accuracy of information if the data is obsolete or there are integrity issues. If all or part of the components of agentic AI are implemented by data processors, this could involve the communication of data to third parties beyond the purposes of the processing. Furthermore, when accessing unstructured data sets, some of the information may be relevant, but other information may be irrelevant or inappropriate from various perspectives.

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

Interconnection to Internet services allows agents to interact with the environment outside the organization, both to gather information and to send information (requests, commands, or locally stored data), increasing their capabilities in terms of both action and information processing.

The existence of bidirectional data communications with multiple participants, without the necessary control by the entity, can significantly increase vulnerabilities, such as the ability to access the control of agentic AI through multiple channels.

As for local information sent abroad, excessive freedom in the use of tools that collect internal information could lead to the communication of unnecessary information by failing to prepare the agent to discern what information is relevant and what is not.

Connecting to the outside world not only to carry out actions but also to obtain information could involve the use of inappropriate sources that are inaccurate, unrealistic, obsolete, partial, biased, or misinformed. This is especially true if there are no procedures in place to verify the reliability, origin, and consistency 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 collected.

B. SERVICE INTEGRATION

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

- ***Service management***

Even when services come from the same provider, the nature of the industry often causes each of them to evolve independently, with non-homogeneous terms and contracts, incompatibilities, service discontinuity, and interface changes. This implies greater complexity in the management of tools, both for the organization's ICT services, the user, and the management of responses by the agents themselves. It also involves the creation of complex data flows and numerous systems that store data at rest in the short and long term (see section "IV.CC. Memory").

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. There are also other functional problems such as integrating heterogeneous APIs, variable latencies, name collisions, nested parameters, misinterpreted dependencies, confusion of models when creating confusing *prompts*, availability, resilience, instabilities and lack of robustness, emergence of cyber vulnerabilities, inconsistency in access and usage privileges, instability of service quality, etc.

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

There are agent-based AI services that are easy to deploy and intuitive, with tools that allow tasks to be designed and components to be connected very quickly, even for end users. These types of environments are common in software prototyping in other contexts and facilitate the deployment of systems such as AI agents.

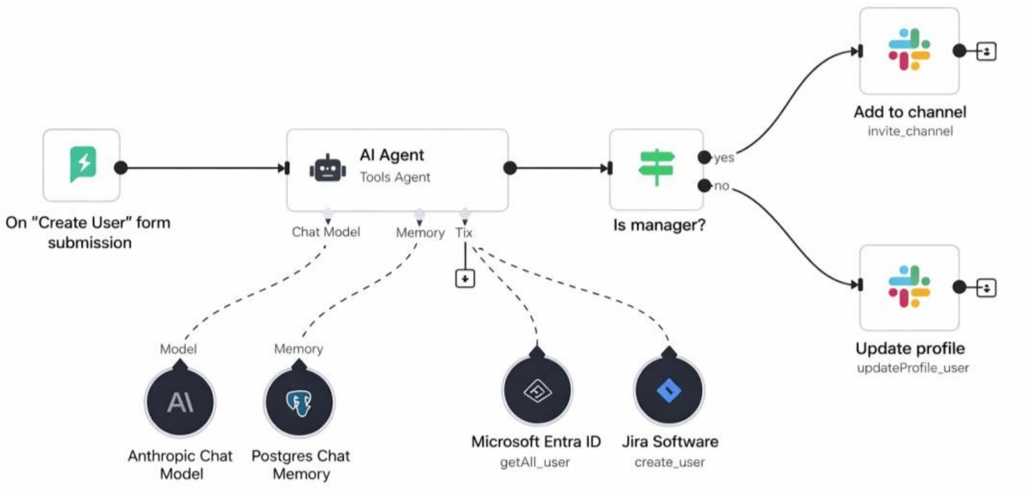


Figure 8 n8n development environment (source: https://n8n.io)

This leads to the temptation for unqualified users to be dazzled by its possibilities and deploy it outside the entity's governance and information policies. The ease of having a solution that seems to work with little effort could create a perception of triviality regarding the implications and impacts for data protection (and for the organization in general) by obscuring the inherent complexity of these developments in many respects.

Introducing an agentic AI system into the processing of a data controller involves redesigning an organizational process in which at least the functional, ICT, and quality managers should be involved, in addition to the DPO when appropriate.

The impacts of errors in the deployment of agent-based AI systems in treatment can affect everything from actual effectiveness to regulatory compliance, reliability, explainability, stability, and robustness of processes, scalability, and availability, the vulnerabilities generated in treatments, the lack of control over data flows, the extension and preservation of such data, the consequences of breaches, the lack of preparedness for incident management, etc.[18]

With mobile devices, the problem of BYOD (*Bring Your* Own Device) arose in the workplace, and with AI chat, the problem of BYOAI (*Bring* Your Own *Artificial* Intelligence) arose, with agentic AI, the problem of BYOAgentic (Build Your Own Agentic) arises, due to the lack of organizational policies and professionals qualified in management and technology, and the use of mature methodologies in the design of processes and applications.

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

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, agents with memory can recognize patterns, adapt over time, and use previous feedback, which is key in goal-oriented applications. Language models alone do not have memory; it must be integrated as an additional component. One of the main challenges is to manage memory efficiently, storing only relevant information to maintain fast responses and low latency.

There are two very different types of memory in agentic AI. One is the memory that enables agent functionality. The other is the memory that enables agentic AI management and allows control mechanisms to be implemented. It consists of all the system operation logs, logs for each of the system components, and logs for the services accessed by the agentic AI.

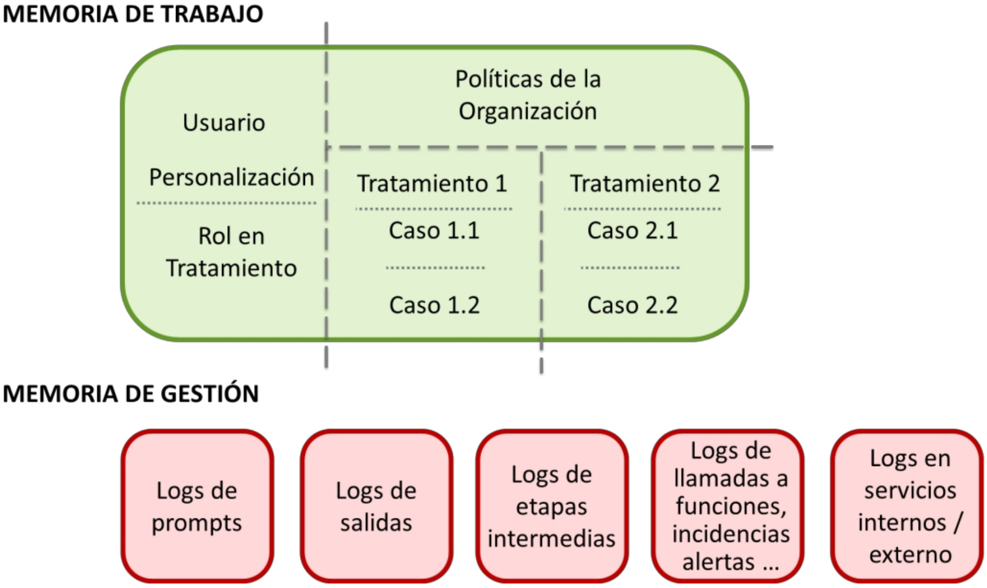


Figure 9 Memory in agent-based AI

- ***Working memory***

Agents use different types of "working" memory, short-term and long-term (depending on the type of agent, we could also talk about medium-term memory). The most simplified approach is that short-term memory allows us to remember previous interactions within an execution cycle[20] and long-term memory allows us to

[19] Currently, services that provide access to LLMs do incorporate memory as they evolve toward 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.

systems to retain information across different conversations or sessions.

Long-term memory can be categorized as follows:

- Semantic memory: involves the retention of specific facts and concepts and can be used to personalize applications by remembering facts from past interactions, creating a continuously updated "profile" with specific information about the user.
- Episodic memory: allows past events or actions to be remembered and is used to enable 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.
- Procedural memory involves remembering the rules used to perform tasks. An effective approach to refining these instructions is reflection or *metaprompts* using an IAG, where the agent refines its own instructions based on its interactions.

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

- Files, SQL or vector databases.
- Division into fragments and context windows that can handle complex entries without getting lost, focusing on the most relevant parts.
- Incorporation of metadata and tagging (dating, users, categories, etc.) to quickly filter the necessary information.
- Retrieval-augmented generation (RAG) technique that allows a knowledge store to be queried 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 implementing a treatment in the organization using agentic AI systems, the information stored in memory could be classified into:

- Organization memory for all processing: this is the information that the organization considers relevant in order to carry out automation within the organization. This unique context may be important in relation to data protection in order to ensure consistency and completeness (see chapter on Measures).
- Memory for each specific processing operation, which is relevant for a single processing operation and not for a different one. It may also contain specific context information set by the organization.
- Memory for each case dealt with in the processing, such as processing that provides a service to a customer and is not relevant to

other cases (depending on the processing, an approximation by case or by customer could be given).

- User memory for the same processing, which may be aspects of personalization or also categorized by processing and by case.

The logical organization of memory could take different forms, from a large repository where data from users of the agent and personal data from each processing operation are stored, leaving the agent in control of which data it will use at any given time:

