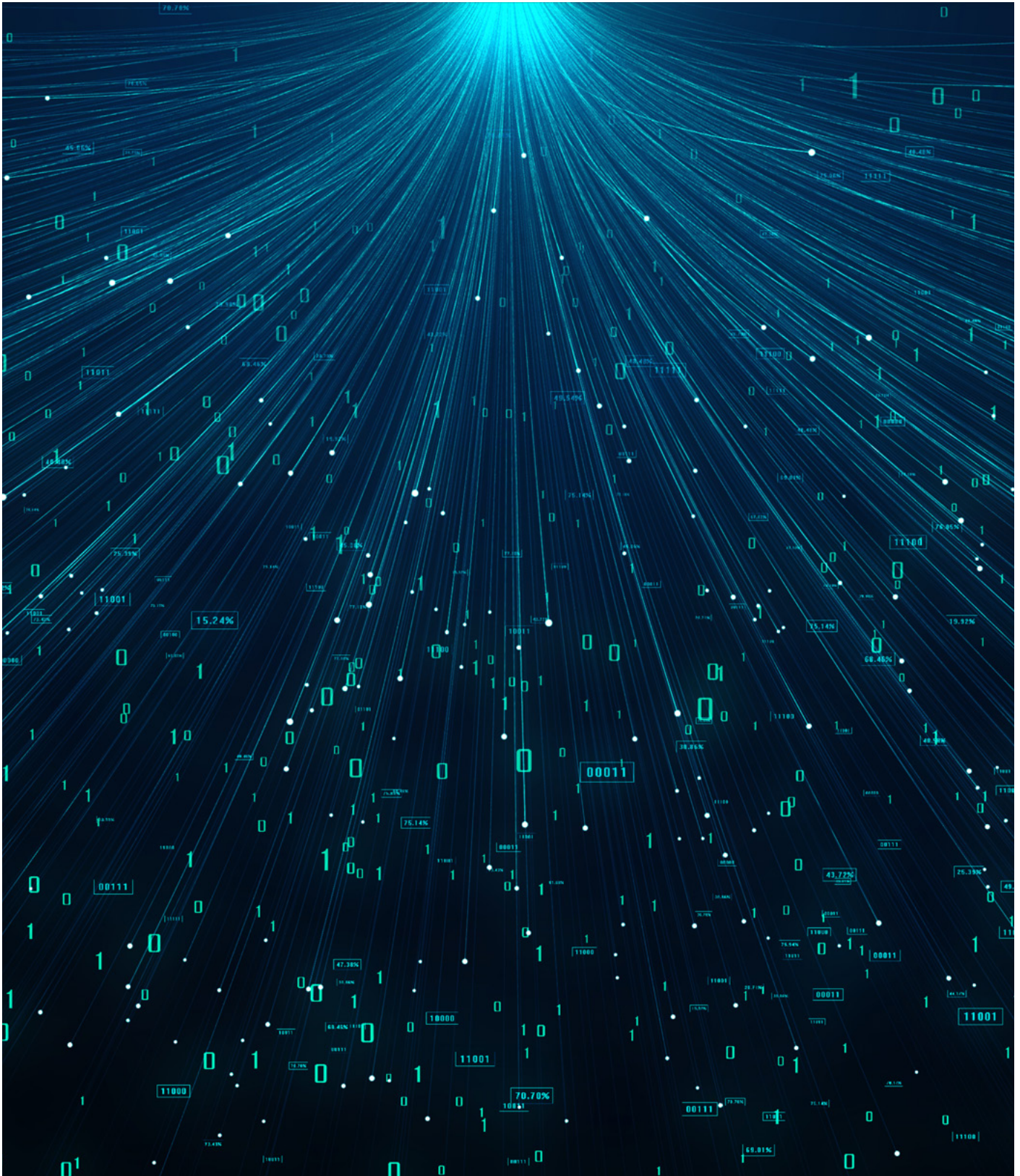


1

# Definition of an AI agent

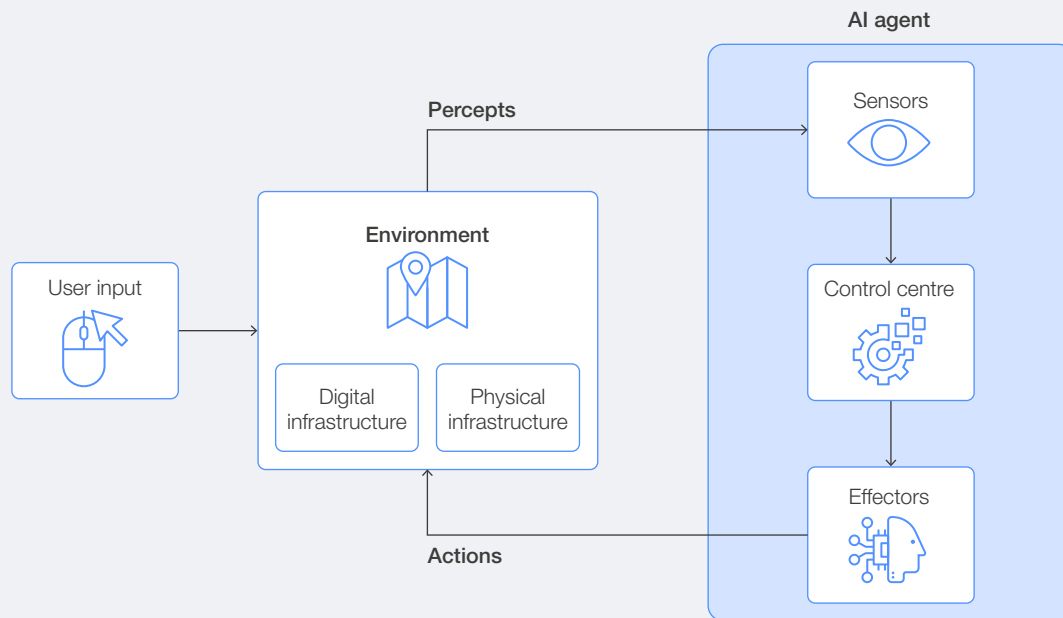
An AI agent responds autonomously to inputs and its reading of its environment to make complex decisions and change the environment.



Based on the definition of the International Organization for Standardization,<sup>5</sup> an AI agent can be broadly defined as an entity that senses **percepts** (sound, text, image, pressure etc.) using **sensors** and responds (using **effectors**) to its **environment**. AI agents generally have the autonomy (defined as the ability to operate

independently and make decisions without constant human intervention) and authority (defined as the granted permissions and access rights to perform specific actions within defined boundaries) to take actions to achieve a set of specified goals, thereby modifying their environment.