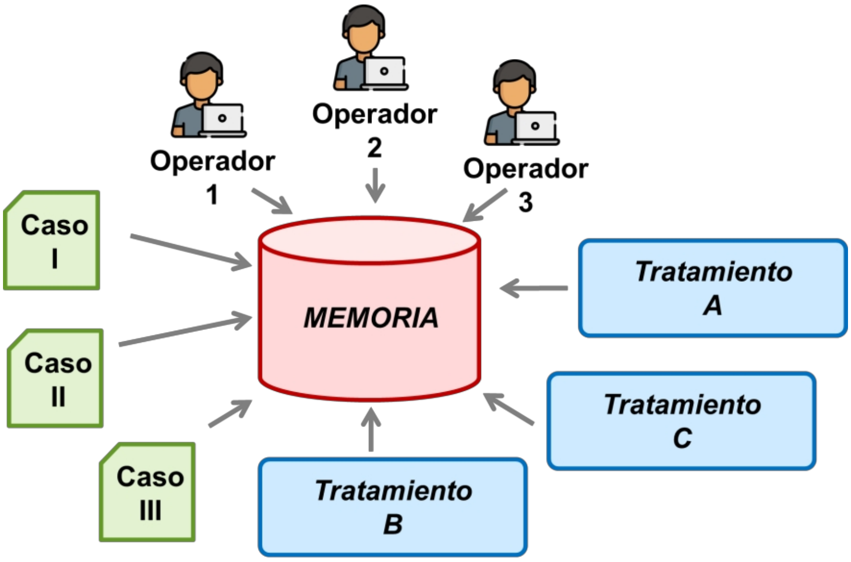


Figure 10 Organization of memory as a single logical repository

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

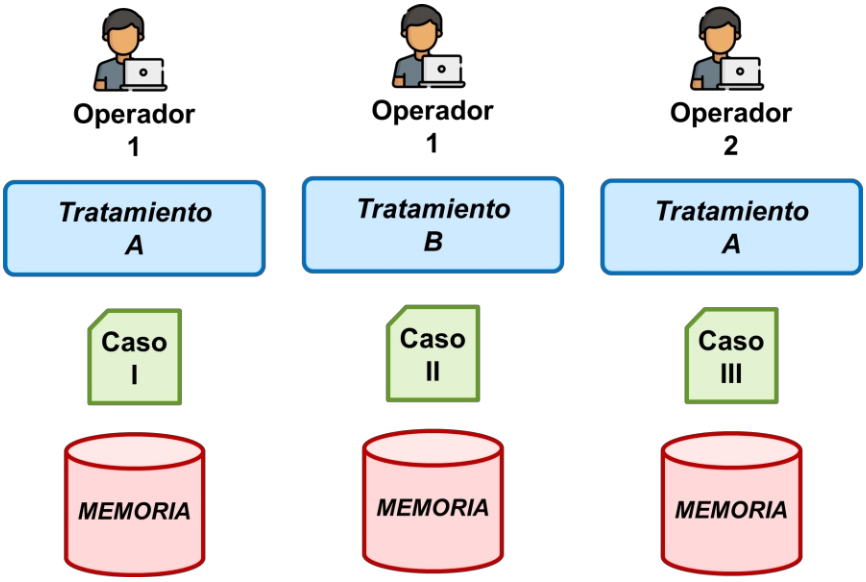


Figure 11 Fully granular distribution of memory

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

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

- Relevance: clear and effective policies must be established regarding what is to be stored in memory for each processing operation. Relevance can be described with *prompts* when analyzed with an LLM or other techniques. Among other things, it should be ensured that there is compartmentalization between different processing contexts (at least), for example, that user credentials provided for one purpose are not accessible for another purpose in which a credential request appears[21].
- Consistency and completeness of context: the information stored must be of sufficient quality (including in relation to bias, relevance to the context, up-to-date, without contradictions), particularly if it is to be used to make inferences or decisions about individuals. This applies to both long-term memory and summaries produced in short-term memory[22].
- Conservation: the information stored must be the minimum necessary for the agent's operation. This refers not only to information relating to trade secrets or industrial property, but in particular to the personal information of the user, the subjects undergoing treatment, or third parties, including information that could infer a profile of any of them.
- Integrity: the stored information allows the results of inferences to be manipulated and the agent's own actions to be changed, meaning that it may be subject to manipulation 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 do not fall within the scope of data protection).

- ***Management memory***

We must also consider the impact that shadow memory can have, which is common in the use of digital systems, such as activity logs. This memory also plays a role in the operation of agentic AI, as it must be exploited to analyze malfunctions, incidents, attacks, alerts, etc.

Depending on how it is used, the memory stored in the logs can be either a privacy measure or have a critical impact or present a risk.

- Data protection measure by design: by enabling auditability of all actions, traceability, repeatability, accountability, deterrence against abuse, etc.

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

- Critical impact, for example, when records store excessive information about users, which becomes hypervigilance and goes beyond preserving privacy and cybersecurity.
- Risk in the event that persons authorized to manage such records fail to comply with their confidentiality obligations, unauthorized processing occurs due to personal data breaches or the use of captured personal information for other purposes (e.g., adjustment of LLMs).

A unique aspect arises when an agentic AI system is used to implement different types of processing. In this case, some of its components, such as LLMs, will use logs to store the activity of all processing operations. For example, they will store the prompts and inferences made on all of them, becoming nodes for collecting personal information from individuals whose data is subject to multiple processing operations, but they could also store data from

This could have a greater impact when these 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 data subjects and the impact in the event of personal data breaches is increased.

- ***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 operations, and also records what accesses or operations are being performed on such personal data, it must be designed to allow for the exercise of all rights under the GDPR, including access, rectification, erasure, restriction, and objection.

D. AUTONOMY

Agentic automation means that agents can act autonomously, without receiving explicit instructions from a human user. This autonomy allows them to decide how the task will be executed, what steps it will be subdivided into, what 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 capability gives agentic AI a great capacity to complete objectives.

Autonomy is significant in the AI agent's interaction with the environment: accessing and updating data repositories, exchanging data between participants, combining said data, managing other processes, and generating results, decisions, evaluations, or other generative content, both internally within the organization and externally outside of it.

The level of autonomy of the agent in the treatment is a design decision made by the person responsible, and could be[23]:

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

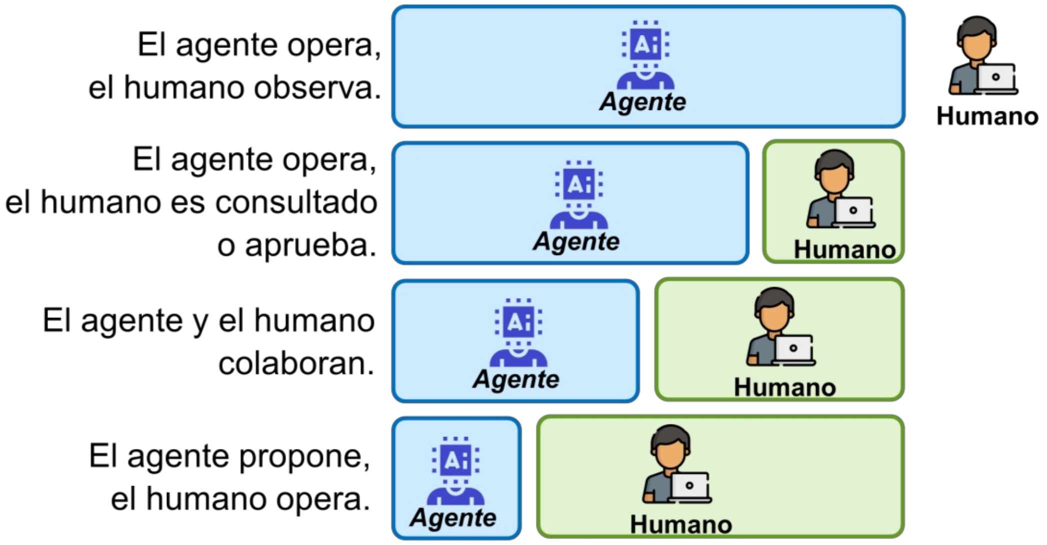

El agente opera, el humano observa.
Agente
Humano
El agente opera, el humano es consultado o aprueba.
Agente
Humano
El agente y el humano colaboran.
Agente
Humano
El agente propone, el humano opera.
Agente
Humano

Figure 12 Levels of autonomy of agents

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

- Whether the data accessed, updated, or exchanged complies with the principles of minimization, accuracy, and limitation of processing.
- If such actions are automated decisions in accordance with Article 22 of the GDPR.
- If such actions have a serious impact on the individual and if they are reversible within the framework of the processing (actions such as deleting the natural person's data from the organization's systems).
- If the necessary human oversight is in place.
- Whether the task has been properly organized, subdivided, and executed to ensure that the processing as a whole actually fulfills the purpose.
- If there is transparency regarding its execution: quality of results, explainability, repeatability, traceability, auditability, and auditing, among others.
- Whether revocation mechanisms are provided in agentic AI and/or within the framework of the processing.

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

- ***Transparency and human oversight***

Users and developers may find it difficult to understand how some AI agents make decisions. A lack of transparency in internal reasoning processes (since decisions emerge from chains of inference distributed among various agents and tools), and limited comprehension by human operators who are not sufficiently qualified, 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 regarding 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 AI agents' decision-making processes can create more pronounced obstacles to achieving meaningful explainability and the transparency necessary for various objectives: demonstrating effectiveness, evidence of robustness, guarantees for customers, legal protection from liability for actions, and, among others, compliance with data protection in terms of citizens' rights.

At the same time, automation bias intensifies, leading users to accept the system's decisions without sufficient critical analysis and reinforcing the 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 to enable, and in some cases require, effective, continuous, and meaningful oversight.

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

Task decomposition mechanisms and interaction between multi-agent systems, together with the orchestration of activities between these agents, enable the execution of highly complex tasks, adding flexibility and adaptability that allows 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 those that are necessary, and in the proper order, taking into account the impact this may have on data subjects (and on other objectives of the organization). There will be processing operations in which decomposition and orchestration will be predefined, at least to a certain level. In others, a reasoning-oriented LLM can do all the decomposition.

[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)

It should be noted 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 this purpose. On the other hand, there is no possibility of contamination between different incompatible learned models (e.g., sub-tasks of administrative procedures from different jurisdictions).

Technical complexity can lead to instability in emergent behavior, i.e., unpredictable and undesirable dynamics characteristic of complex systems, which cannot be anticipated or explained solely by analyzing their individual components. As a result, unforeseen outcomes or infinite planning loops may occur.

Likewise, dependence on sequential calls to external tools can generate round-trip loops that accumulate latency, especially in multi-step tasks where each phase depends on the results of the previous one.

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

Compounding errors are a 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 constructed *query*, receives incomplete data, processes it as complete, and makes erroneous inferences, which lead to the execution of wrong tasks, etc.

With regard to compound errors, whether from internal sources, external sources, or intermediate inference results, they can generate information or lines of reasoning that do not comply with the entity's policies or regulatory requirements, such as access to excessive or insufficient quality data, erroneous inferences, trade secrets, the organization's ethical values, objectives, biases, financial information, and, among others, data protection considerations. All this information can produce results that deviate from the purpose and create harm 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 or centralized planning procedure, which could communicate through shared memory or messaging protocols, the breach of a single element of the aforementioned could compromise all the processes that use that system.

[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

- ***Non-repeatable behavior***

Inference is the process by which large language models (LLMs) and other AI models, including agent-based systems, make decisions. In classical machine learning, *one-shot* inference implies that a given input produces a reproducible and deterministic output, provided that 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 information sources, the services accessed, their versions, task planning, memory, and the possible commands generated from all these elements, it is not possible to anticipate the system's output.

This problem is not inherent to the nature of agentic artificial intelligence, but rather responds to specific implementations in which adequate control of the constructed system is not available. 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) divert the entire plan, producing different trajectories in each execution. In this context, reasoning loops or infinite loops, opportunistic strategy changes, and emergent behaviors that were not explicitly programmed may appear.

Consequently, unpredictable inference in agentic AI systems is a significant problem, as it prevents accurate anticipation and control of the system's behavior 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.

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

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

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

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

email or cloud information) 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 AIs 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.

Granting excessive permissions to agents is a critical factor, as a model with broad privileges can be used as a pivot point between different systems, so that an initial compromise leads to unauthorized access to databases, internal services, or sensitive credentials, amplifying the impact of the incident.

Likewise, the absence of isolation mechanisms between services significantly increases the risk of remote execution of commands and code with adequate privilege levels from untrusted inputs. For example, the agent (in exchange for access to services such as news, either by mistake or motivated by an attack) could be giving consent or entering into contracts on behalf of a user or establishing new user-responsible 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 hinders the effective application of access controls, auditing, and privilege revocation, and increases both the risk of internal threats and the exposure to external attacks, especially in highly distributed and dynamic environments typical of AI agents.

V. ASPECTS OF COMPLIANCE WITH DATA PROTECTION REGULATIONS

A personal data controller could choose one (or more) agent-based AI systems from among the means to be used to implement the processing. Whether it is a fully local, fully remote, or any intermediate service, it will be a system that forms part of an entity's technological infrastructure and could implement operations (including all operations) of one or more processing activities.

When AI agents are used in processing operations, the following issues may arise:

- The appearance of more participants than the controllers or processors who are originally part of the processing.
- Greater extension in the type and categories of data of the subjects who were the object of the processing, including additional profiling.

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

- Greater extension of subjects whose data is processed, beyond the subjects who should be the subject of processing, which could be collected from the environment accessible to the agent or from the agent's own memory.
- Greater processing of data from users (employees of the organization) who interact with AI agents.
- Less transparency in processing.
- Data retention in more stakeholders and systems.
- New purposes.
- Automated actions affecting data subjects.
- New impacts or risks to the rights and freedoms of data subjects.
- Others, depending on the processing and how agentic AI is implemented therein.

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

A. DETERMINATION OF PROCESSING RESPONSIBILITIES

The controller is the person who, alone or jointly with others, determines the purposes and means of the processing, regardless of the form of those 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 new risks in processing that could arise from the use of agentic AI systems.
- Analyze the proportionality of critical impacts[27]that could arise from using agents.

When an AI agent runs entirely locally, there would be no further analysis of liability. However, in most cases, 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 the entire agentic AI as a service provided by another entity.

[27] An impact with absolute certainty of occurrence is called a critical impact. For example, in the processing of personal data that is legitimate for some reason to record all communications made by a person, it 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 due to a breach, there would be an additional impact, which is why there is a risk. Critical impacts usually derive from the definition of the processing itself and could also be reduced by taking measures to make it proportionate to the purpose of the processing, in many cases by changing the definition of the processing itself, but it is not a risk, it is a certainty.

AI agents enable the automation of processing operations with the support of IAG. In order to study the liability relationships of a processing operation, it is necessary to analyze case studies that will be found by the same controller who is related to other entities that would provide the same service when implementing the processing without AI agents.

Therefore, it should be analyzed without prejudice to the additional data processing that may occur through 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 in the processing:

- The agent may access third-party services to obtain non-personal information, such as a service schedule, the value of financial assets, historical data, etc. In addition, the implementation of the agent may not allow such services to link them to a specific user (there are no identifiers, cookies, or history linked to the specific user, as this is filtered out in the use of the agent's action). In this case, the entity providing such a service will have no role in the data protection framework, without prejudice to other regulatory areas.
- The agent may send non-personal information to third-party services to perform any process: information storage, text translation, rule analysis, reasoning about a task, etc. (where a language model could be involved). If, as in the previous case, the service does not link it to a user, there would be no data protection relationship. However, if it links it to a user for the purpose of processing, for example, to save the context of interactions with the user, it would be a data processor.
- In the above case, if 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 records held by public administrations or entities. Without prejudice to the legitimacy of access, the relationship that could be established would be one of communication between controllers. However, if the agent is accessing the services of a car rental agency contracted by the entity to obtain billing information for an employee, this would be a controller-processor relationship.
- The agent may access a third-party service to transmit personal information relating to data subjects. In this case, the relationship between the controller and the other entity must be analyzed regardless of the use of agent AI systems. For example, if an agent at a healthcare center provides services to an insurance company with respect to

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

of healthcare expenses incurred under an insurance contract, it automatically contacts the insurer's service to transmit the data, this would be a controller-to-controller relationship[29].
- In the event that the agent itself is a service provided by another entity, and insofar as it processes users' personal data and/or customers' or citizens' personal data within the framework of the original processing by the controller, the entity providing the service will be the processor, as in the previous example of the travel agent.

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

To the extent that the incorporation of agentic AI systems involves a relationship with Internet services or services from other entities (whether processors or third parties), the same issues will arise as when there is no automation in the processing:

- The use of personal data provided in the processing for purposes other than those of the original data controller. For example, retraining of LLMs, security, or others.
- The creation of new relationships of responsibility with, where applicable, users or those whose personal data is 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 operations are legitimate. The second case could also be legitimate, provided that there is no misunderstanding on the part of the user as to with whom they are establishing the relationship and that the agentic AI is not allowed to give consent in an automated manner without some form of control.

In all cases, the data controller must take care to monitor these situations, and this will depend on the type of agentic AI solution incorporated into the processing. If the agentic AI system is implemented by the controller itself, measures may be taken to configure the AI agents to determine which services will be accessed (see chapter "VII. Measures"), evaluate contracts or terms of service, review data protection clauses, cookies where applicable, determine the lawfulness of such processing and the guarantees of regulatory compliance, analyze the risk to data subjects, and assess whether the use of such a service is proportionate or whether it is more appropriate to seek alternatives. It is necessary to assess the degree of regulatory compliance of the alternatives analyzed, in particular in relation to, among others, Article 28 of the GDPR, international transfers, data retention, etc.

[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

In the event that the entire agentic AI service is outsourced to another entity, the same obligations mentioned above apply, in addition to those relating to the chain of sub-processors.

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

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

B. Transparency

In the event that the use of agentic AI systems in processing involves additional recipients of the data to those provided for in the processing itself, which will occur in many cases, their identity must be duly disclosed. If, for example, this means that personal data, whether of users or of persons subject to processing, is sent to an AI service of another entity, both categories of persons must be adequately informed.

Similarly, you must be informed of any changes that, due to the use of agentic AI systems in the processing, affect the retention periods for personal data or, when it is not possible to determine this precisely, the criteria used to establish said period. You must also be informed of any additional automated decisions (see section "V.F. Automation of decisions") or additional international transfers (see section "V.I. International transfers").

When the incorporation of agent-based solutions or artificial intelligence systems into processing involves the further 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 additional relevant information, in accordance with the provisions of Article 13(2) of the 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 for exercising those rights.

C. Legitimacy, minimization, and lifting of prohibitions

The inclusion of agentic AI systems in processing could involve additional data processing, although not necessarily. For example, if the administrator who managed travel is replaced by an AI agent within the organization itself, and this agent has interfaces with the same services that were consulted manually, the result will be the same data processing.

Furthermore, the use of agentic AI can result in, or guarantee, less data processing, as it may be the case that the user's data processing has been suppressed when accessing such services via the Internet, since cookies or profiling could not be carried out. In any case, the use of an AI agent is not an end in itself.

If the implementation of agentic AI systems does not involve additional processing beyond the original processing, it will not be necessary to seek legitimation for their inclusion in the processing. It should be noted that, as agentic AI systems are composed of digital systems, some of which are very complex, they will involve more cybersecurity processing, which will be covered by legitimate interest as long as it is aimed at that purpose, necessary, and proportionate.

If additional processing is required, it must have a legitimate basis and, in the case of special categories of data, a circumstance that overrides the prohibition. If the basis is legitimate interest, it must pass an assessment that the purposes of that legitimate interest are clearly identified, the processing is necessary for the purposes of the legitimate interest or interests pursued, and that the legitimate interest or interests are not overridden by the interests or fundamental rights and freedoms of the data subjects (also known as a "balancing test").

If it is based on consent, it could be argued that measures are necessary for the management of such consent (see section "VII.G. Consent management").

Data minimization must be considered from the design stage of processing operations and transferred to the design or configuration of agents. For example, suppose that, within the framework of a processing operation, it is necessary to determine whether an employee is on the guest list for an event. To do this, the guest list could be downloaded and the personal data of the employee and, incidentally, of everyone else on the list could be processed. It would be worth considering whether it is possible to achieve the same purpose without processing the entire guest list (for example, by asking the organizer if the employee is on the list) or even without exposing anyone with "zero-knowledge" strategies. In this example, it has not been established whether the processing is being carried out by a human operator or by 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 limitation of treatment must also be addressed from the design stage. For example, take a customer service treatment in which the action of agentic AI is

When a claim is made by 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 in order to provide a better response in the future. It is necessary to consider whether it is necessary to store personal data from past customers in the memory that the agent will use to perform future actions. If so, it is also necessary to consider the legitimacy of processing personal data from other customers that may be stored in long-term memory in the development of each execution of the agent AI.

D. RECORD OF PROCESSING ACTIVITIES

The record of processing activities (RPA) is an essential tool for managing compliance with data protection regulations, as it serves as a catalog of personal data processes.

When a decision is made to replace traditional means with agent automation in a processing operation, the RAT must be updated to determine, for example, whether this entails a change in the categories of personal data being processed, whether it is necessary to update the information relating to the categories of recipients to whom the personal data have been or will be disclosed, including recipients located in third countries or international organizations, whether new transfers of personal data should be detailed, whether retention periods should be modified, or whether the general description of the technical and organizational security measures referred to in Article 32(1) of the GDPR should be updated.

The GDPR requires minimum information in the RAT, but not maximum. It is advisable to integrate the RAT into the entity's quality control system process catalog as a management tool to ensure and demonstrate compliance. In this 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 use of agentic AI systems in processing should not undermine the exercise of rights, and the necessary measures must be implemented to guarantee them. This means understanding how personal data storage and operations work and providing for measures and procedures to exercise those rights (see section on Measures).

It should be noted that the memory of the agentic AI system stores personal data within the framework of one or more processing operations. In addition, the logs will store information on both the users of the agentic AI and the persons subject to the processing, and may even store data on persons who should not be subject to processing. The configuration of both memory systems and agentic AI must be technically capable from the design stage to allow for the management of data subjects' rights.

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

- The design of the agentic AI, in which the data controller will have been involved, at least in choosing an agentic AI system as a means of implementing processing operations.
- The configuration of the specific system, for example, with *prompts* defined by the system administration, or scheduled events (by the user or the organization) that initiate automatic operations.
- *Prompts* made by users of agentic AI, which can trigger many operations on personal data.

In the latter case, it should be noted that *prompts* made by individuals may be subject to a duly justified request for access.

It is also necessary to consider that access to services outside the organization that act as data processors may result in the storage of personal data in their log files or in the possible memories of their own agents. Personal data may also belong to users of agentic AI, especially when the entire agentic AI system is a service contracted to a processor.

F. AUTOMATION OF DECISIONS

The automation of decisions and the degree of autonomy of the agent is a matter of treatment design, including technical design factors, human intervention design, and actual implementation. The controller can manage how decisions made by the agent or agentic AI will be handled, which actions will be allowed to be automated without supervision, and also the measures to manage these design decisions (see chapter "VII. Measures").