FIGURE 1: The core components of an AI agent



Source: World Economic Forum

Figure 1 highlights how an agent is made up of several core components, including:

- **User input:** the external (e.g. human, another agent) input that the AI agent receives. This could be instructions such as typing via a chat-based interface, voice-based commands or pre-recorded data.
- **Environment:** the bounds in which the AI agent operates. It serves as the area in which the agent applies its sensors and effectors to percept and modify its surroundings based on the inputs received and the actions decided upon by the control centre. The environment can be **physical infrastructure** such as the mapped area of an autonomous vehicle or **digital infrastructure** such as the intranet of a business for a coding agent.
- **Sensors:** mechanisms through which the agent perceives its environment. Sensors can range from physical devices (e.g. cameras or microphones) to digital ones (e.g. queries to databases or web services).
- **Control centre:** typically makes up the core of the AI agent along with the model, such as an LLM. The control centre helps process
- information, make decisions and plan actions. Based on the capabilities of the AI agent, the control centre involves complex algorithms and models that allow the agent to evaluate different options and choose the best course of action.
- **Percepts:** the data inputs that the AI agent receives about its environment, which could come from various sensors or other data sources. They represent the agents' perception or understanding of its environment.
- **Effectors:** the tools an agent uses to take actions upon its environment. In physical environments, effectors might include robotic arms or wheels, while in the digital environment, they could be commands sent to other software systems, such as generating a data visualization or executing a workflow.
- **Actions:** represent the alterations made by effectors. In physical environments, actions might be pushing an object, whereas in digital environments they could be linked to updating a database.



2

# The evolution of AI agents

Developers have transformed AI from rule-based systems to active agents capable of learning and adapting while engaged in a task.



The development of AI agents began in the 1950s,<sup>6</sup> and since then they have evolved from simple rule-based systems to sophisticated autonomous entities capable of complex decision-making. Early AI was characterized by deterministic behaviour, relying on fixed rules and logic that made these systems predictable but unable to learn or adapt from new experiences.

Advances in AI research introduced systems that could handle larger datasets and manage uncertainty, leading to probabilistic outcomes and non-deterministic behaviour. This shift enabled more flexible and dynamic decision-making, moving beyond rigid frameworks.

The 1990s marked a significant turning point, as machine learning applications became

more widespread. AI systems began to learn from data, adapt over time and improve performance. The introduction of neural networks during this period laid the foundation for deep learning, which has since become essential to modern AI.

Since 2017, the rise of LLMs has transformed AI's capabilities in natural language understanding and generation. These models use vast amounts of data to produce human-like text and engage in complex language-based tasks.

Today's AI agents use various learning techniques, including reinforcement learning, or transfer learning, allowing them to continuously refine their abilities, adapt to new environments and make more informed decisions.

## 2.1 Key technological trends

Over the past 25 years, the increase in computing capacity, the availability of large quantities of data on the internet and novel algorithmic breakthroughs have enabled significant developments in the base technologies behind recent advances in the capabilities of AI agents. These are briefly described below.

### Large models

Large language models (LLM) and large multimodal models (LMM) have revolutionized the capabilities of AI agents, particularly in natural language processing and the generation of text, image, audio and video.

The emergence of large models has been driven by several technological advances and by the transformer architecture, which has paved the way for a deeper understanding of context and word relationships, considerably improving the efficiency and performance of natural language processing tasks.<sup>7</sup> In summary, advanced AI models have enabled better understanding, generation and engagement with natural language.

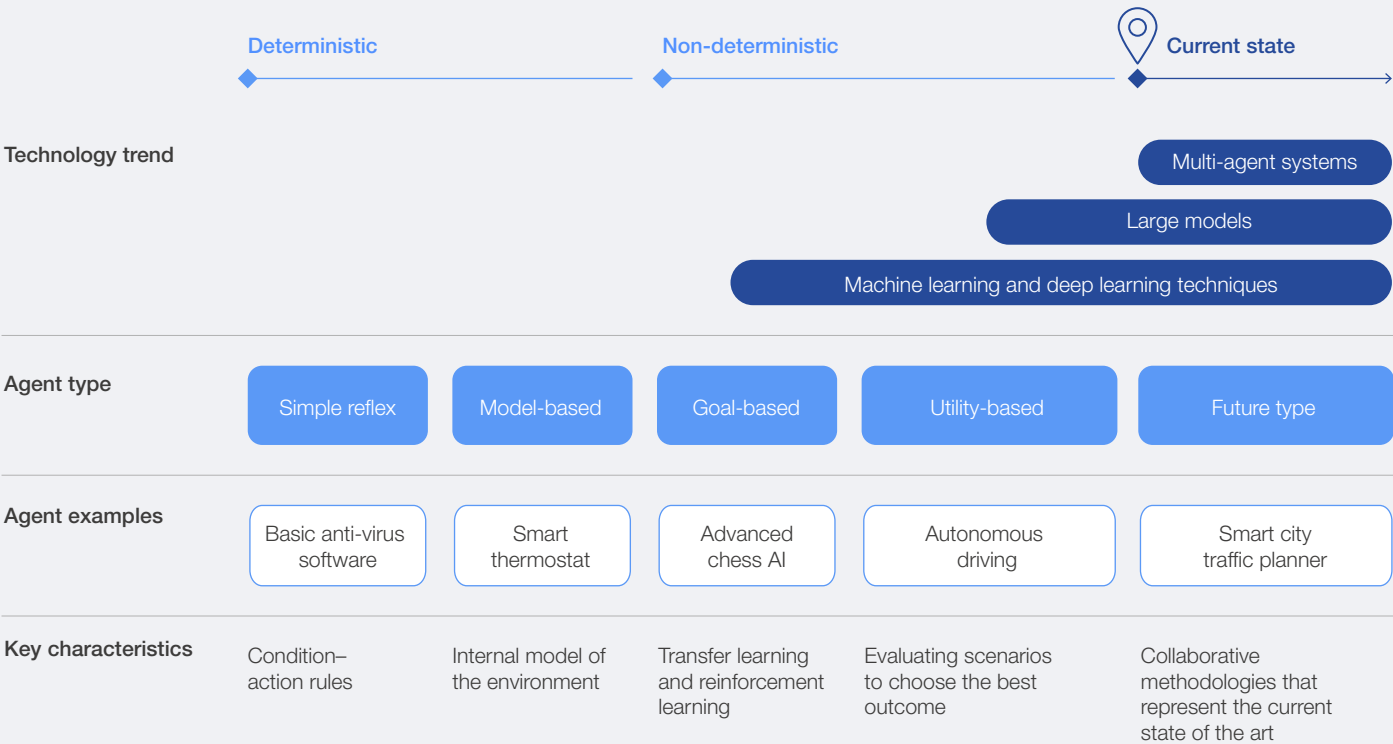
### Machine learning and deep learning techniques

A range of techniques have greatly improved AI models through increased efficiency and greater specialization. Some examples of machine- and deep-learning techniques include:

1. **Supervised learning:** facilitates learning from labelled datasets, so the model can accurately predict or classify new, previously unseen data.<sup>8</sup>
2. **Reinforcement learning:** enables agents to learn optimal behaviours through trial and error in dynamic environments. Agents can continuously update their knowledge base without needing periodic retraining.<sup>9</sup>
3. **Reinforcement learning with human feedback:** enables agents to adapt and improve through human feedback, specifically focusing on aligning AI behaviour with human values and preferences.<sup>10</sup>
4. **Transfer learning:** involves taking a pretrained model, typically trained on a large dataset (e.g. to recognize cars) and adapting it to a new but related problem (e.g. to recognize trucks).<sup>11</sup>
5. **Fine-tuning:** involves taking a pretrained model and further training it on a smaller, task-specific dataset. This process allows the model to retain its foundational knowledge while improving its performance on specialized tasks.<sup>12</sup>

These and other learning paradigms are often used in combination and have dramatically expanded the problem-solving capabilities of AI agents in various areas of application. The evolution of AI agents is detailed in Figure 2, while the agent types are further expanded in the following section.