▪ *Article 22 of the GDPR*

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

There are personal data processing operations that do not involve automated decisions. for example, an AI agent may be used in the organization to track and select events via the Internet and prepare summaries and analyses based on the company's objectives and employee categories, sending them to the devices of those employees based on the interests they have declared, without this constituting an automated decision within the meaning of Article 22.

However, if such decisions do exist, the conditions that allow them (Article 22.2 of the GDPR) and the measures that will need to be implemented (Article 22.3 of the GDPR) and the limitations on the use of special categories of data (Article 22.4) and on decisions on minors (cons. 71 of the GDPR). In addition, information on the existence of automated decisions, including profiling (Articles 22, 13(2)(f), and 14(2)(g) of the GDPR), providing, at least in such cases, meaningful information on the logic involved, as well as on the significance and the envisaged consequences of such processing for the data subject.

Automated decision-making should also be assessed given that, 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 concerning him or her or similarly significantly affects him or her." Even if the decision-making process does not affect individuals' legal rights, it could still fall within the scope of Article 22 if it produces an equivalent or significantly similar effect in its consequences. For data processing to significantly affect an individual, the effects of the processing must be sufficiently important to warrant attention.

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

- significantly affect the circumstances, behavior, or choices of the individuals concerned;
- have a prolonged or permanent impact on the person concerned; or
- in the most extreme cases, lead to the exclusion or discrimination of individuals.

- ***Other automated actions***

The use of agentic AI in processing may involve risks to 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 by email or file transfer services may have an impact on the confidentiality of personal data.

This problem, and others that may arise from automated actions, must be taken into account in risk management for the rights and freedoms of data subjects. In particular, reversibility of certain actions by AI agents should be built into the design.

G. RISK MANAGEMENT

As with any innovative treatment or process that incorporates modifications to its implementation or new technological systems, adequate risk management is necessary.

Risk management involves a critical analysis of the future impact of the processing, beyond the context of the organization, in order to manage potential problems (threats) before they become real problems, i.e., before they materialize. It

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

This 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 prioritized, and then efforts must be coordinated and decisions made to avoid or minimize their likelihood or impact.

Therefore, it goes beyond the scope of the system, in this case the agentic AI system, and covers all elements of 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 modeling framework[31] modeling framework, which will enable data controllers to identify threats of Linking, Identification, Inaccuracy, Non-repudiation, Exclusion, Detection, Data Breach, Deception, Disclosure, and Unawareness/Unintervenability.

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

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

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

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

- ***Rule 2***

A simplified approach to setting a minimum threshold of safeguards that must never be exceeded was set out in 2021 in relation to the execution of applications in browsers from a cybersecurity perspective alone and became known as Rule 2[32]. It has subsequently been reformulated for AI agents by various authors[33]in the following form:

[31] AEPD, "Introduction to LIINE4DU 1.0: A new methodology for modeling threats to privacy and data protection," October 2024. Available at: https://www.aepd.es/guias/nota-tecnica-introduccion-a-liine4du-1-0.pdf

[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/

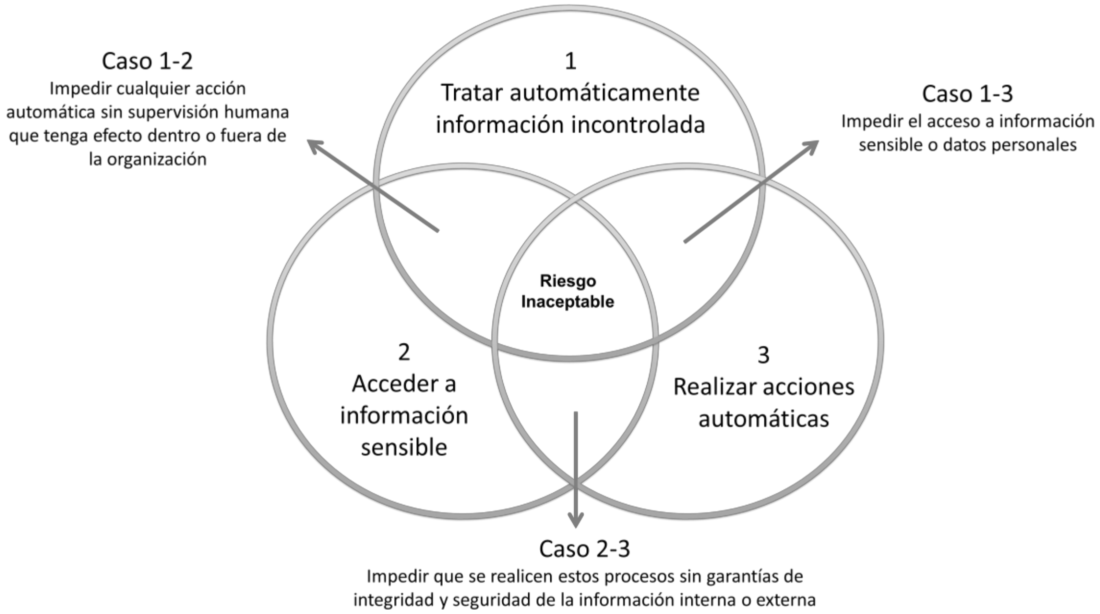

Caso 1-2
Impedir cualquier acción automática sin supervisión humana que tenga efecto dentro o fuera de la organización
1
Tratar automáticamente información incontrolada
Caso 1-3
Impedir el acceso a información sensible o datos personales
Riesgo Inaceptable
2
Acceder a información sensible
3
Realizar acciones automáticas
Caso 2-3
Impedir que se realicen estos procesos sin garantías de integridad y seguridad de la información interna o externa

Figure 13 Rule 2

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

- The specific implementation of this AI agent allows emails to be received without any guarantee that there is no technical or social engineering attack.
- The agent could access sensitive information in the user's systems without restrictions, and
- The agent could then automatically initiate actions (such as creating a reply email, manipulating the agent's long-term memory, rewriting sensitive information in other repositories within the organization, etc.).

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

Rule 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 could trigger access to sensitive information, any automatic action without human supervision that has an effect inside or outside the organization must 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 of automatically processing uncontrolled information that could trigger automatic actions, the agent must prevent access to sensitive information or personal data.

- ***Risk of processing***

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

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

In short, with regard to the above examples, risk management must be carried out on the processing in which the agent system is a means of protecting the rights of data subjects, where one part is the management of the security risks of the agent system itself.

- ***Side effects of processing***

The implementation of treatments, especially those involving novel techniques, may give rise to undesirable side effects that are beyond the objectives of the data controller[34]. These side effects may affect the individuals undergoing treatment or result from the processing of data belonging to users of 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 from customers or users of the agentic AI (also using "shadow leak" techniques, see the Threats chapter), 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 of such information would depend on the sensitivity of the processing implemented with such agentic AI, and would be worse if there is no memory sharing between processes.

- ***Data protection impact assessment***

AI agents are undoubtedly a new technology, but this does not necessarily mean that a data protection impact assessment (DPIA) is required in all cases. It will depend on the type of processing involved and the type of AI agent system to be used. It could be the case that when a type of AI agent system is used for different types of processing, for

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

Some treatments will not require an EIPD, others will, and for those that already had an EIPD that had been successfully completed, it will need to be reviewed[35].

- ***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 separate, risk mitigation actions 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, data protection obligations from the design stage onwards would not be fulfilled, and at the very least, the effectiveness of risk management for rights and freedoms in relation to the processing of personal data would be compromised.

The following two chapters 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 controller shall implement appropriate technical and organizational measures to effectively apply 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 the processing, and the selection of the type of agentic AI system and the configuration of the system and its components must take all these factors into account from the outset.

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

The aspects that, in particular, will need to be managed are data minimization, avoiding the "default memory" of unnecessary data or uncontrolled user activity logs, prohibiting the reuse of data for secondary purposes without legitimacy, paying attention to the design of the processing of special categories and their retention, and considering human supervision, among others. Chapter "VII. Measures" below details techniques for implementing data protection by design and by default, as well as for managing risk.

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

However, the application of data protection techniques from the design stage and by default should be applied beyond a purely reactive manner, i.e., in circumstances that prevent processing, but also proactively. Proactive application involves introducing measures that will improve the protection of the rights and freedoms of data subjects through the use of agentic AI over other traditional forms of 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 using SML in conjunction with other systems, in the intermediate stages of reasoning chains for categorization, sanitization, minimization, and alerting in data exchanges. Another example is that, in stages where human intervention is necessary (such as signing a decision about a person), anonymized information can be provided in certain aspects to avoid bias in that intervention. Another example would be to access sensitive data that is necessary within the framework of processing, but without exposing it to human operators.

In these reactive and proactive actions, it is essential to involve DPOs and data protection advisors who are duly qualified in understanding these technologies and the measures that can be adopted from the design stage.

I. International transfers

If the inclusion of agentic AI systems in a processing operation results in additional transfers of personal data to a third country or an international organization, it must be ensured that these are carried out with the safeguards provided for in Chapter V of the GDPR and that adequate information is provided, including through the record of processing activities, including the identification of the third country or international organization of destination and, in the case of transfers referred to in Article 49(1), second paragraph, of the GDPR, documentation of the appropriate safeguards applied.

If such guarantees do not exist, it will be necessary to consider redesigning the agents or choosing another type of agentic AI.

VI. THREATS

As previously analyzed, the integration of AI agents into corporate processes introduces a new and expanded attack surface that goes beyond simply deceiving IAG models. This risk surface is considerably more complex, as it originates both from legitimate and authorized processing of personal data and from possible unauthorized manipulation resulting from operational autonomy, system interconnection, and access to multiple sources and tools.

Next, and taking the vulnerabilities described above as a reference, we present some of the main threats associated with the implementation of treatments that incorporate agentic AI that have implications on data protection, without going into others that could affect other objectives of an organization, such as cybersecurity for the protection of the organization itself (not the data subjects), effectiveness and efficiency, fraud, labor, 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 constantly changing in 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 corporate image, this analysis focuses specifically on those that directly affect personal data protection and compliance with associated regulatory obligations.

A. Arising from authorized processing

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

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

The underlying threat that can prevent the effective implementation of the GDPR in the organization is the failure to integrate agentic AI as a system that must be managed within the framework of process governance and the entity's information and quality assurance policies.

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

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

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

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

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

In particular, it is important to involve DPO and data protection advisors who are duly qualified in understanding these technologies.

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

Implementing agentic AI processing without having properly configured the organization's or users' data access policy, especially when dealing with unorganized information repositories, 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 action process that should not be taken into account.
- Communication of data to third parties outside the purposes of the processing, when agentic AI, or any of its components, such as LLMs, have access to such data. This could also occur when agentic AI invokes Internet services.
- Processing of inaccurate or obsolete data, by including historical information about individuals that is not relevant in the inferences.
- Exposure of personal data of users of agentic AI, when accessing, for example, explicit or implicit contact lists, CVs and aspects of professional activity, browsing history, etc.
- Exposure of data belonging to third parties who are not the legitimate subject of the processing, for example, when accessing emails or meeting minutes that include addresses and data belonging to third parties.
- Integrity issues, given that data can be modified, enriched, or altered.

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

The drift of a reasoning process could give rise 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 parties.
- Breach of the principle of minimization, both by processing excessive data and generating inferences about new categories of data.
- Processing of special categories of data in breach of Article 9 of the GDPR, for the same reasons as in the previous point.
- Breach of the principle of accuracy, both by using obsolete or erroneous information about individuals and by inferring erroneous personal data.
- Breach of the principle of processing limitation.

- Personal data breaches.
- Automated decisions in breach of Article 22 of the GDPR.
- Risk of high-impact and/or irreversible actions affecting individuals.
- 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 natural person's data from the organization's systems).
- Lack of transparency that allows for reporting and assurance regarding the quality of 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 that can be made, without debugging 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 principles of minimization, accuracy, and limitation of processing, in addition to compromising data security.

— *Misalignment*

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

Goal misalignment can occur when pursuing a purpose without considering the actual objectives of the treatment (e.g., biased inferences), behavioral misalignment (e.g., disclosing sensitive information to third parties), emergent misalignment, or harmful behaviors (e.g., malicious advice).

— *Feedback loops and bubble effects*

Feedback can occur when the agent is generating content that, stored in long-term memory, can in turn be used to generate new content. AI agents generate feedback *loops* that optimize autonomous adaptation, but they also generate risks such as amplified biases, behavioral 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 the *feedback* is manipulated.

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

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

This can cause these types of loops to condition human supervision, with the impact that this 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 external information access tools that can be invoked, the number of accesses that can be made, without debugging the arguments of the functions to limit the amount and category of data accessed, and without filtering and analyzing the information being accessed in relation to the processing could violate the principle of minimization and expose data security.

In particular, not including controls in "deep *research*" agents, which can autonomously analyze hundreds of sources on the Internet, could lead to massive, automated *scraping* of scattered 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 its forwarding to other systems, violating the principle of data minimization.

- ***Shadow leak exfiltration***

Shadow-leak consists of the silent and progressive leakage of sensitive information, such as data, internal context, memory, rules, or secrets, through seemingly legitimate interactions, fragmented and innocuous queries, and partial responses from the model. Each response, considered in isolation, appears secure and authorized, without causing obvious violations or triggering security mechanisms, but their combination allows confidential information to be reconstructed.

For example, the attack can be carried out by exfiltrating memory or context through repeated queries about past decisions, successive reformulations, or the inference of patterns stored in the agent's memory; also by inferring 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, overly informative error messages, or differential behaviors depending on the context.

- ***Shifting all responsibility to the user or human oversight***

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

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

This phenomenon is not unique to agentic AI and arises when attempts are made to compensate for underlying problems in the design of treatments, agentic AI systems, or governance in general by shifting them to human supervision.

Users of agentic AI within the framework of an organization's processing, such as those who supervise certain actions, must have clearly assigned responsibilities, but within certain limits. Neither role can replace the data controller's duty of care in the design of the processing and the selection of the agentic AI used as a means.

- ***Lack of compartmentalization of the agent's memory***

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

- Excessive processing of personal data, by incorporating data corresponding to other processing operations involving the same subject into the inference or action process.
- Communication of data corresponding to other processing to third parties involved in the present processing.
- Processing of personal data of the user of the agentic AI within the framework of a processing operation in which they are not an interested party (or such data are not necessary).

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

Closely related to all of the above is the failure to consider the case 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.

Furthermore, the absence of data filtering and sanitization mechanisms, such as the removal of hidden metadata, will expose personal data and sensitive information. Such metadata may contain references to authors, locations, editing histories, or technical identifiers that facilitate the identification of individuals or internal processes.

- ***Excessive data retention***

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

- ***Automation bias***

Even if the treatment has been designed to include human supervision, there is a possibility that the implementation of such supervision may be incorrect due to multiple factors (lack of resources to interpret the results, lack of training or motivation, implementation of black box in the agentics, etc.).

Among these factors is automation bias, which can be exacerbated by users' trust in the system and a lack of information.

- ***Profiling users of agentic AI***

The existence of long-term memory, metadata, and information stored in the various logs of each component or service allows for the creation of detailed behavioral profiles that could reveal sensitive patterns. These behaviors could be, for example, those of employees in the context of the employment relationship.

- ***Availability and resilience***

When operations depend on interfaces with Internet services that are not under the organization's control, and for which no alternatives are available, the system may be compromised by changes in the operation of those systems, their service quality parameters, their data formats, or the continuity of the service itself.

- ***Access to agentic AI by unqualified users***

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

- ***Commitments in the supply chain***

A lack of diligence in selecting compromised language models, vulnerabilities in libraries, and software components can compromise personal data and confidential information processed by agentic AI.

B. ARISING FROM UNAUTHORIZED PROCESSING

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

- ***Prompt injection***

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

- Direct: In a direct *prompt* injection attack, an actor, who may even be a legitimate user[37], enters inputs specifically designed to induce the agent's LLM to behave in a manner not intended by its designers. Through this mechanism, the agent can be instructed to ignore organizational guidelines and policies, allowing for excessive or biased processing of personal data.
- Indirect: An indirect *prompt* injection attack hides malicious instructions in the data sources consulted by the agent, rather than entering them directly as a user *prompt*. For example, you can

[37] The user is assumed to be authorized to access the agentic AI, but not authorized to attack it, so we consider this to be unauthorized processing in this section.

be introduced into a PDF file, email, or web page as instructions invisible to humans, but which the agent's LLM interprets as legitimate commands or information to be taken into account in decision-making, which can lead to data exfiltration, circumvention of controls over automated decisions, inaccurate inferences, or biases.

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

Prompt injections can be used to attack agent-based AI systems in various ways, some of which (which can be combined with each other) are:

- *Memory poisoning and RAG*

It consists of introducing malicious documents into the internal repositories that AI consults to enrich its responses. In this way, such content is stored as persistent knowledge. When 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 human 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 an incoming email) without the user having to interact with the chat or click on any links. The AI simply reads the message and the malicious content is activated. For example, an attacker sends an email with invisible instructions (e.g., white text on a white background) and when the agent analyzes the email to summarize it, the system obeys the hidden command. It is a "zero-click" attack because it occurs 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 disguised as a parameter in the URL of an image that the agent attempts to load from the attacker's server. The attacker only needs to check 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 like a digital "worm," abusing the legitimate user's permissions and tokens.

— *Social engineering targeting AI*

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

— *Long pipeline attacks*

Instead of a direct attack, the adversary could introduce malicious information early in the reasoning chain, knowing that:

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

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

— *Context confusion*

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

— *Delayed trigger attacks*

In this case, content that appears harmless at first can be used, but is only activated 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 remote takeover of *the workflow* for purposes such as stealing authentication tokens, faulty input validation, open keys, or unauthorized data sharing.

— *Screen takeover*

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

— *Ransomware and deletion attacks:*

If control of the agentic AI system is taken over to manage files, it can be instructed to execute commands to delete data en masse, selectively delete data, or block access to critical resources (data or services) that cause disruption.

availability or that prevent actions from being taken or decisions from being made about people with the necessary quality.

- ***Availability and resilience 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 may occur that paralyze the agent, creating an availability gap when processing personal data, or inducing it to generate erroneous responses that affect the decisions it may make about individuals.

- ***Illegal access to agent memory***

Unlike memory poisoning, in this case the objective is data extraction, although some of the attack methods described above may be used. Unauthorized access to the agent's memory, including information contained in the agent's activity logs, its components, or accessed services, allows an attacker to obtain personal data from subjects being processed, third parties, or the users of the agent AI themselves.

VII. MEASURES

There are multiple measures that allow you to reap the benefits of including agentic AI as a means of processing while ensuring and being able to demonstrate that the processing complies with the GDPR. Below is a non-exhaustive list of measures which, as in previous cases, focus more on the unique features of the agentic system than on specific aspects of its components. 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 enable compliance with data protection regulations in processing operations that use agentic AI systems (such as consent management).
- Secondly, to reduce the critical impacts that may arise in a processing operation in order to pass a proportionality test in the context of, for example, the assessment of legitimate interest, compatibility of purposes, or PDPI.
- Finally, to mitigate the risk to the rights and freedoms of data subjects that may arise in processing operations where some or all of the operations are based on agentic AI.

To meet these objectives, it is necessary to select objective measures that either eliminate vulnerabilities or reduce or limit the impact of specific threats that generate risk. Therefore, they must be selected because they objectively fulfill their purpose, and the accumulation of measures without evidence-based analysis ("*checkbox* security" or "*security* theater") must be avoided.

A. Governance and management processes

The existence of an information governance framework within the entity, which includes agentic AI systems and 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 enables, among other things, compliance with data protection regulations, as well as other objectives of the entity and regulatory obligations that may be applicable depending on the case. Governance must be unique. What is important is to ensure that the governance elements arising from the use of agentic AI in processing can be "mapped" onto existing ones or, if not, created.

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

- ***Accepting the possibility of failure***

The reality of personal data processing leads us to conclude that it could have an unforeseen impact on the rights and freedoms of data subjects, both through authorized operations and collateral effects, as well as through unauthorized processing[40]. The more complex the implementation of processing operations, the greater the likelihood of errors and unintended consequences, including failures that go beyond personal data breaches, to the point where we must assume that these will occur.

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

Following the principle of fail-safe (in the sense of "safe," not "security"), it is necessary to design treatments, adapt the systems that form part of the means of treatment, and prepare response plans for measures to minimize impact and manage incidents when they occur.

- ***The Data Protection Officer***

Within this governance framework, it is important to include a DPO or data protection advisor who is familiar with data protection regulations, the characteristics of the processing operations concerned, the technical and organizational measures that can be implemented to ensure data protection by design and by default, and to guarantee

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

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

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

compliance and manage critical impacts and risks to people's rights and freedoms.

- ***Basic elements to be incorporated into the organization's 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, identify, and, where appropriate, integrate into the roles already defined in the organization, those roles related to agentic AI systems (such as functional managers or AI managers).
- Anticipate the possible side effects of including agentic AI in processing.
- Include compliance issues, critical impact, and risk that may be involved in including agentic AI systems in processing operations.
- Determine the use cases for each treatment and the different user profiles.
- Criteria for selecting agents, their components, and connections to the outside world.
- Control the redesign of personal data processing when agentic AI systems are included.
- Consider human supervision when necessary.
- Make the necessary adaptations to the internal services to which it will be connected.
- Formalize relationships with third parties that enable the deployment of agents (model developers, agentic AI providers, and other external services), ensuring that measures are in place to fulfill their own responsibilities. In particular, clarify the distribution of obligations in the terms and conditions or contracts between the organization, service quality levels, and functionalities to maintain control, privacy, safety, cybersecurity, and oversight.
- Control the deployment, continuous monitoring, maintenance, and withdrawal of agentic AI systems.
- Identify the roles of external entities with regard to data protection.
- Anticipate new cases that may arise in relation to rights of access, rectification, erasure, restriction of processing, portability, and objection, and respond to these rights in a timely and appropriate manner.
- Integrate with incident management processes and compliance with obligations relating to personal data breaches.
- Adapt training plans.
- Maintain continuous monitoring, supervision, and auditing of the processing that incorporates agentic AI services with procedures for

Clear response and accountability for acting on deviations, incidents, or regulatory non-compliance.