FIGURE 2: Evolution of AI agents’ capabilities







Source: World Economic Forum

2.2 Types of AI agents

This section outlines different types of AI agent and traces their evolution, highlighting the key technological advances that have supported their development. AI agents can be considered as either deterministic or non-deterministic, based on their defining characteristics, which are outlined below.

TABLE 1: Defining characteristics of deterministic and non-deterministic AI agents

Deterministic AI agents	Non-deterministic AI agents
<b>Rule-based:</b> operate with fixed rules and logic, meaning the same input will always produce the same output.	<b>Data-driven and probabilistic:</b> make decisions based on statistical patterns in data, with outcomes that are not fixed but instead are probabilistic.
<b>Predictable behaviour:</b> the decision-making process is transparent and consistent, which makes the outcomes predictable.	<b>Flexible and adaptive:</b> able to learn from data, adapt to new situations and handle uncertainty, often resulting in varied outcomes for similar inputs.
<b>Limited adaptability:</b> these systems cannot learn from new data or adjust to changes; they follow only predefined paths.	<b>Complex decision-making:</b> use algorithms that factor in probabilities, randomness or other non-deterministic elements, allowing for more nuanced and complex behaviours.

Type	Definition	Examples
 <p><b>Simple reflex agents</b></p>	<p>Simple reflex agents operate based on a perception of their environment, without consideration of past experiences.<sup>13</sup> Instead, they follow predefined rules to map specific inputs to specific actions. The implementation of condition–action rules allows for rapid responses to environmental stimuli.</p> <p>These early agents are simple rule-based machines or algorithms designed to provide static information and unable to adapt or change course.</p>	<ul style="list-style-type: none"> <li>– Basic spam filters using keyword matching</li> <li>– Simple chatbots with predefined responses</li> <li>– Automated email responders that send prewritten replies following specific triggers</li> </ul>
 <p><b>Model-based reflex agents</b></p>	<p>Model-based reflex agents are designed to track parts of their environment that are not immediately visible to them.<sup>14</sup> They do this by using stored information from previous observations, allowing them to make decisions based on both current inputs and past experiences. By basing their actions on both current perceptions and their internal model, these agents are more adaptable than simple reflex agents even though they are also governed by condition–action rules.</p>	<ul style="list-style-type: none"> <li>– Smart thermostats that optimize energy usage by adjusting to current and historical temperature data, as well as user preferences</li> <li>– Smart robotic vacuum cleaners that use sensors and maps to navigate efficiently, avoiding obstacles and optimizing cleaning paths</li> <li>– Modern irrigation systems that use sensors to collect real-time data on environmental factors such as soil, moisture, temperature and precipitation, to optimize water dispensation</li> </ul>
 <p><b>Goal-based agents</b></p>	<p>Goal-based agents are able to take future scenarios into account. This type of agent considers the desirability of actions' outcomes and plans to achieve specific goals.<sup>15</sup> The integration of goal-oriented planning algorithms allows the agent to make decisions based on future outcomes, making them suitable for complex decision-making tasks.</p>	<ul style="list-style-type: none"> <li>– Advanced chess AI engines that have the goal of winning the game, planning moves that maximize the probability of success and considering a long-term strategy</li> <li>– Route optimization systems for logistics that set goals for efficient delivery and plan optimal routes by setting clear priorities</li> <li>– Customer service chatbots that set goals to resolve customer issues and plan conversation flows to achieve their goals efficiently</li> </ul>
 <p><b>Utility-based agents</b></p>	<p>Utility-based agents employ search and planning algorithms to tackle intricate tasks that lack a straightforward outcome, thereby going beyond simple goal achievement.</p> <p>They use utility functions to assign a weighted score to each potential state, facilitating optimal decision-making in scenarios with conflicting goals or uncertainty. Rooted in decision theory, this method allows for more advanced decision-making in complex environments. These agents can balance multiple, possibly conflicting objectives according to their relative significance.<sup>16</sup></p>	<ul style="list-style-type: none"> <li>– Autonomous driving systems that optimize safety, efficiency and comfort while evaluating trade-offs such as speed, fuel efficiency and passenger comfort</li> <li>– Portfolio management systems such as robot-advisers that make financial decisions based on utility functions that weigh risk, return and client preferences</li> <li>– Healthcare diagnosis assistants that analyse patient medical records, label patient data (e.g. tumour detection) and optimize treatment strategy recommendations in cooperation with doctors</li> </ul>