- Agile information channels on updates to agentic AI systems, their use in processing, alternatives, and incidents involving governance roles and, to the extent necessary, users.

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

B. CONTINUOUS EVIDENCE-BASED EVALUATION OF THE AGENT

As processing operations become automated, the supervision of these processes must be automated to the same extent, or even more so, with regard to compliance with policies and measures adopted.

Audit automation should include all measures selected by the organization in relation to management for compliance with data protection regulations. This process requires a structured approach and covers both agentic AI systems as a whole and within the framework of processing, as well as the individual evaluation of each of the components and services that comprise it.

This could include reviewing the functionalities of each component in the event of changes to these elements, for regulatory compliance and risk mitigation purposes. Evaluation methods may include *benchmark testing*, *human-in-the-loop* assessments, 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 services being evaluated and in the agentic AI systems that incorporate them.

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

Functionality criteria must make it possible to identify when the agentic AI system and its components are behaving correctly or incorrectly, and provide objective measures that can serve as benchmarks. In particular, criteria and metrics for transparency, reproducibility, control, compliance, and traceability.

- ***Golden testing practices***

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

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

It allows for repeatable verification and evaluation of deviations in the event of changes in legal terms and system functionality, as well as increasing the explainability and transparency of the system in well-defined contexts.

- ***Contracts and other legal links***

The reality of contracting services on the Internet is that, in many cases, contracts do not comply with local legal regulations, the terms of contracts are changed unilaterally, and even the subject matter of the contract is altered without prior notice (version changes, discontinuity of functionalities, etc.). Furthermore, the dynamic nature of any digital service and application requires a review each time there is an update to the terms and contracts, as well as to the technical aspects of the services themselves, in order to determine how to comply with the GDPR.

Therefore, for those components or services that have an impact on data protection, the data controller must assess the conditions both at the time of design decisions and dynamically or automatically during the life cycle, in order to determine the legal changes to the components and assess their suitability.

From there, decisions can be made on how to implement each type of agentic AI and for each processing operation.

- ***Apply the precautionary principle.***

When deploying agentic AI solutions, an "incremental approach" can be adopted, for example, by gradually incorporating treatments, starting with those with the lowest risk, in limited cases, etc. It is also possible to opt for the use of agentic AI systems that have already been tested in similar organizations and to consult incidents and problems that have already arisen in the organization's environment in order to anticipate them and take appropriate measures.

The precautionary principle can also be applied at the level of agentic AI operation, such as the activation of 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 perform specific explainability audits, both of the model used and of the joint operation of the agents or agentic AI.

Explainability can be achieved through "white box" analysis (analysis of orchestrator code, data flow verification, etc.) and through "black box" testing, so it is closely related to "*Golden tests*."

- ***Human intervention***

If human supervision has an impact on processing, it will be necessary to implement regular audits of the effectiveness of such supervision.

C. Data minimization

The principle of minimization seeks to limit the processing of personal data to what is strictly necessary, and this can be achieved when agents are properly designed and configured so that, by default, they do not attempt to be effective simply by "brute force" data volume.

- ***Definition of policies for accessing the organization's information***

For each processing operation in which 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. In other words, an information policy incorporating the "*need to know*" principle must be implemented for agentic AI.

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

- ***Data cataloging and cataloging***

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

This identification can be at the dataset level (e.g., files, emails) or at the data field level (e.g., email recipients). Identification is performed by adding data (metadata) to the original data.

Therefore, cataloging 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, cataloging must characterize the quality of the stored information (accuracy, relevance, age, scope, biases, regulatory constraints on use, objective context, etc.).

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

- ***Cataloging unstructured sources***

Unstructured sources represent a high percentage of data in an entity (e.g., emails, meeting minutes and recordings, reports, etc.) and are characterized by their lack of a fixed format, which complicates their indexing, scalability, and searchability, in addition to requiring significant resources for large volumes.

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

From there, the information could be pre-processed to extract specific data for the agents' tasks. In particular, anonymization or removal of personal data that is not necessary.

- ***Minimization granularity***

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

For example, in the context of artificial intelligence, in an automatic response to email messages, applying minimization in the operation of syntactic review of an email message before it is sent means not processing the recipient's name, and in the operation of sending the email, it means processing that name but not analyzing the content of the message[42].

Minimization should focus on both general subject data and information about the users of the systems themselves. In particular:

- A design that avoids personal profiling of users
- Elimination of metadata that is not useful in the pipeline or *pipeline* phases.
- Disassociation of user actions when they are not necessary.

- ***Filtering of data flows***

In relation to the above, the analysis can not only be performed on data at rest, but could also be considered for data in transit between the agent's different actions when they involve communication of data with third parties and external parties. In other words, in the intermediate stages of the reasoning chains, filtering of the information exchanged could be incorporated for categorization, sanitization, minimization, and alerting in data exchanges.

This would not only prevent the detection of, for example, internally generated *prompt* injections, but also the exposure of personal data, excessive use of data, mass access to information, etc. Another case would be to determine whether the agents' operations involve the processing of special categories of data that are not necessary for the processing.

[42] In classified information processing, this is known as the "need-to-know" principle, which applies to each of the parties involved in the processing.

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

- ***Shadow leaks***

In order to minimize "*shadow* leaks"[43], measures should be implemented, such as the use of *data loss prevention* (DLP) tools aimed at minimizing exposure of the internal context of the system, limiting the disclosure of explanations relating to reasoning or operating rules, applying correlation controls between queries to detect improper relationships, using generic responses to meta-questions[44] and continuously monitoring long-term query patterns in order to identify anomalous behavior or potential risks.

- ***Pseudonymization of users***

Pseudonymize user interaction with the agent, so that single-use *tokens* are used for interaction between components or for access to external services when authentication is required. This will, among other things, prevent the control and profiling of users (see next section) and prevent the AI agent from granting effective consent 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 may even constitute a profile of them. To this end, the following could be considered:

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

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

[44] Queries that do not directly seek functional information about the system, but rather attempt to gain knowledge about its internal workings, rules, sources, reasoning mechanisms, limits, or 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 offers.

D. Memory control

Control of the memory of the agentic AI system is closely related to data minimization strategies, guarantees of explainability and repeatability of inferences or profiling of individuals, and traceability capabilities for applying consent management, rights exercises, and processing limitations.

Control of the agent's memory must be exercised over both short-term and long-term memory.

- ***Memory management***

Introduce the ability to access, catalog, and manage memory content, allowing, for example, searches by content and quality parameters, deletion, setting processing limitations or usage alerts, including access traceability, auditing, 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 should be considered.

The level of granularity of compartmentalization will depend on the processing, clearly defining which memory will be commonly used by any agentic AI operation in the organization as it implements its policies, and which data and information will need to be separated between processing, users, and different cases. The rigidity of this compartmentalization, from a physical division, a rigid logical division, or a cataloging search, will depend on the processing and the 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, consistency 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 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 operation and could be, for example, on the division into subtasks, access to certain tools, or final decisions.

- ***No selective log policy***

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

This policy means that the information recorded in the component is minimal and only relates 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 may be recorded at the processing log level, for each processing independently, and in accordance with the data controller's information policies.

- ***Establishment of strict retention periods***

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

- ***Deactivation of memory storage***

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

- ***Apply memory sanitization strategies***

Apply long-term memory sanitization or cleansing techniques by automatically checking for harmful content, expiring unused or obsolete entries, analyzing information consistency, searching for and removing unnecessary user credentials, distilling information, analyzing and removing biases, and implementing strategies to force the user/administrator to perform periodic cleanups.

E. AUTOMATION

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

The degree of autonomy that the agentic AI system may have must be established by the person responsible for each of the processing operations, taking into account the context, scope, purposes, and risk to the rights and freedoms of individuals, as well as regulatory compliance in relation to automated decisions. The decision must be appropriately justified based on evidence and documented.

[45] This is different from sending a *prompt* saying that the information that has been stored should not be taken into account.
[46] In English, the concept is defined as "*sanitization*."

- ***Effective and secure design of chains of reasoning***

The design of reasoning chains must be controlled and validated. If the reasoning chain is developed using LLMs, it is necessary to assess their capacity to achieve the level of quality required to address the contexts of the treatments in which agentic AI will be used. In addition, it must be ensured that there is no possibility of contamination between different incompatible learned models (e.g., subtasks of administrative procedures from different jurisdictions) in the development of the reasoning chain.

Where appropriate, assess the need to implement reasoning chains, either in full or at a higher level of abstraction, in *a hardcoded* manner by the administrator. For example, manually divide a processing operation into subtasks[47]and let the reasoning agents work out the details of those subtasks.

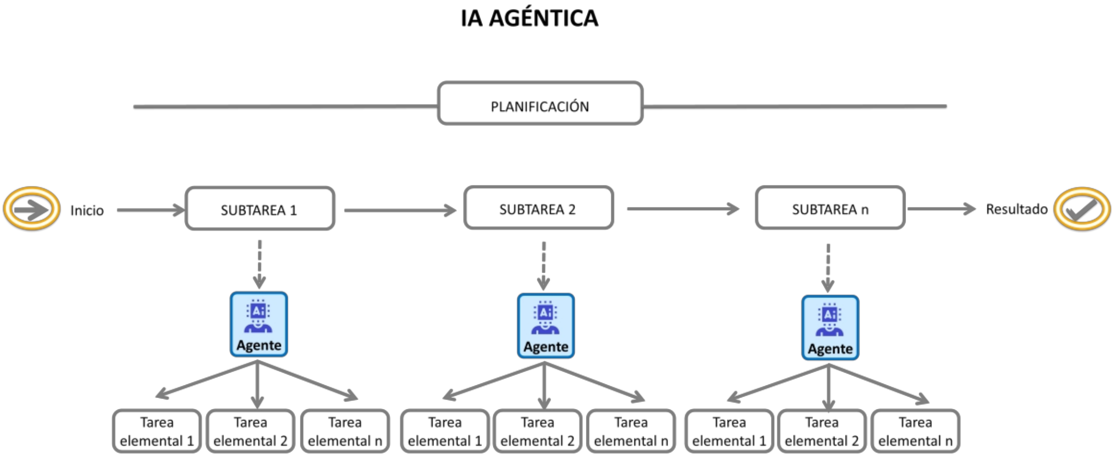


Figure 14 Example of two levels of task decomposition

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

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

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

[48] Security mechanisms and restrictions that guide the behavior of AI models to prevent harmful, biased, or inappropriate outputs, similar to barriers on a highway that prevent vehicles from veering off course.