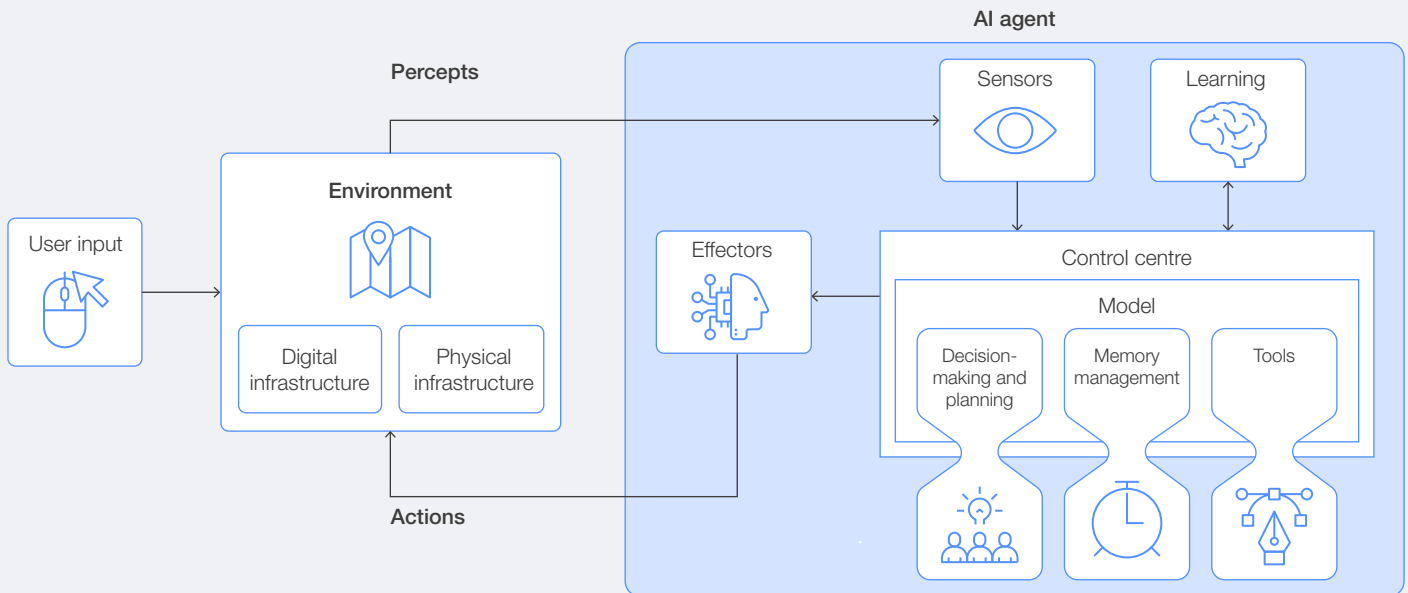


## 2.3 Advanced AI agents

The architecture of many current AI agents is often based on or linked to LLMs, which are configured in complex ways. Figure 3 presents a simplified

overview of the key components leading to current breakthroughs in AI agents and their growing range of capabilities.

FIGURE 3: Key components of advanced AI agents



Source: World Economic Forum

The AI agent begins with **user input**, which is directed to the agent's control centre. The user input could be a prompt given to carry out an instruction. The **control centre** directs the user input to the model, which forms the core algorithmic foundation of the AI agent. This model could be an LLM or an LMM, depending on the application's needs. The model then processes the input data from the user's instructions to generate the desired result.<sup>17</sup>

At the core of the architecture is the control centre, a crucial component that manages the flow of information and commands throughout the system. It acts as the orchestration layer, directing inputs to the model and routing the output to appropriate tools or effectors. In simple terms, this layer orchestrates the flow of information between 1) user inputs, 2) decision-making and planning, 3) memory management, 4) access to tools and 5) the effectors of the system enabling action in digital or physical environments.<sup>18</sup>

The decision-making and planning component of an AI agent uses the model's outputs to assist in decision-making and planning of multistep processes. In this segment, advanced features such as chain-of-thought (CoT) reasoning are implemented, which allows the AI agent to engage in multistep reasoning and planning. CoT

is a technique where an AI agent systematically processes and articulates intermediate steps to reach a conclusion, which enhances the agent's ability to solve complex problems in a transparent manner, as each step of the model's underlying reasoning is reproduced in natural language.<sup>19</sup>

**Memory management** is vital for the continuity and relevance of operations. This component ensures that the AI agent remembers previous interactions and maintains context. This is essential for tasks that require historical data to inform decisions or for maintaining conversational context in chatbots.

**Tools** enable the AI agent to access and interact with multiple functions or modalities. For example, in an online setting, an AI agent could have access to external tools such as web searches to gather real-time information and scheduling tools to manage appointments and send reminders, as well as project management software to track tasks and deadlines. In terms of modalities, an AI agent could use natural language processing tools alongside image recognition capabilities to perform tasks that require understanding of text-based as well as visual-based data sources.

Once decisions are made or plans set, the **effectors** component of the AI agent executes the required actions. This could involve interacting

with the physical world (in robotics), executing a software function or providing recommendations and decisions to human users.

The **learning** component is intrinsic to the model and enables the AI agent to improve its performance over time as the model gathers more input, using machine learning and deep learning techniques as mentioned in section 2.1.

The **application** layer surrounds the control centre, models and other components, acting as the interface between the AI agent and its environment. It interprets the outputs from the control centre and adapts them to specific tasks or

domains. For example, in a healthcare AI agent, the application layer would translate model outputs into diagnostics, treatment recommendations or medical alerts through an appropriate user interface.

In summary, when the varying components of an advanced AI agent come together, they represent the agent's ability to model the environment, maintain memory or knowledge storage with beliefs and preferences, as well as inherent abilities to learn, plan, make decisions, perceive (sense), act (interact) and communicate with the agent's surroundings.



### Example of an advanced AI agent: AI agent infotainment system

An AI agent in a car's infotainment system acts as a smart assistant, activated through voice commands to manage navigation, entertainment, climate controls and other vehicle settings. It processes live traffic, weather and driver preferences to optimize routes, suggesting alternatives around delays or hazards. The agent

personalizes entertainment based on user habits, recommends nearby stops such as restaurants or fuel stations and proactively provides updates such as low fuel alerts or optimal recharging points for electric vehicles – all while ensuring the driver remains focused on the road.

## 2.4 AI agent system

An AI agent system is an organized structure that integrates multiple heterogeneous (e.g. rule- and goal-based agents) or homogeneous (e.g. goal-based only) AI agents.<sup>20</sup> Each agent is typically specialized, possessing its own capabilities, knowledge and decision-making processes, while sharing data to collaboratively achieve the goal of the system.

Several designs are possible, such as:

- Mixture-of-agents, where each agent is called sequentially, with agents processing the outputs from each previous agent<sup>21</sup>

- Central orchestration, which coordinates calls of agents and manages the inputs and outputs accordingly

The AI agent system is designed to ensure that each agent contributes to the overall objective, whether it involves managing complex real-time processes such as autonomous driving, optimizing industrial processes or coordinating activities; for example, in smart city infrastructure. By dividing the workload among specialized agents, the system can handle dynamic environments and adapt to changing conditions, ensuring optimal performance.



### Example of an AI agent system: Autonomous vehicle AI agent system

A human user gets into an autonomous vehicle (AV). The AV is comprised of an AI agent system that includes agents for perception, path planning, localization for finding its specific place on the road and control to steer and brake.

The perception and localization agents are dedicated to continuously mapping the environment through sensors, the global positioning system (GPS) and cameras. The planning agent calculates the optimal trajectory by factoring in real-time traffic, weather and road conditions. The control agent

handles the vehicle's core mechanics, such as braking, accelerating and steering.<sup>22</sup> The AI agent infotainment system serves as the interface with the passenger, and handles elements such as processing voice commands and adjusting routes, climate, entertainment or other in-car settings based on user preferences.<sup>23</sup>

All agents work together in a coordinated and centralized manner to ensure the vehicle reaches its destination safely and efficiently, prioritizing both passenger comfort and safety.<sup>24</sup>



## 2.5 The future of AI agents: Towards multi-agent systems

Multi-agent systems (MAS) consist of multiple independent AI agents as well as AI agent systems that collaborate, compete or negotiate to achieve collective tasks and goals.<sup>25</sup> These agents can be autonomous entities, such as software programs or robots, each typically specialized with its own set of capabilities, knowledge and decision-making processes. This allows agents to perform tasks in parallel, communicate with one another and adapt to changes in complex environments.