- ***Catalog and whitelists of services***

This involves having a catalog of services, which may include different LLMs, identifying versions 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 treatments (for developers, if function calls are predefined, or to limit the tools that can be invoked by the LLM otherwise). The catalog should cover external services but also internal ones, for example, data repositories or services that allow access to the operator's screen.

- ***Limitation of accessible services***

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

- ***Control over the execution of tools***

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

- Control of the parameters used to invoke tools, implementing *guardrails* and rigid formats to detect erroneous or biased parameters.
- Control of tool responses, implementing new *guardrails* on input content.
- Requiring human oversight for certain tools that could have a greater impact.

- ***Criteria and control points for human intervention***

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

- High-impact actions and decisions, such as editing sensitive data, final decisions in high-risk areas (such as healthcare or the legal field), 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, such as permanently deleting data, sending communications, or making payments.

- Atypical or unusual behavior, such as when an agent accesses a system or database outside their scope of work, or when they select a delivery route that doubles the average distance.
- User-defined. Agents can act on behalf of users with different levels of risk tolerance. In addition to the limits defined by the organization, users can be offered the option to set their own limits, for example, requiring approval for purchases exceeding a certain amount.

- ***Reversibility of AI agent actions***

Assess the need to implement measures that allow certain actions to be reversed, for example, if the agent can modify personal data.

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

Include adjustable autonomy controls for each treatment using agentic AI, based on impact or risk, ranging from autonomous execution in low-impact and low-risk treatments to mandatory human intervention with high granularity in treatment operations when they are high.

- ***Effective human oversight***

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

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

- Competence and authority: you have the authority or assigned task 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 treatment and the automated system used, in terms of its capabilities and limitations.
- Independence: assess whether there are pressures from within or outside the organization that influence the person's challenge of the decision.
- Diligence in the exercise of their competence: in particular, whether they are subject to automation bias.
- Means to exercise their competence and qualifications.
    - That the procedures and technical means exist to intervene at the appropriate time and in the appropriate manner.
    - That it has the necessary information in a timely manner to be able to exercise its qualification, in particular, to know the consequences and risks of decisions in general, and those being made for specific cases and all aspects that condition the automated decision.

These include the specific individual's data, but could also include the procedures for collecting input data, the data implicit in the model that generates the decision, contextual data that has not been taken into account in the automated decision, as well as the capabilities and limitations of the decision-making system. They also include data that the person, in their capacity as a qualified professional, considers necessary to take into account for the specific case and that has not been considered in the automated decision.

- Have the resources to exercise their expertise: agentic AI decisions must be explainable, for example, applications that allow them to analyze the information in the format being used for the automated decision, etc.
- Have the necessary time to exercise their expertise for each of the decisions within their competence.

- ***Escalation routes***

Human oversight could be complemented by automated real-time monitoring to escalate any unexpected or anomalous behavior. Escalation involves implementing protocols and techniques to transfer control of automated processes to a human operator when high-risk situations, uncertainty, or anomalies are detected.

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

- ***Four-eyes principle***

In cases of automated processes with a significant impact on people's rights and freedoms, the principle of double verification by different people may be applied, providing an additional layer of trust in the human oversight mechanism and promoting critical awareness on the part of the operator.

F. AGENT CONTROL FROM THE DESIGN STAGE

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

- ***Documentation***

Maintain a dynamically updated record with integrity control (not necessarily on paper) of the process of responsibilities, decisions, actions taken, designs, architecture, operational events, and evolution.

- ***Qualified professionals***

Use a team of qualified professionals to deploy agentic AI systems in treatment; it is not just a matter of implementing an AI agent, but also of taking into account the implications of automating organizational processes, which requires personnel with knowledge of data science, process quality, operational context, security, and regulatory compliance, among other areas.

- ***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 it was transformed, 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 Lineage*."

In this sense, the more complex the data lifecycle and the more parties involved in it, the more valuable it is to incorporate traceability into the processing.

Traceability can serve purposes other than data protection, such as controlling trade secrets, intellectual and industrial property, and improving contracts. On the other hand, it can fulfill the following objectives from the point of view of the GDPR:

- Comply with the GDPR's transparency requirements for data subjects.
- Enable the effective exercise of data subjects' rights, in particular the management of consent.
- Enable the controller to fulfill their obligations (e.g., to ensure the principles of processing limitation, purposes consistent with the legal basis, or control of processors/subprocessors).
- Have evidence that data is processed in each processing operation performed in agentic AI, in its intermediate phases. In particular, if special categories of data are being used.
- Controls over employees involved in processing, now as users of agentic AI, to prevent abuse and bias.
- Demonstrate diligence and transparency to data subjects and supervisory authorities.

Therefore, measures to ensure these capabilities are related to data cataloging and involve keeping logs of the information processed by all reasoning processes, the sources accessed, and the services used in the inference, both input and output. In particular, it means having detailed control over the data and purposes for which external services access information.

This is particularly relevant both for transparency in data processing and for purposes of analyzing the reproducibility of inferences, controlling the information that is processed by services about users, monitoring regulatory compliance, implementing information policies, etc.

- ***Verification and validation testing***

Although verification and validation tests are well-known techniques in systems engineering, and not specific to artificial intelligence systems, it is important to remember that they exist and remain 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 consists of checking whether the system is being built correctly, i.e., whether it complies with requirements, design specifications, and standards using static techniques such as reviews, inspections, and code analysis (for example, checking that internal and external data flows are actually as stated). Validation checks that the actual needs of the user in a given context are met in relation to established quality metrics, using 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 constructing *prompts*. This involves defining a structured set of step-by-step instructions detailing how an AI agent should act within the treatment framework to achieve consistent and more predictable results and avoid malicious prompts.

For example, *prompts* could be structured as follows: initial interpretation, classification, validation criteria, preliminary breakdown of the problem, selection of tools, criteria for information search, cross-checking, data cleaning, evaluation, etc. All of this with predefined fields.

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

- ***Repeatability mechanisms***

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

of reasoning, as well as pseudo-randomness configurations in "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 management, authentication, and privileges***

The management of users' digital identities, agentic AI, and its components is a traceability and auditing tool. It also enables the necessary management to prevent unauthorized escalation of agent privileges, identity theft, and access control violations.

The basic principle to be applied in the agentic AI environment is that of least privilege, and the following strategies should be implemented:

- Implement secure authentication mechanisms for both users and agentic AI and its components. For example, require cryptographic identity verification for agents, implement granular RBAC and ABAC, multi-factor authentication (MFA) for accounts with high privileges, enforce continuous re-authentication in long sessions, prevent privilege delegation between agents except as authorized in predefined flows, mutual authentication in AI-to-AI interactions, limit the persistence of credentials or the temporality of agent credentials, etc.-AI and between agents, limiting the persistence or temporality of agent credentials, etc.
- Restrict privilege escalation and identity inheritance: for example, use dynamic access controls that expire elevated permissions, develop AI-based behavior profiles to detect inconsistencies in role assignment and agent access patterns, require human validation for high-risk AI actions involving changes in authentication, detect role inheritance anomalies in real time, apply time restrictions to privilege escalation, etc.
- Detect and block AI impersonation attempts: for example, detect inconsistencies in identity verification, monitor unexpected role changes, detect, log, and alert suspicious deviations in authentication attempts or failed attempts, as well as cascading or recursive execution patterns of tools activated between agents, isolate agents that generate suspicious protocol traffic.

- ***Strict control over updates.***

Having control over which updates are made to each element of the agentic AI system and decision-making power over when those updates go into production to avoid incompatibilities, instabilities, and lack of robustness, new treatments, changes in functionality, the emergence of new vulnerabilities, legal changes with regulatory non-compliance, uncontrolled international transfers, etc.

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

One prevention mechanism is to include version control systems with the possibility of rollback.

In the case of updates, consider the use of *sandboxing* (see section below) and continuous evaluation (previous sections).

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

With regard to the ability to perceive and act on the environment, measures can be used to restrict the breadth of the external context with which the agent interacts.

In its most restrictive form, the application of the *sandboxing* principle would involve the implementation of *Secure Processing Environments* (SPE[51]). In its most lenient form, we would find an implementation without restrictions on the permissions of the agentic AI system to interact with the environment. The use of sandboxing in the execution of tools invoked by agentic AI is a common intermediate application. Depending on compliance obligations, impacts, or risks, an architecture between the two extremes should be established.

One possible implementation of *sandboxing* is to use confined environments, such as containers or microVMs, to isolate agent execution. Another is to use 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 testing phases.

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

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

- ***Data extraction flow control***

Introduce controls that require explicit user action for data communications to third parties or mass data transmissions in those treatments where this is necessary.

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

[51] The Data Governance Regulation defines secure processing environments in Article 2.20) as "secure processing environment" means the physical or virtual environment and organizational 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, download, and export of data, as well as the calculation of derived data using computational algorithms;2.

- ***Circuit breakers and hard limits on steps***

Circuit breakers in agent-based AI are programmed safety mechanisms that automatically interrupt the execution of an agent when they detect predefined anomalies, such as infinite loops, deviations from objectives, massive data access, attempts at massive information exchanges, deviation from objectives, etc.

- ***Calibration and alignment controls***

Calibration problems, insofar as they may affect the processing of personal data (excessive, sensitive, inaccurate, without legitimacy), can be avoided by introducing measures between the intermediate stages of the reasoning chains that evaluate actions and data in relation to quality parameters, alignment with policies and regulations, and business interests. These measures could be implemented based 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. Human oversight could also be included among the possible measures of this type.

G. Consent management

In the case of consent-based processing, the data subject must also be able to give, modify, or withdraw their consent within the framework of a complex chain of reasoning, which may include multiple repositories and data sources from multiple entities.

H. Depending on the complexity of the processing, its impact, and its risks, management could take different forms, closely related to what is set out in sections "VII.D. Memory control

The control of the memory of the agentic AI system is closely related to data minimization strategies, guarantees of explainability and repeatability of inferences or profiling of individuals, and the capacity for traceability to apply consent management, rights exercises, and processing limitations.

The agent's memory must be controlled in terms of both short-term and long-term memory.

- ***Memory management***

Introduce the ability to access, catalog, and manage memory content, allowing, for example, searches by content and quality parameters, deletion, setting processing limitations or usage alerts, including access traceability, auditing, etc.

- ***Memory compartmentalization***

In the case of the same agentic AI in the organization, the opportunity to have memory compartmentalized and managed for different purposes should be considered.

treatments, different cases within treatments, and/or for different users.