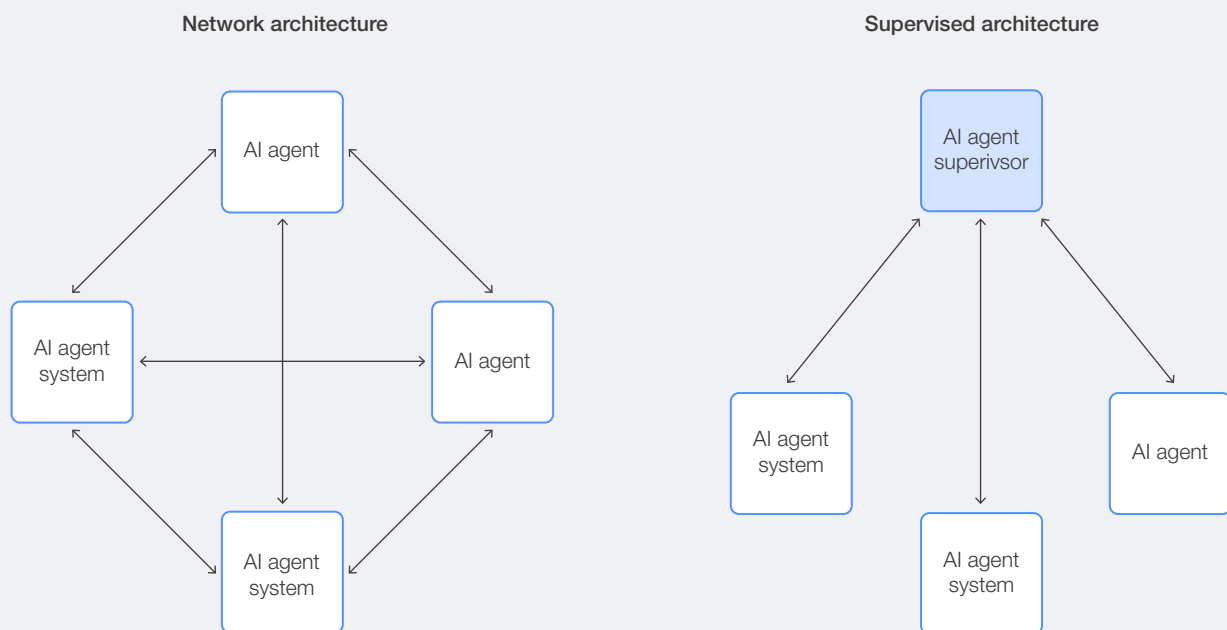
The architecture of a MAS is determined by the desired outcomes and the goals of each participating agent or system. There are several architectural types,<sup>26</sup> for example:

- **Network architecture:** In this set-up, all agents or systems can communicate with one another to reach a consensus that aligns

with the MAS's objectives. For example, when autonomous vehicles (AVs) park in a tight space, they communicate to avoid collision. In this case, the MAS objective to prevent accidents aligns with each AV's goal of safe navigation, allowing them to coordinate effectively and reach consensus.

- **Supervised architecture:** In this model, a "supervisor" agent coordinates interactions among other agents. It is useful when agents' goals diverge, and consensus may be unattainable. The supervisor can mediate and prioritize the MAS's objectives while considering each agent's unique goals, thereby finding a compromise. An example could be when a buyer and seller agent cannot reach agreement on a transaction, which is then mediated by an AI agent supervisor.

FIGURE 4: Examples of MAS architecture



Source: World Economic Forum

While current efforts largely focus on developing AI agents within closed environments or specific software ecosystems, the future is likely to see multiple agents collaborating in different domains and applications. In MAS, different types of agent could work together to tackle increasingly complex tasks that require multistep processes, integrating expertise from various fields to achieve more sophisticated outcomes.

These agents can communicate and interact within a broader adaptive system, enabling them to handle both specific tasks and complex situations more efficiently than a single agent, or even an AI agent system, could on its own.

In some cases, multi-agent systems address the limitations of single-agent systems, such as scalability issues, lack of resilience in the event of

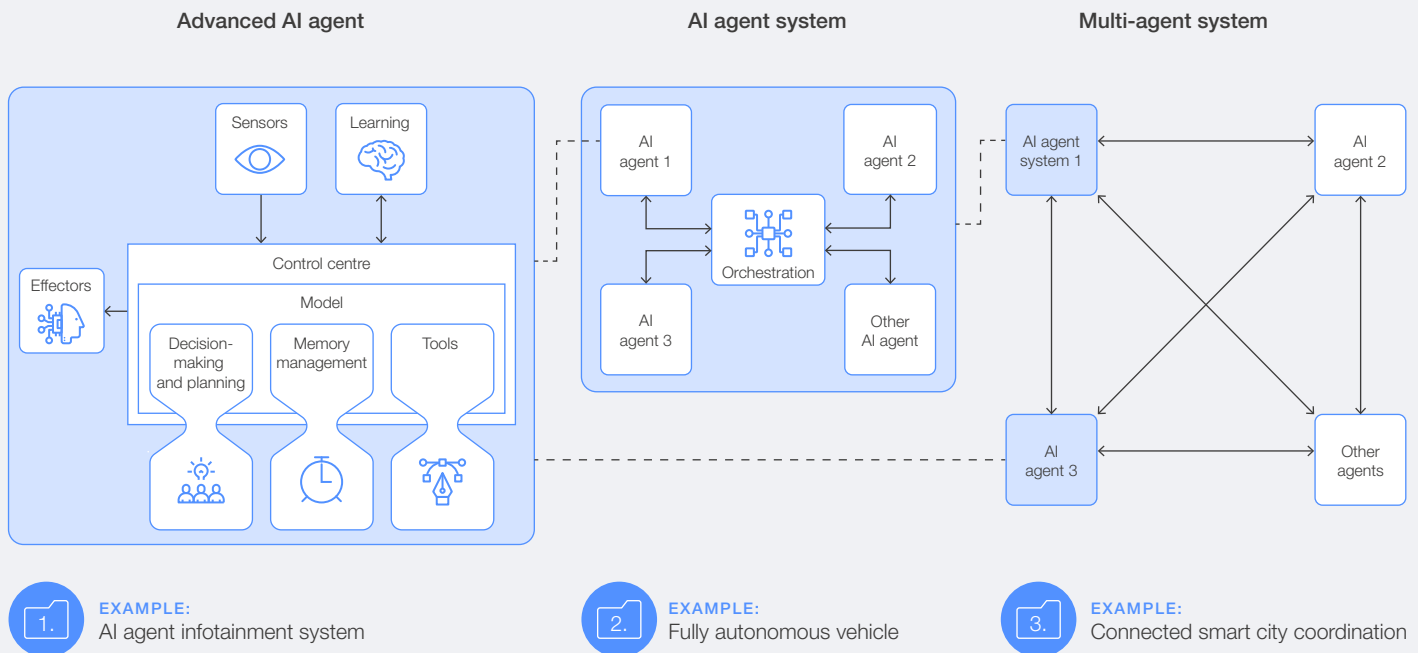
failure or errors and limited skill sets. By distributing tasks among multiple agents, MAS could increase both efficiency and capability.

In theory, multi-agent systems are highly adaptable, as agents can be dynamically added or removed, allowing the system to respond to changing environments and requirements. This scalability is

crucial for applications that need to grow or evolve over time without extensive re-engineering.

In many ways, multi-agent systems can be considered as a future type of system that could coordinate agent actions among multiple users or organizations through human-comprehensible language or to-be-determined AI agent protocols.

FIGURE 5: The structure and relationships among the AI agent, AI agent system and multi-agent system



Source: World Economic Forum



### Example of a multi-agent system: Smart city traffic management with vehicle-to-everything (V2X) communication

In a smart city, a multi-agent system (MAS) manages traffic flow in real time, using vehicle-to-everything (V2X) communication, enabling vehicles to interact with other vehicles, pedestrians and road infrastructure.<sup>27</sup> Each traffic signal is controlled by an AI agent system that communicates with nearby signals, public transport systems, emergency services and parking services to check availability. Vehicles, equipped with their own AI agent system, share data such as speed, location and road conditions, allowing for coordinated actions to enhance road safety, traffic efficiency and

energy usage. For example, if an accident occurs, AI agents can reroute traffic, adjust signal timings, notify emergency services and communicate with vehicles and pedestrians to avoid the area, all with minimal human intervention. This system optimizes traffic flow, improves road safety and reduces energy consumption by dynamically adapting to real-time conditions. For instance, if a parking lot is full, the system can direct vehicles to available parking further away, even if it conflicts with the driver's and the onboard AI agent's preference for proximity.

## Interoperability of multi-agent systems

One technical challenge in multi-agent systems is associated with enabling effective communication between different AI agents and AI agent systems.<sup>28</sup> In some cases, interactions are limited by the boundaries of native application environments, restricting the potential of AI agents to narrower and more specialized subdomains, where control is more easily retained.

The interoperability of AI agents relies on common communication protocols, which are the rules and standards governing how AI agents exchange information. These protocols can generally be categorized in two types:

- **Predefined protocols:** these are based on established agent communication languages and ontologies. Since they are predefined, the communication patterns are predictable

and consistent; however, they may not adapt well to dynamic environments where new communication needs arise.<sup>29</sup>

- **Emergent protocols:** these allow agents to learn how to communicate effectively based on their experiences, often using reinforcement learning techniques. This enables agents to adapt their communication strategies to changing environments and tasks.<sup>30</sup> However, decoding and understanding emergent communication remains an ongoing research challenge.<sup>31</sup>

A good understanding of the messages exchanged between AI agents is essential, otherwise it could affect the overall reliability of multi-agent systems. This inconsistency could lead to misunderstandings or misaligned actions when agents collaborate, especially in complex environments requiring precise coordination. To enhance the transparency of multi-agent interaction, the information exchanged needs to be easily accessible and interpretable by humans.

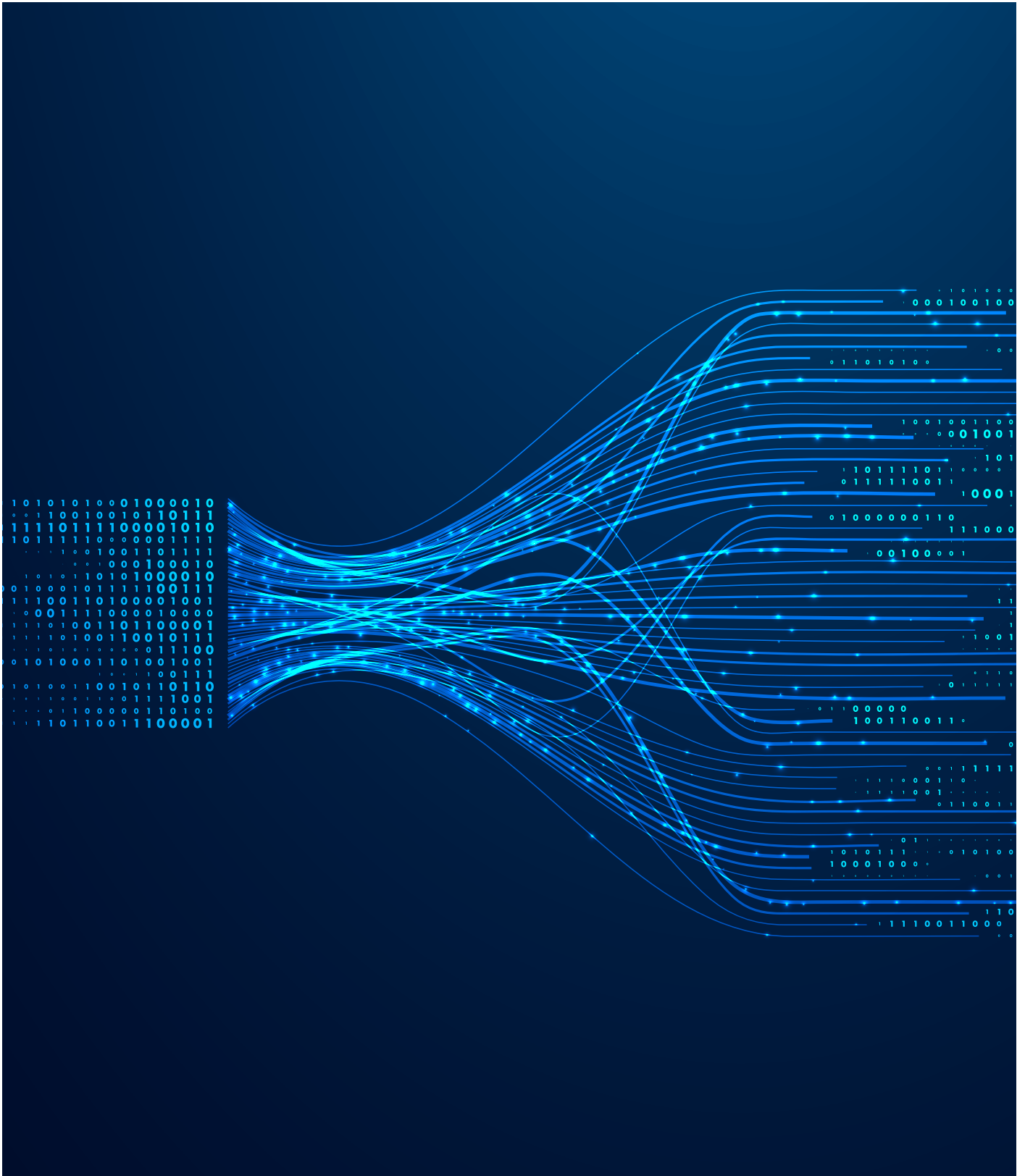




3

# Looking ahead

AI agents have the potential to tackle challenging tasks with great efficiency. But they carry associated risks such as malfunction, malicious use and unwanted socioeconomic effects.



## 3.1 Key benefits

By scaffolding capabilities such as reasoning, planning and self-checking on top of LLMs, more capable AI agents emerge that hold the potential to dramatically increase users' productivity and absolve them from certain tasks. This could involve completing tasks beyond users' skill sets, such as specialized coding, or partially or fully offloading tedious tasks that can be done more cheaply, quickly and at a greater scale than before. Additionally, the application of AI agents can play a crucial role in addressing the shortfall of skills in various industries, filling the gaps in areas where human expertise is lacking or in high demand.

Key characteristics of greater autonomy increasingly allow AI agents to tackle open-ended, real-world challenges that at one time were beyond them – for example, helping in scientific discovery, improving the efficiency of complex systems such as supply chains or electrical grids, managing rare and unusual scenarios in processes that are too infrequent to justify traditional automation, or enabling physical robots that can manipulate objects and navigate physical environments.<sup>32</sup>



### Examples of the benefits of applications of AI agents include:



#### Software development

AI agents can help generate, run and check code and other artefacts needed, allowing software developers to focus on higher value-added activities.



#### Healthcare

AI agents could improve diagnostics and personalized treatment, reducing hospital stays and costs through data analysis and decision-making support. For example, in under-resourced areas, AI agents could help alleviate the workload of clinical specialists by assisting doctors in developing tailored treatment plans.<sup>33</sup>



#### Enhanced customer experience

AI agent-based chatbots or virtual assistants can offer personalized, round-the-clock support, increasing customer satisfaction. They have the potential to provide consistently accurate responses, helping businesses maintain communication quality and resolve customer issues efficiently.<sup>34</sup>



#### Education

AI agents could help personalize learning experiences by adapting content to each student's needs, offering real-time feedback and supporting teachers with grading and administrative tasks. This allows educators to focus more on creative and interactive learning experiences.



#### Finance

AI agents could help enhance fraud detection, optimize trading strategies and offer personalized financial advice. They can analyse large datasets to identify patterns and trends, providing faster and more accurate insights for decision-making.

## 3.2 Examples of risks and challenges

While AI agents have the potential to offer numerous benefits, they also come with inherent risks, as well as novel safety and security implications. For example, an AI system independently pursuing misaligned objectives could cause immense harm, especially in scenarios where the AI agents' level of autonomy increases while the level of human oversight decreases. AI agents learning to deceive human operators, pursuing power-seeking instrumental goals or colluding with other misaligned agents in unexpected ways could pose entirely novel risks.<sup>35</sup>

Agent-specific risks can be both technical and normative. Challenges associated with AI agents stem from technical limitations, ethical concerns and broader societal impacts often associated with a system's level of autonomy and the overall potential of its use when humans are removed from the loop. Without a human in the loop at appropriate steps, agents may take multiple consequential actions in rapid succession, which could have significant consequences before a person notices what is happening.<sup>36</sup>

AI agents can also amplify known risks associated with the domain of AI and could introduce entirely new risks that can be broadly categorized into technical, socioeconomic and ethical risks.

### Technical risks

Examples of technical risks include:

- **Risks from malfunctions due to AI agent failures:** AI agents can amplify the risks from malfunctions by introducing new classes of failure modes. LLMs, for example, can enable agents to produce highly plausible but incorrect outputs, presenting risks in ways that were not possible with earlier technologies. These emerging failure modes add to traditional issues such as inaccurate sensors or effectors and encompass capability- and goal-related failures, as well as increased security vulnerabilities that could lead to malfunctions.<sup>37</sup>

Capability failures occur when an AI agent fails to perform the tasks it was designed for, due to limitations in its ability to understand, process or execute the required actions. Goal-related failures occur when a system is highly capable but nevertheless pursues the wrong goal. These issues can be caused by:

- **Specification gaming:** When AI agents exploit loopholes or unintended shortcuts in their programming to achieve their objectives, rather than fulfilling their goals.<sup>38</sup>

- **Goal misgeneralization:** When AI agents apply their learned goals inappropriately to new or unforeseen situations.<sup>39</sup>
- **Deceptive alignment:** When AI agents appear to be aligned with the intended goals during training or testing, but their internal objectives differ from what is intended.<sup>40</sup>
- **Malicious use and security vulnerabilities:** AI agents can amplify the risk of fraud and scams increasing both in volume and sophistication. More capable AI agents can facilitate the generation of scam content at greater speeds and scale than previously possible, and AI agents can facilitate the creation of more convincing and personalized scam content. For example, AI systems could help criminals evade security software by correcting language errors and improving the fluency of messages that might otherwise be caught by spam filters.<sup>41</sup> More capable AI agents could automate complex end-to-end tasks that would lower the point of entry for engaging in harmful activities. Some forms of cyberattacks could, for example, be automated, allowing individuals with little domain knowledge or technical expertise to execute large-scale attacks.<sup>42</sup>
- **Challenges in validating and testing complex AI agents:** The lack of transparency and non-deterministic behaviour of some AI agents creates significant challenges for validation and verification. In safety-critical applications, this unpredictability complicates efforts to assure system safety, as it becomes difficult to demonstrate reliable performance in all scenarios.<sup>43</sup> While failures in agent-based systems are expected, the varied ways in which they can fail adds further complexity to safety assurance. Failsafe mechanisms are essential but could be harder to design due to uncertainty on potential failure modes.<sup>44</sup>

### Socioeconomic risks

Examples of socioeconomic risks include:

- **Over-reliance and disempowerment:** Increasing autonomy of AI agents could reduce human oversight and increase the reliance on AI agents to carry out complex tasks, even in high-stakes situations. Malfunctions of the AI agents due to design flaws or adversarial attacks may not be immediately apparent if humans are not in the loop. Additionally, disabling an agent could be difficult if a user lacks the required expertise or domain knowledge.<sup>45</sup>

Pervasive interaction with intelligent AI agents could also have long-term impacts on individual



and collective cognitive capabilities. For example, increased reliance on AI agents for social interactions, such as virtual assistants, AI agent companions, therapists and so on could contribute to social isolation and possibly affect mental well-being over time.

- **Societal resistance:** Resistance to the employment of AI agents could hamper their adoption in some sectors or use cases.
- **Employment implications:** The use of AI agents is likely to transform a variety of jobs by automating many tasks, increasing productivity and altering the skills required in the workforce, thus causing partial job displacement. Such displacement could primarily affect sectors reliant on routine and repetitive tasks, in industries such as manufacturing or administrative services.
- **Financial implications:** Organizations could face higher costs associated with the deployment of AI agents, such as expenses for securing software systems against cyberthreats and managing associated operational risks.

## Ethical risks

Examples of ethical risks include:

- **Ethical dilemmas in AI decision-making:** The autonomous nature of AI agents raises ethical questions about their decision-making capabilities in critical situations.
- **Challenges in ensuring AI transparency and explainability:** Many AI models operate as “black boxes”, making decisions based on complex and opaque processes, thereby making it difficult for users to understand or interpret how decisions are made.<sup>46</sup> A lack of transparency could lead to concerns about potential errors or biases in the AI agent’s decision-making capabilities, which would hinder trust and raise issues of moral responsibility and legal accountability for decisions made by the AI agent.

## 3.3 Addressing the risk and challenges

To enable the autonomy of AI agents for cases where it would greatly improve outcomes, several challenges must be addressed. These challenges include safety and security-related assurance, regulation, moral responsibility and legal accountability, data equity considerations, data governance and interoperability, skills, culture and perceptions.<sup>47</sup> Addressing these challenges requires a comprehensive approach throughout the stages of design, development, deployment and use of AI agents as well as changes across policy and regulation. As advanced AI agents and multi-agent systems continue to evolve and integrate into various aspects of digital infrastructure, associated governance frameworks that take increasingly complex scenarios into consideration need to be established.

In assessing and mitigating the risks of potential harm from AI agents, it is essential to understand the specific application and environment of the AI agent (including stakeholders that may be affected). The risks of potential harm from an AI agent stem largely from the context in which it is deployed.<sup>48</sup> In high-stakes environments such as healthcare or autonomous driving, even small errors or biases can lead to significant consequences for the users of such systems. Conversely, in low-stakes contexts, such as customer service, the same AI agent might pose minimal risks, as mistakes are less likely to cause serious harm.

Within the context of a specific application and environment, it is important to adopt a risk analysis methodology that systematically identifies, categorizes and assesses all of the risks associated with the AI agent. Such an approach helps ensure that appropriate and effective mitigation mechanisms and strategies can be implemented by relevant stakeholders at the technical, socioeconomic and ethical levels.

## Technical risk measures

Examples of technical risk measures:

- **Improving information transparency:** Where, why, how, and by whom information is used is critical for understanding how a system operates and why certain decisions are made by the agent. Measures can be implemented to improve the transparency of AI agents such as the integration of behavioural monitoring and implementation of thresholds, triggers and alerts that involve continuous observation and analysis of the agent’s actions and decisions. Implementing behavioural monitoring helps to ensure that failures are better understood and properly mitigated when they occur.<sup>49</sup>

## Socioeconomic risk measures

Examples of socioeconomic risk measures:

- **Public education and awareness:** Developing and executing strategies to inform and engage the public are essential to mitigate the risks of over-reliance and disempowerment in social interactions with AI agents. These efforts should aim to equip individuals with a solid understanding of the capabilities and limitations of AI agents, allowing for more informed interactions, along with healthy integrations.
- **A forum to collect public concerns:** Acceptance and involvement, trust and psychological safety are crucial to tackle societal resistance and for the proper adoption and integration of AI agents into various processes. Without sufficient human “buy-in”, the implementation of AI agents would face significant challenges. In addressing societal resistance and creating wider trust in AI agents and autonomous systems, it is important that public concerns are heard and addressed throughout the design and deployment of advanced AI agents.<sup>50</sup>
- **Thoughtful strategies for deployment:** Organizations can embrace deliberate strategies around increased efficiency and task augmentation rather than focusing on outright worker replacement efforts. By prioritizing proactive measures such as retraining programmes, workers can be supported in transitioning to new or changed roles.

## Ethical risk measures

Examples of ethical risk measures:

- **Clear ethical guidelines:** Prioritizing human rights, privacy and accountability are essential measures to ensure that AI agents make decisions that are aligned with human and societal values.<sup>51</sup>
- **Behavioural monitoring:** Implementing measures that allow users to trace and understand the underlying reasoning behind an AI agent’s decisions is necessary to mitigate transparency challenges.<sup>52</sup> Behavioural monitoring can make system behaviour and decisions visible and interpretable, which enhances overall user understanding of interactions. This approach also strengthens the governance structure surrounding AI agents and helps increase stakeholder accountability.<sup>53</sup>

As the adoption of AI agents increases, critical trade-offs need to be made. Given the complex nature of many advanced AI agents, safety should be regarded as a critical factor alongside other considerations such as cost and performance, intellectual property, accuracy, and transparency, as well as implied social trade-offs when it comes to deployment.

The level of autonomy of advanced AI agents is likely to continue to increase due to ever more capable models and reasoning capabilities.<sup>54</sup> The complexities of more advanced systems call for a multidisciplinary approach that includes diverse stakeholders, from scientists and researchers to psychologists, developers, system and service integrators, operators, maintainers, users and regulators, all of whom are needed to establish appropriate risk management frameworks and governance protocols for the deployment of more sophisticated AI agent systems.

This white paper has taken a first step in outlining the landscape of frontier AI agents, but further research is needed to provide more details on the safety, security and socioeconomic implications as well as the novel governance measures required to address them.



# Conclusion

AI agents are becoming more autonomous in their operation and decision-making, bringing potential benefits and risks.

The development of AI agents has been marked by significant milestones, from the early days of simple reflex agents to sophisticated multi-agent systems. Recent advances in LLMs and LMMs have resulted in the next evolution of AI agents, which have moved from basic systems that react to immediate stimuli to complex entities capable of planning, learning and making decisions based on a comprehensive understanding of their environment and user needs.

The ongoing development of AI agents is fundamentally linked to increased autonomy, improved learning capabilities, enhanced decision-making abilities and multi-agent collaboration. As the architecture and emerging use cases for AI agents continue to proliferate, the shift towards multi-agent systems that can collaborate in increasingly complex environments is likely to continue.

Increased autonomy plays an important part in the evolution of AI agents and creates novel opportunities for new applications while also presenting unique risks to society. The introduction of AI agents will likely reduce the need for human involvement and oversight in some areas, bringing a more efficient approach to tedious tasks. However, a reduction in human oversight could also increase the risk of accidents. Furthermore, increased automation of workflows could be a way for malicious actors to exploit novel vulnerabilities, while also exacerbating socioeconomic and ethical risks.

The rapid advance of AI agent capabilities is set to be followed by a wave of innovation in AI agents, which could have the ability to transform the global economy and the roles of human labour in new and significant ways.

Further research is necessary to explore the safety, security and societal impacts of AI agents and multi-agent systems, emphasizing both technical solutions and organizational governance frameworks. These efforts are critical for mitigating risks associated with the ongoing development, deployment and increasing use of more sophisticated AI agents in a range of domains.

At this point, it is vital for stakeholders to come together throughout technical, civil society, applied and governance-facing communities to research, discuss and build consensus on novel governance mechanisms.

This white paper has offered an initial exploration of the rapidly evolving landscape of AI agents, aiming to promote deeper understanding of this emerging field and spark conversation on responsible adoption and diffusion practices. Through equitable development, deployment and governance, the growing presence of advanced AI agents holds the promise of driving positive societal transformation for many years to come.