The level of granularity of the compartmentalization will depend on the processing, clearly defining which memory will be commonly used by any agentic AI operation in the organization as it implements its policies, and which data and information will need to be separated between processing, users, and different cases. The rigidity of this compartmentalization, from a physical division, a rigid logical division, or a cataloging search, will depend on the processing and the 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, consistency between different actions of the organization, or the appearance of biases.

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

- ***Selective log policy***

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

This policy means that the information recorded in the component is minimal and only relates 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 may be recorded at the processing log level, for each processing independently, and in accordance with the data controller's information policies.

- ***Establishment of strict retention periods***

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

- ***Deactivation of memory storage***

In certain processing operations and according to their needs, allow the deactivation of persistent memory by default or its deactivation by the user. The granularity of the deactivation may be at the level of subtasks that can be considered high risk in order to avoid the storage of personal data irrelevant to future processing or to prevent the persistence of malicious injections.

deactivation may be granular at the level of subtasks that can be considered high risk in order to avoid the storage of personal data that is irrelevant for future processing or to prevent the persistence of malicious injections.

- ***Apply memory sanitation strategies***

Apply long-term memory cleansing or purging techniques by automatically checking for harmful content, expiring unused or obsolete entries, analyzing information consistency, searching for and removing unnecessary user credentials, distilling information, analyzing and removing biases, and implementing strategies to force the user/administrator to perform periodic cleanups.

I. AUTOMATION

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

The degree of autonomy that the agentic AI system may have must be established by the person responsible for each of the processing operations, taking into account the context, scope, purposes, and risk to the rights and freedoms of individuals, as well as regulatory compliance in relation to automated decisions. The decision must be appropriately justified based on evidence and documented.

- ***Effective and secure design of reasoning chains***

The design of reasoning chains must be controlled and validated. If the reasoning chain is developed using LLMs, it is necessary to assess their capacity to meet the quality standards required to address the contexts of the treatments in which agentic AI will be used. In addition, it must be ensured that there is no possibility of contamination between different incompatible learned models (e.g., sub-tasks of administrative procedures from different jurisdictions) in the development of the reasoning chain.

Where appropriate, assess the need to implement chains of reasoning, either in full or at a higher level of abstraction, *hardcoded* by the administrator. For example, manually divide a task into subtasks and let the reasoning agents work out the details of those subtasks.

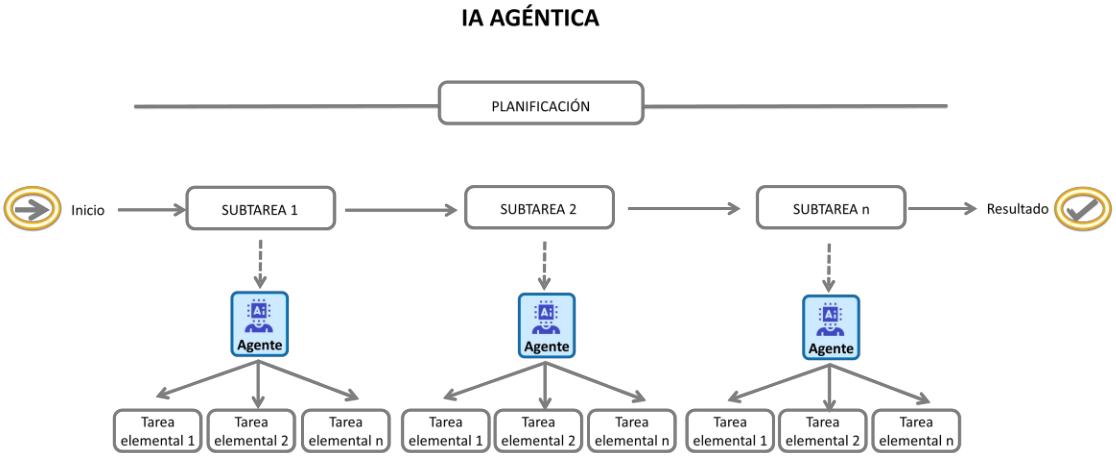


Figure 14 Example of two levels of task decomposition

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

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

- ***Catalog and whitelists of services***

This involves having a catalog of services, which may include different LLMs, identifying versions 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 treatments (for developers, if function calls are predefined, or to limit the tools that can be invoked by the LLM otherwise). The catalog should cover external services but also internal ones, for example, 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 catalog for specific treatments. In this way, each treatment 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 a priori in the form of a RAG.

▪ *Control over the execution of tools*

The invocation of tools and services on the Internet are *de facto* partial outputs of agentic AI that can be transparent to users. The controls that could be established over 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.
- Requirement for human oversight in relation to certain tools that could have a greater impact.

▪ *Criteria and control points for human intervention*

Significant criteria and control points or limits of action requiring human approval should be defined at the design stage, especially before sensitive actions are executed. This may include:

- High-impact actions and decisions, such as editing sensitive data, final decisions in high-risk areas (such as healthcare or the legal field), or actions that could lead 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, such as permanently deleting data, sending communications, or making payments.
- Atypical or unusual behavior, such as when an agent accesses a system or database outside their scope of work, or when they select a delivery route that doubles the average distance.
- User-defined. Agents can act on behalf of users with different levels of risk tolerance. In addition to the limits defined by the organization, users can be offered the option to set their own limits, for example, requiring approval for purchases exceeding a certain amount.

▪ *Reversibility of AI agent actions*

Assess the need to implement measures that allow certain actions to be reversed, for example, if the agent can modify personal data.

▪ *Level of autonomy according to treatment*

Include adjustable autonomy controls for each treatment using agentic AI, based on impact or risk, ranging from autonomous execution in low-impact and low-risk treatments to mandatory human intervention with high granularity in treatment operations when they are high.

- ***Effective human oversight***

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

When evaluating human intervention, it is advisable to take into account:

- Competence and authority: you have the authority or assigned task that allows you to alter the outcome of the automated decision.
- Preparation and training: do you have 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 terms of its capabilities and limitations?
- Independence: assess whether there are pressures from within or outside the organization that influence the person's challenge to the decision.
- Diligence in the exercise of their competence: in particular, whether they are subject to automation bias.
- Means to exercise their competence and qualifications.
  - That the procedures and technical means exist to intervene at the appropriate time and in the appropriate manner.
  - That they have the necessary information in a timely manner to be able to exercise their qualification, in particular, to know the consequences and risks of decisions in general, and those being made for specific cases, and all aspects that condition the automated decision. These include the data of the specific individual, but could also include the procedures for collecting input data, the data implicit in the model that generates the decision, the contextual data that has not been taken into account in the automated decision, as well as the capabilities and limitations of the decision-making system. Also included is any data that the person, in their capacity, deems necessary to consider for the specific case and that has not been considered in the automated decision.
  - That it has the resources to exercise its qualification: the decisions of agentic AI must be explainable, for example, applications that allow it to analyze the information in the format being used for the automated decision, etc.
  - It must have the necessary time to exercise its qualification for each of the decisions within its competence.

- ***Escalation routes***

Human supervision could be complemented by automated real-time monitoring to escalate any unexpected or anomalous behavior. Escalation involves the implementation of protocols and techniques to transfer control of automated processes to a human operator when high-risk situations, uncertainty, or anomalies are detected.

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

- ***Four-eyes principle***

In cases of automated processes with a significant impact on people's rights and freedoms, the principle of double verification by different people may be applied, providing an additional layer of trust in the human supervision mechanism and promoting critical awareness on the part of the operator.

" and☒"Traceability."

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

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

In some processing operations, the use of both "white" and "black" lists could be considered, allowing the precise definition of subjects' preferences regarding certain processing operations.

J. TRANSPARENCY

The GDPR establishes minimum transparency measures that are mandatory.

However, to pass a proportionality test or reduce risk, additional measures can be implemented. In order to demonstrate to the subject that they can trust the processing operations to the AI agent system (as a user, employee, customer, etc.), measures such as the following could be adopted: real-time information on data flow, information on which data subjects' data are in the repositories or third-party services that are processing the data, access to records of processing activities and data communications, information on intermediate events in the reasoning chain, context used in the result, human intervention performed, possibility of requesting review or human action, access to certifications, audits, or EIPDs of processing.

K. LITERACY

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

- Management level, in the knowledge necessary to make appropriate evidence-based decisions regarding the inclusion of AI agents in treatments.
- Level of ICT managers responsible for the development, procurement, deployment, operation, maintenance, and withdrawal of such systems, so that, in particular, the implications of data protection and the techniques and organizational measures to implement them are understood and identified.
- Level of users with different roles in processing operations involving agentic AI systems, with knowledge of the possibilities, implications, and limitations of these tools.

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

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

VIII. FINAL THOUGHTS

AI systems, such as agentic AI, are here to stay. Pretending to ignore their existence, both from a competitive organizational standpoint and in terms of regulatory authorities, would mean missing out on strategic opportunities.

Understanding this technology is necessary in order to make rational, evidence-based decisions about its implementation. Understanding a technology involves more than just becoming a user; it means understanding its fundamentals, its implications, its limitations, and how it is implemented. Both the irrational rejection of all the advantages offered by agentic AI and the leap of faith to uncritically accept any type of implementation in personal data processing could be harmful.

In particular, with objective analysis, the chosen implementation of agentic AI allows for more than just ensuring data protection, i.e., only a reactive approach to threats and vulnerabilities. An implementation of agentic AI that takes data protection into account from the design stage allows for the definition of agent-based processing that incorporates Privacy Enhancing Technologies (PETs) that offer superior guarantees (or could even enable) manual processing. In itself, 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 providers accessed by the organization.

In this regard, the involvement of the DPO and data protection advisors are key elements. To this end, DPOs must be knowledgeable about processing and process management principles, be able to understand the fundamentals of the tools being used, be aware of the various technical and organizational alternatives for implementing safeguards, and be able to envision the opportunities they may offer for the protection of rights. In addition, they must be involved in decisions regarding the design of processing operations and the AI agent systems selected to implement them, since the possible measures for managing compliance and data protection risk are related to the fulfillment of the entity's other objectives and obligations and must be addressed in an integrated manner.

In conclusion, we find ourselves faced with a technology that is rapidly evolving and requires analysis and experience, both in terms of its impact, its measures, and its opportunities for data protection. Therefore, please consider this text as an introductory study that does not claim to be exhaustive.

IX. REFERENCES

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

Guidelines of the Article 29 Data Protection Working Party on automated individual decision-making 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*. Published in NeurIPS: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Spanish Data Protection Agency (AEPD). Compliance with the GDPR for processing that incorporate Artificial (2020) https://www.aepd.es/documento/adecuacion-rgpd-ia.pdf

Spanish Data Protection Agency (AEPD). Risk management and impact assessment in personal data processing. (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 Audits of Processing Operations 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*. Published in ICLR 2023. https://arxiv.org/pdf/2210.03629

Spanish Data Protection Agency (AEPD). *Assessment of 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 modeling 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. (n.d.). *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* Singapore (2026) https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf