# Business, *meet* agentic AI

## Confidence in *autonomous* and *agentic* systems

Capgemini

# Foreword

In Capgemini's AI Futures Lab, we study advances in AI and the implications that those advances will have on the world. It's a fascinating and awe-inspiring work where we are confronted every day by the amazing possibilities coming over the horizon, but the potential implications of autonomous and agentic systems are unparalleled. The reason that we're so excited about this is not particularly because of the technology behind it or any particular AI model that's used, but because it fundamentally shifts our relationship with technology.

***The reason agentic and autonomous systems are transformative is because they allow users of technology to shift from defining the solution to simply stating their problem.***

For the entirety of our history with computing, if we wanted to make a computer do something we needed to describe in great detail how to solve that problem, either by programming it ourselves or relying on experts who could. That was, by definition, an exclusive arrangement where only people who understood technology could get the most out of it. The era of autonomous and agentic AI presents us with a new vision of the future, one where users of technology can command technology to solve problems that they themselves have no idea how to solve. This is the version of AI that we were always promised by science fiction, where anyone can harness the full power of AI.

**Robert Engels**
Head of AI Futures Lab
*robert.engels@capgemini.com*

**Mark Roberts**
Deputy Head of AI Futures Lab
*mark.roberts@capgemini.com*

# Executive summary

Today's technology landscape is quickly changing. Autonomous and agentic AI systems, systems that can make decisions and take action on their own, are becoming increasingly important. These systems aren't just a small step forward; they represent a major shift in how people interact with and experience technology.

Autonomy is a game changer. By allowing autonomous and AI agents to take actions we unlock amazing new opportunities, but these do not come without risks.

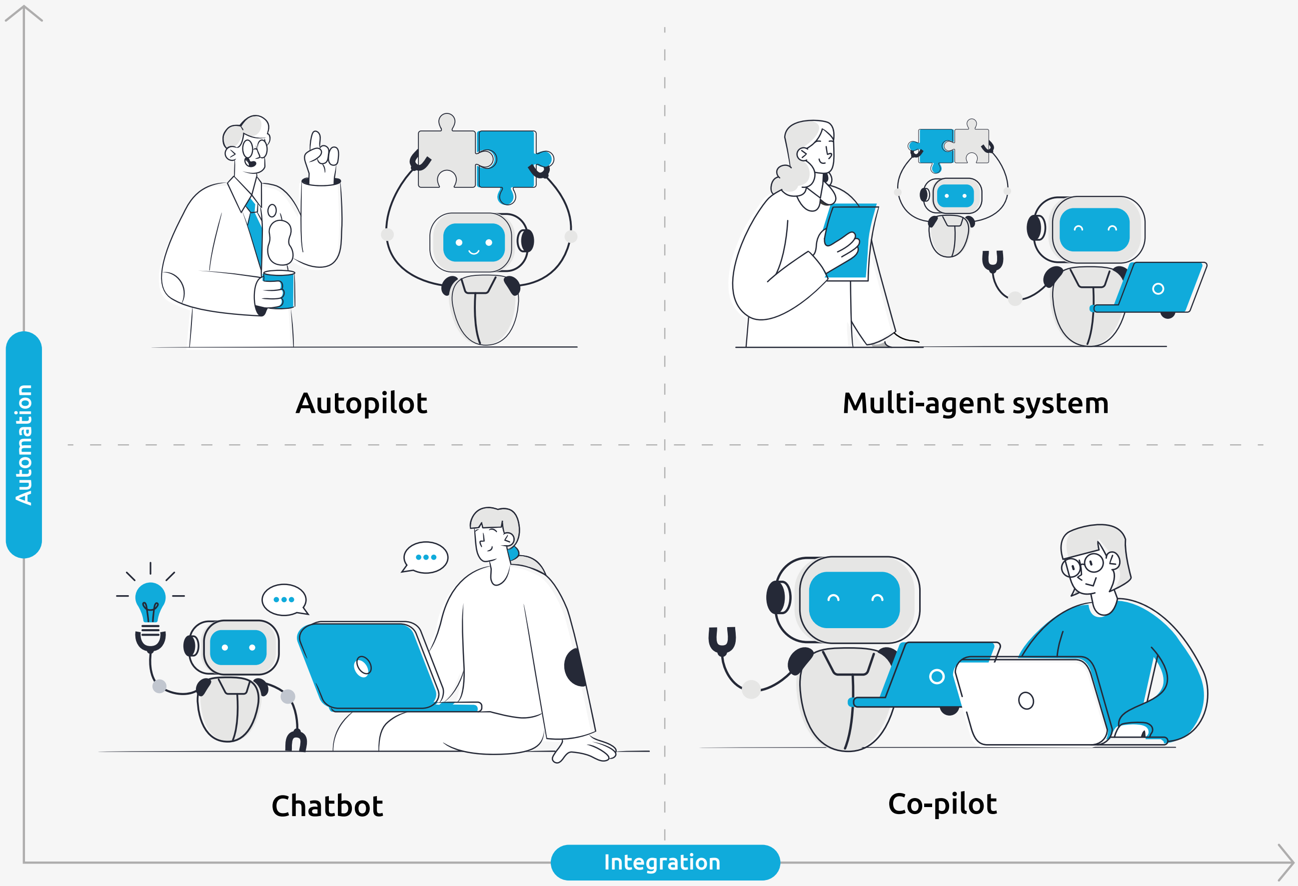
This white paper builds on our 2024 edition of Unleashing Confidence in AI. *Confidence in autonomous and agentic systems* looks at how autonomous AI systems are changing our understanding of artificial intelligence. While many basic ideas remain the same, autonomy brings new challenges that require a fresh perspective.

In this new era where autonomous AI systems interact and co-exist with human society, ensuring AI is reliable and meets human expectations is of utmost importance. This white paper expands these categories to address the special needs of autonomous and agentic AI systems.

# Table of Contents

# Journey to multi-agency

Autonomous AI systems have existed long before today's AI boom. For many years, systems with different levels of independence and integration have gradually developed. Understanding this history helps us appreciate the full gravity of today's advancements.



Though the era of agentic AI began 50 years ago, the development of simple chatbots showing limited independence and integration have brought the technology into recent focus. These basic systems, like early website help tools, had limited capabilities and limited access to information which was not already included in the models they used.

As integration improved, co-pilot systems emerged. These more advanced tools could access and interpret data across multiple systems, offering more helpful assistance. However, for the most part, these systems only gave advice and couldn't act on their own.

The next step forward was autopilot systems. These tools had enough independence to take specific actions, but lacked the complete integration capabilities required to handle entire processes. These systems underlined the value of letting machines act independently.

Today, we're seeing multi-agent AI systems that combine high levels of both independence and integration. These advanced systems can create great value, but need careful oversight and risk management. The move from single AI agents to systems with multiple AI agents working together marks a fundamental change in both the capability and complexity of this technology.

This history shows an important fact: the shift from single agents to multi-agent systems isn't just about having more agents; it's about creating systems with entirely new properties and emergent behavior when multiple agents work together in shared environments.

# Understanding agent properties

## What makes something an agent?

Agents are central to modern AI discussions, and terms like "agent" and "agentic" are used everywhere, often with varying meanings. Despite their recent popularity, these concepts have deep roots in computer science with well-established definitions.

> An agent is any entity that works on behalf of another entity, working to accomplish high-level objectives often using specialist capabilities. Agents have a degree of *autonomy* and *authority* to *take actions* that modify their world.

Being precise about what we mean by "agent" is important. As competing definitions make it hard to communicate effectively about these systems. Creating a shared understanding is essential for building connected agent ecosystems that work well together.

Naturally, different fields focus on different aspects of agents. Business perspectives often focus on what agents can accomplish and how they affect organizations. Technical definitions typically focus on what agents can do and how they work. Bridging these viewpoints requires a definition that makes sense to everyone.

At its core, an agent is any entity that works on behalf of another entity, be it another agent, a human, or maybe even other organisms.

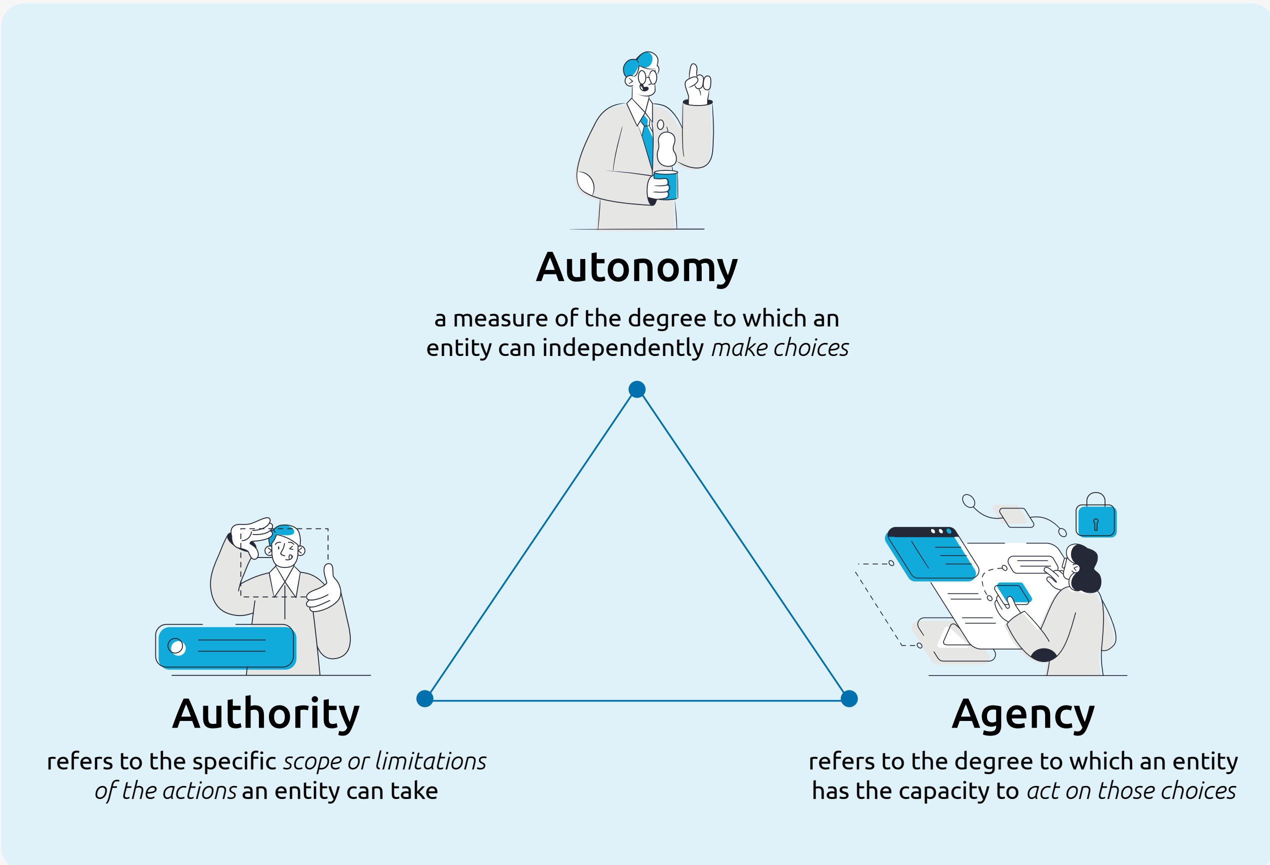
This highlights a crucial distinction: The ability to take action is what makes something an agent. An AI system might provide sophisticated analysis and recommendations, but if it can't execute actions on its own, it's an assistant or a co-pilot, but not an agent. It's important to note that an agent doesn't have to use artificial intelligence. Many non-AI systems are agents, from simple thermostats to your car's driving assistance systems to complex industrial control systems. Similarly, many AI agents also lack the action capabilities needed to be agents.

AI agent systems are build around 3 main aspects (or views):

- Agent view: how much autonomy, agency, authority
- System view: what are the systemic properties agents work in
- Payload view: what are the capabilities

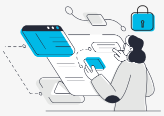
In the next sections, we will build the foundation for analyzing, designing, and managing autonomous AI systems across different situations and applications.

## The agent view: autonomy, authority, and agency



There are three related concepts that define what an agent can do, each representing a different aspect of capability:

**Autonomy** describes the extent to which an entity can make independent decisions on its own without outside direction. Autonomous systems evaluate situations, create options, and choose actions without human involvement.



**Agency** describes an entity's ability to act on those decisions by taking actions that affect its environment. Agency is the link between making a decision and creating a result.

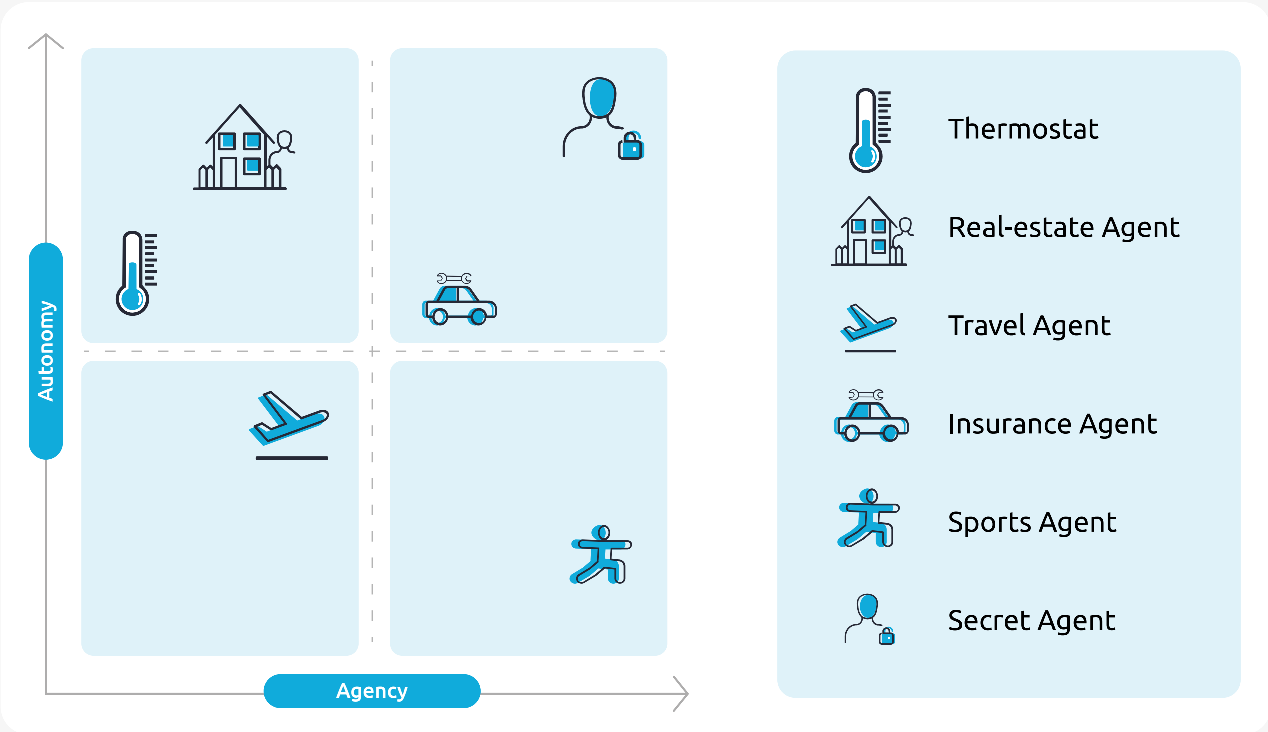
**Authority** defines what actions an entity is allowed to take. Authority sets boundaries around what an agent can do.

Rather than being binary attributes, these properties exist on sliding scales. For example, consider a standard thermostat: It has high autonomy (independently deciding when to turn on heating), high agency (directly controlling the heating system), but limited authority (it can only do one thing within narrow limits). This example shows how autonomy, agency, and authority can combine in different ways depending on an agent's purpose and design.

It's important to remember that these attributes are design choices, not inherent system properties. When designing agentic systems the decision on appropriate levels of autonomy, agency, and authority is made based on intended functions, risk considerations, and oversight requirements. These deliberate choices shape what an agent can do, what limits it has, and how effective it is. Together, these three dimensions provide a framework for systematically evaluating agent capabilities and risks. An agent with high autonomy, agency, and authority might deliver exceptional value, but would also present significant risks that require careful oversight. The best configuration depends entirely on the specific context, goals, and risk tolerance of where the agent will be used.

## Human society as an analogy for AI systems

Deciding appropriate levels of autonomy and agency isn't unique to artificial systems. Human societies have developed nuanced frameworks for delegating authority to various professional agents, providing useful models for the design of AI systems. For example, our societal constructs, such as governments and councils, all include checks and balances that ensure that autonomy, agency and authority are regulated in line with the risks they pose.




Understanding how we grant authority to human agents offers valuable insights for AI governance. Real estate, travel, and insurance agents are some common examples that show how authority is carefully calibrated according to risk, expertise, and trust requirements.

Real estate agents have high levels of autonomy to market properties, but limited agency regarding final price negotiations and contract terms. Travel agents typically have the agency to make bookings within specified limits, but often lack the agency to make significant itinerary changes without explicit approval. Insurance agents typically have greater agency and autonomy due to their specialized knowledge and the complexity of insurance products.

Looking at extreme examples, sports agents have high agency but restricted autonomy – negotiating contracts only when explicitly authorized to do so by their clients. On the opposite end, intelligence operatives have high autonomy with high agency, receiving broad mission objectives while keeping almost complete discretion about methods and execution.

These examples highlight a basic principle: How much autonomy and agency we give to agents depends on purpose, risk, and trust. The same principle applies to artificial agents. Systems will work within the boundaries set by their designers, optimizing for the objectives they're assigned. Carefully considering these boundaries is an essential part of responsible system design.

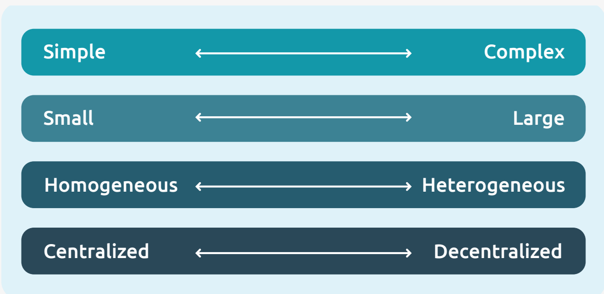
# The system view: coordinating multiple agents

While individual agents can provide significant value in many situations, the full potential of agentic approaches emerges through multi-agent systems – coordinated groups of specialized agents operating in shared environments. The shift from isolated agents to integrated systems introduces new dimensions of complexity, capability, and governance requirements.

A multi-agent system (MAS) consists of multiple independent agents operating within a common environment, working together to achieve goals beyond what any individual agent could accomplish. These systems, sometimes called agentic architectures or frameworks, represent the cutting edge of autonomous system development.

## The four key dimensions of multi-agent systems

There are four key dimensions that shape the capabilities, behaviors, and governance requirements of multi-agent systems:




- **Simplicity vs. complexity:** System complexity is determined by how agents interact with each other. Simple systems have straightforward workflows with predictable agent interactions and clearly defined information flows. Complex systems involve intricate interdependencies, feedback loops, and emergent behaviors that can't be easily predicted from looking at individual agents. Complexity brings both enhanced capabilities but also governance challenges, requiring sophisticated monitoring and oversight mechanisms.

- **Small vs. large:** The number of agents in a system strongly influences how it operates. Small systems with agents typically show more predictable behaviors and straightforward governance requirements. Large systems with numerous agents offer enhanced capabilities and the ability to scale to larger challenges, but introduce coordination challenges, emergent behaviors, and governance complexities that demand specialized approaches to design, monitoring, and control.

- **Heterogeneity vs. homogeneity:** System composition varies from homogeneous collections of similar agents (often called swarms) to heterogeneous collections of specialized agents with distinct capabilities. Homogeneous systems excel in parallelism and resilience, but struggle with multi-faceted challenges. Heterogeneous systems can leverage specialized agent capabilities to address specific problems, yet require sophisticated coordination mechanisms to ensure effective collaboration. Though diversity in agent capabilities can enhance system adaptability, it can also complicate coordination and governance.

- **Centralization vs. decentralization:** Decision-making authority can be distributed from strictly centralized architectures to fully decentralized systems. Centralized systems maintain clear command structures with decision authority concentrated in designated control points, enabling precise governance, but creating potential bottlenecks and single points of failure. Decentralized systems distribute decision-making broadly across the agent network, enhancing resilience and scalability, but potentially complicating system-wide governance and alignment.

The centralization spectrum represents a fundamental trade-off between control and resilience.

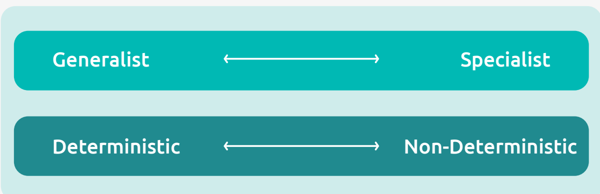
These dimensions might suggest that maximizing along each axis (larger, more complex, more heterogeneous, more decentralized) would create optimal systems. However, such maximization introduces significant challenges for control, predictability, and alignment. The appropriate configuration depends entirely on specific requirements, risk tolerance, and governance capabilities of the implementation context.

# The payload view: what agents can actually do

## What is a payload?

The system and agent views talk only about the overall architecture of a system in which autonomous agents operate, but say nothing at all about what those agents can actually do or how they work internally. The payload within these agentic containers determines the agent's functions, capabilities and behaviors and is independent of the overall system that they operate within.

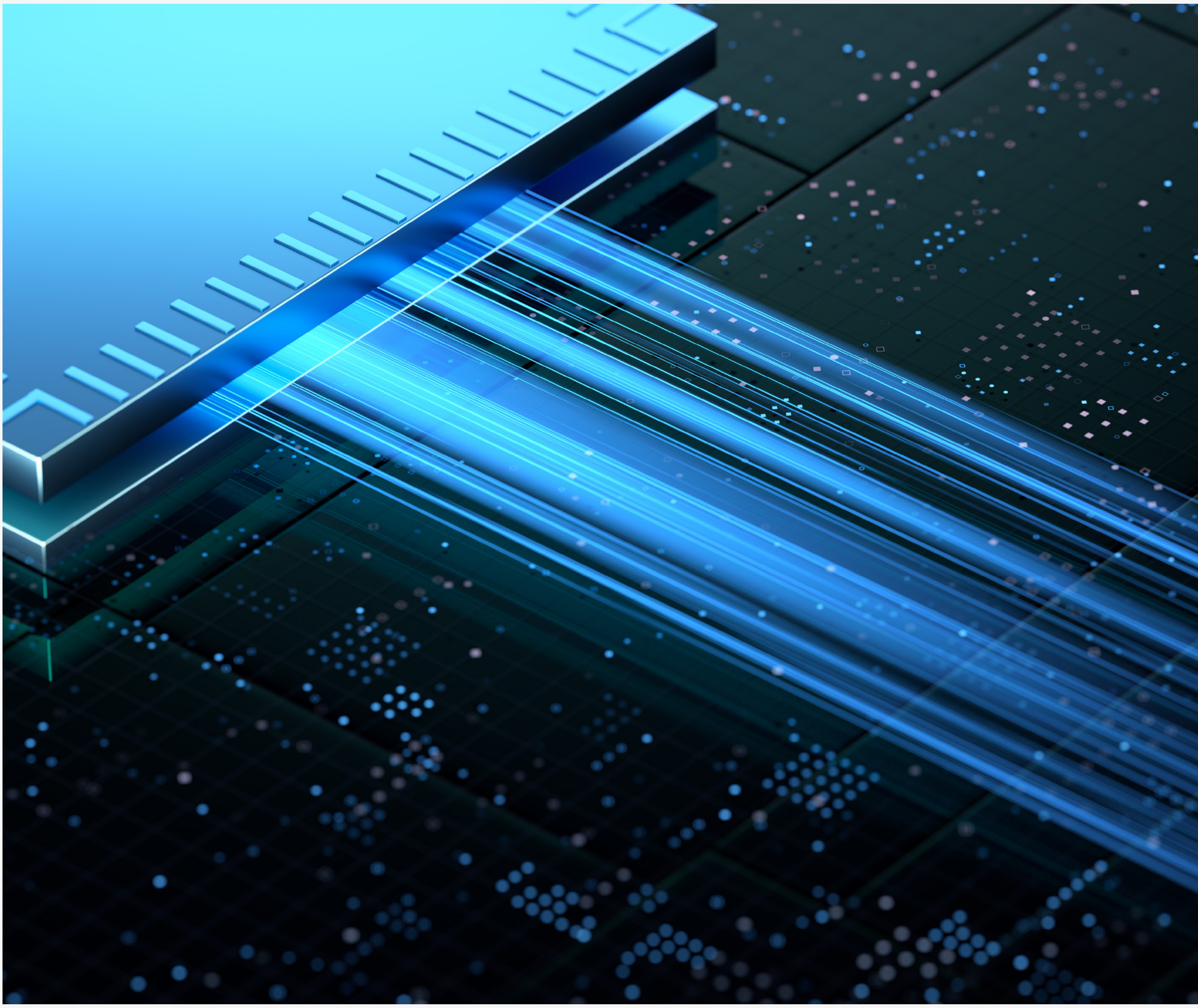
The payload view focuses on the concrete capabilities of individual agents rather than their relationships or system organization. These capabilities ultimately determine the practical utility and impact of the system regardless of its architectural sophistication.




Two fundamental dimensions characterize agent payloads, shaping their operational characteristics and suitability for specific applications:

- **Specialization vs. generalization:** Agents range from highly specialized entities focused on narrow domains to generalized systems capable of addressing diverse challenges. Specialized agents excel within defined domains, delivering high performance, but with limited flexibility. They typically exhibit high agency within their specialization, but limited authority beyond it. Generalized agents demonstrate greater adaptability across varied contexts, but rarely match the performance of specialized agents within specific domains. They often feature higher autonomy due to their broader decision-making capabilities. While individual agents may be specialized or generalized, entire systems can likewise be characterized by their overall specialization profile and the balance they strike between depth and breadth of capability.

- **Determinism vs. non-determinism:** The predictability of agent behavior spans from fully deterministic systems that always produce identical outputs given identical inputs to non-deterministic systems capable of adaptation and evolution over time. Traditional agentic systems have predominantly employed deterministic approaches – a thermostat detecting 20°C consistently triggers the same response between such measurements. Deterministic systems offer reliability and predictability, but limited adaptability. Non-deterministic systems can learn and evolve, potentially



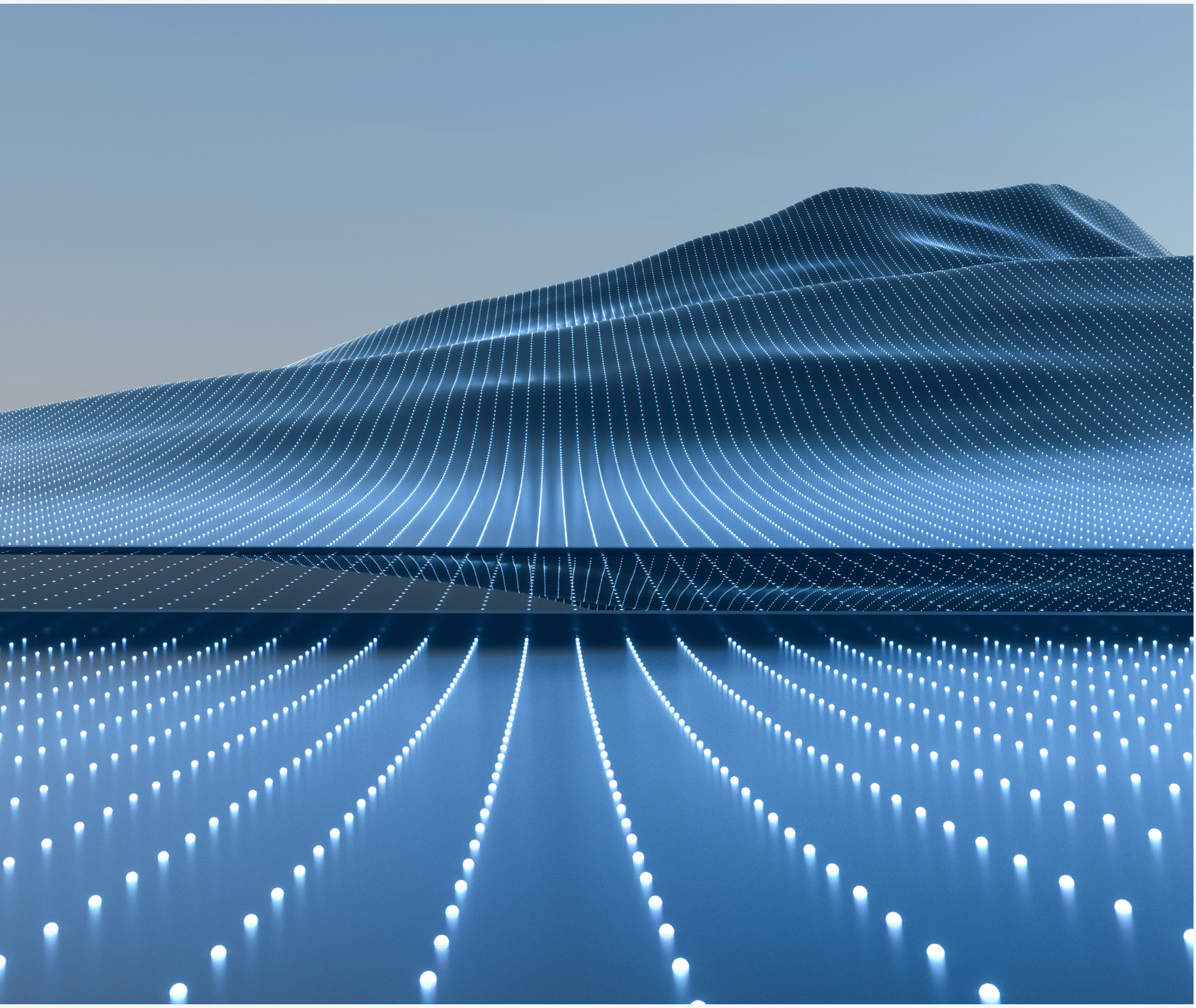
# Making systems autonomous

Given the above definitions, we should consider the minimum logical architecture required for an agent to function. A fundamental architecture consists of three essential layers, each addressing a distinct functional requirement:

**The interpretation layer:**
Enables communication and interaction with the agent's environment, including other systems and human operators. This layer translates external information that arrives through a variety of sensors and other inputs into formats agents can process. The agents then create outputs suitable for external consumption. For human interaction, natural language capabilities provide an effective interface; for system integration, standardized APIs offer efficient communication channels. The interpretation layer determines what information the agent can access and how effectively it can communicate its outputs and actions.

**The understanding/knowledge layer:**
Contains the information and processing capabilities that enable the agent to comprehend its operating environment, predict outcome probabilities, and reason about the systemic impacts of potential actions. This layer must maintain sufficient contextual understanding to support effective decision-making aligned with the agent's purpose. The sophistication of this layer largely determines the agent's capability to make appropriate decisions in complex or novel situations.

**The outcome/action layer:**
Implements the agent's decisions as concrete interventions in its environment. Without this capability to effect change, an entity fundamentally cannot function as an agent regardless of its interpretive or knowledge capabilities. The action layer may employ deterministic or non-deterministic mechanisms depending on the agent's design and purpose. The scope and power of an agent's action capabilities directly correlate with both its potential value and the governance requirements it necessitates.



# LLMs within the agentic landscape

A common misunderstanding is that large language models (LLMs) and agents are the same thing, and the implication that agentic systems must use LLM technology. This view misunderstands both what qualifies as an agent and how language models fit into agent designs.

As we established earlier, being an agent doesn't require AI components at all. Many agents operate effectively without any AI capabilities, let alone anything as sophisticated or heavyweight as an LLM. Agents and multi-agent systems can be built using a wide spectrum of technologies depending on specific requirements rather than following technology trends. Often, a more simple payload allows simpler and more scalable deployment of agentic technologies.

Interestingly, many systems marketed as agents are actually just LLMs without true agentic capabilities. Despite having sophisticated language processing and generation capabilities, these systems lack agency, i.e. the capacity to take action. This confusion in terminology makes it harder to communicate clearly about system capabilities and appropriate governance requirements.

When LLMs are used in genuine agent architectures, their role is often different from what people assume. Rather than functioning as the knowledge/understanding layer as frequently suggested, LLMs typically serve as interpretation layers, excelling at capabilities as multi-modality and natural language processing rather than comprehensive abstraction and world modeling. LLMs are remarkably good at translating between human language and machine-processable formats, but have limited ability to represent world knowledge compared to specialized (non-LLM) knowledge systems.

This raises an important question: If LLMs haven't fundamentally changed what it means to be an agent, why have agent architectures become so popular during the LLM era? The answer lies in the communication capabilities that LLMs provide. Natural language offers a universal interface for communicating with artificial systems about goals, constraints, and preferences. Similarly, natural language helps agents communicate with each other without needing custom integration protocols. LLMs have made agentic systems more popular by making dramatic improvements in interpretation rather than by changing agency itself.

# World models: How agents understand their environment

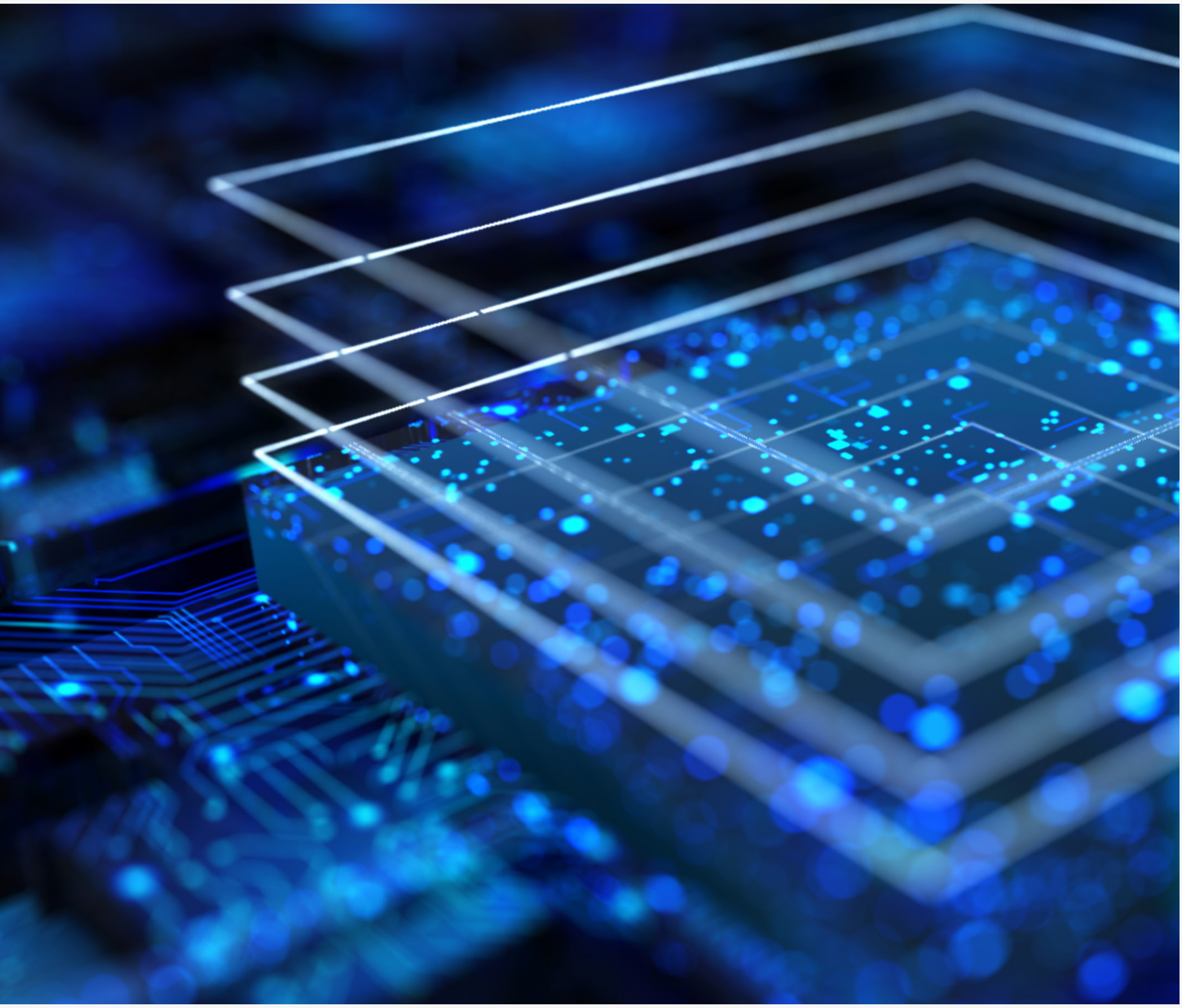
Effective agency requires contextual understanding. Agents operate within specific environments, known as "worlds" in AI terminology. These worlds represent everything that agents can perceive and influence. These environments might be narrowly defined software domains or aspects of physical reality. The agent's internal representation of this environment – its world model – determines its ability to make appropriate decisions and take effective actions.

Consider again the thermostat example. With a simple world model, it recognizes only temperature values. An advanced thermostat with a richer world model might incorporate occupancy patterns, thermodynamic models, weather forecasts, utility pricing dynamics, and user preferences. This enhanced contextual understanding enables seemingly intelligent behavior – preemptively adjusting temperature based on anticipated occupancy or utility rate changes – that builds user trust through apparent comprehension of relevant factors.

Inadequate world models inevitably produce suboptimal performance. A customer service agent lacking contextual understanding of customer history, preferences, and situation will invariably deliver unsatisfactory experiences regardless of its language capabilities or available actions. World models provide the contextual foundation for intelligent decision-making. The same is true for human agents too – if a real-estate agent did not have a good understanding of the local housing market, or a travel agent had a poor grasp of geography and climate, they would be unable to effectively achieve their goals.

Human cognition provides a useful parallel. All humans maintain internal world models that, despite individual variations, share enough commonality to enable collaboration, anticipation, and empathy. These shared models help us coordinate efficiently without having to explain the entire context for every interaction.

For artificial agents, world models serve not just as operational necessities, but as foundations for trustworthiness. They enable users to understand whether agent successes or failures stem from appropriate reasoning rather than lucky coincidence or misaligned understanding. A transparent world model allows users to assess whether an agent's actions reflect genuine understanding or merely superficial pattern matching.



# Purpose and alignment

One of the defining features of an agent is that it works towards a high-level goal. It's this property that enables a new paradigm of how we interact with technology. However, defining a goal is easier said than done, as it relies on being able to represent and communicate rich and precise descriptions of the state of the world. Defining an agent's purpose and measuring its alignment to that purpose are important tasks to consider when deploying agent-based systems. Purpose may be explicitly coded through rules and objectives, or implicitly shaped through learning processes. Regardless of how it's implemented, clarity of purpose remains essential for evaluating performance and determining whether an agent delivers its intended value.

Alignment issues emerge when agent behavior diverges from its intended purpose. This can potentially result from:

- Unclear purpose definition that doesn't capture true objectives
- Incomplete or inaccurate world models that provide insufficient context for good decisions
- Performance metrics that accidentally encourage undesired behaviors
- Unexpected behaviors that weren't anticipated during system design

As agents gain more independence and capability, alignment becomes both more important and more challenging. This explains why alignment research has become a central focus for AI safety specialists concerned with advanced systems. Ensuring that increasingly powerful autonomous AI systems remain reliably aligned with both explicit and implicit human intentions represents perhaps the most important challenge in agent development.

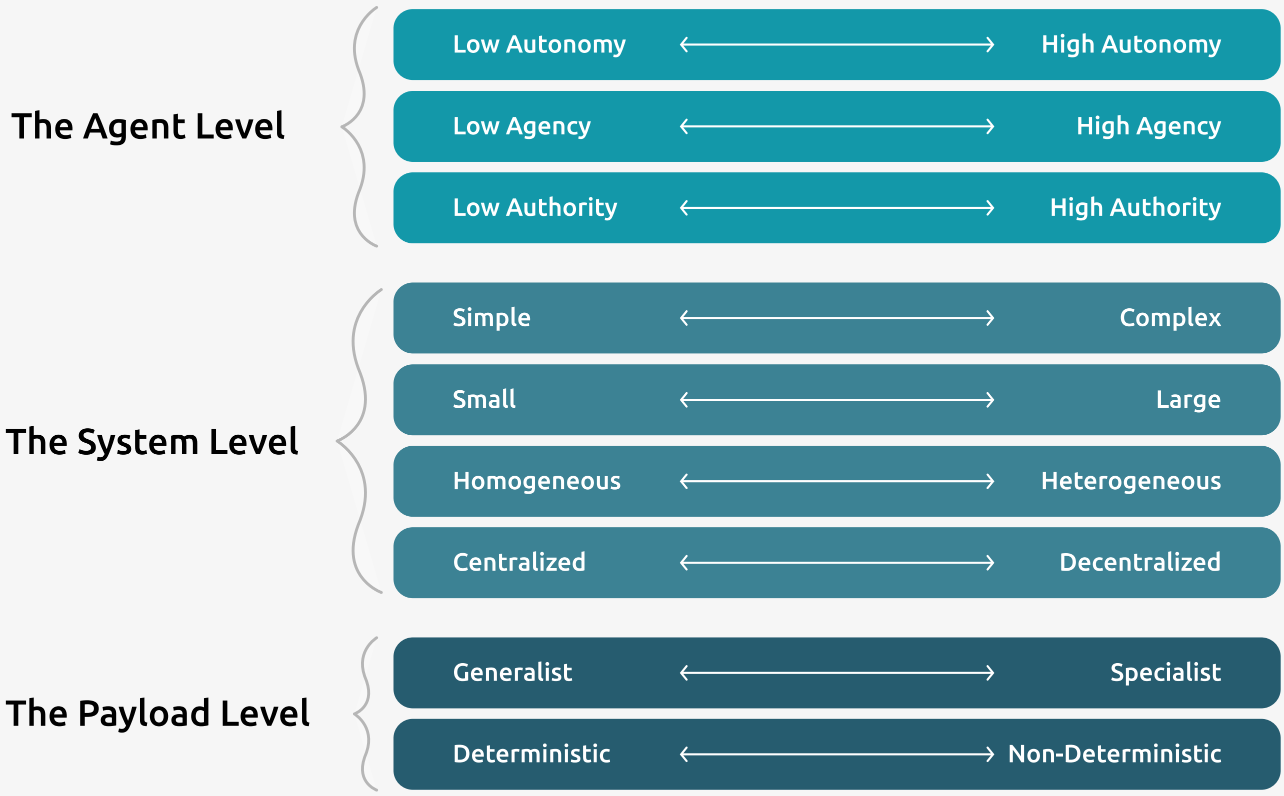
Consider this analogy; in the 2014 book "*Superintelligence*" by Oxford University professor Nick Bostrom, we were introduced to the now famous *Paperclip Maximiser Thought Experiment*. In this experiment, an AI system is given the explicit task of producing as many paperclips as possible. It quickly identifies that humans threaten the completion of this goal, as they may decide to turn it off and they're a great source of minerals that could be used to make even more paperclips. The result: Lots of paperclips, and no humans left to use them.

In this thought experiment, the AI's purpose is clear – it must maximize its paperclip production. However, its alignment is clearly inadequate. It will complete this task at the expense of anyone or anything else's safety and wellbeing. This is because the world model provided for the AI does not provide enough context around the other parameters and ethical considerations it should consider while achieving its goal.

# Bringing it all together

# Technical properties: A unified framework

The dimensions and properties discussed so far combine to form a framework for understanding autonomous and agentic systems. There are many other properties of multi-agent systems, but this unified perspective of the high-level properties enables the systematic analysis of existing systems and the intentional design of new ones across diverse application domains. This framework encompasses:




By thinking about autonomous AI systems along these dimensions, organizations can develop a shared vocabulary for discussing system characteristics, capabilities, and governance requirements. This common language facilitates more effective communication across disciplinary boundaries, enabling the collaborative development and implementation of autonomous AI systems aligned with organizational objectives and values.

## The business view, from technical possibility to organizational value

Technical capability alone cannot ensure the successful implementation of autonomous and agentic systems. Organizational readiness across multiple dimensions determines whether these systems will deliver sustainable value or create unmanageable complications. Effective implementation requires a complete business perspective, integrating technical, operational, ethical, and human factors.

From a business perspective, successful autonomous and agentic systems demand a multifaceted approach that begins with a clear purpose and measurable value proposition. Organizations must articulate which business needs these systems will address and how their impact will be measured. This clarity of purpose must be accompanied by robust governance structures that establish oversight, responsibility, and accountability throughout the system's lifecycle.

Effective risk management forms another critical pillar of business readiness, requiring organizations to systematically identify, assess, and mitigate potential risks before they materialize into problems. Closely related to risk management are ethical considerations. Agentic systems must align with organizational values and ethical standards to maintain trust and integrity both internally and externally.

The technical aspects of integration cannot be overlooked either. Even the most sophisticated autonomous system will fail to deliver value if it doesn't integrate smoothly with existing processes and systems. This integration challenge often highlights skills gaps within organizations, as developing, deploying, and maintaining these systems requires specialized expertise that may need to be cultivated or acquired.

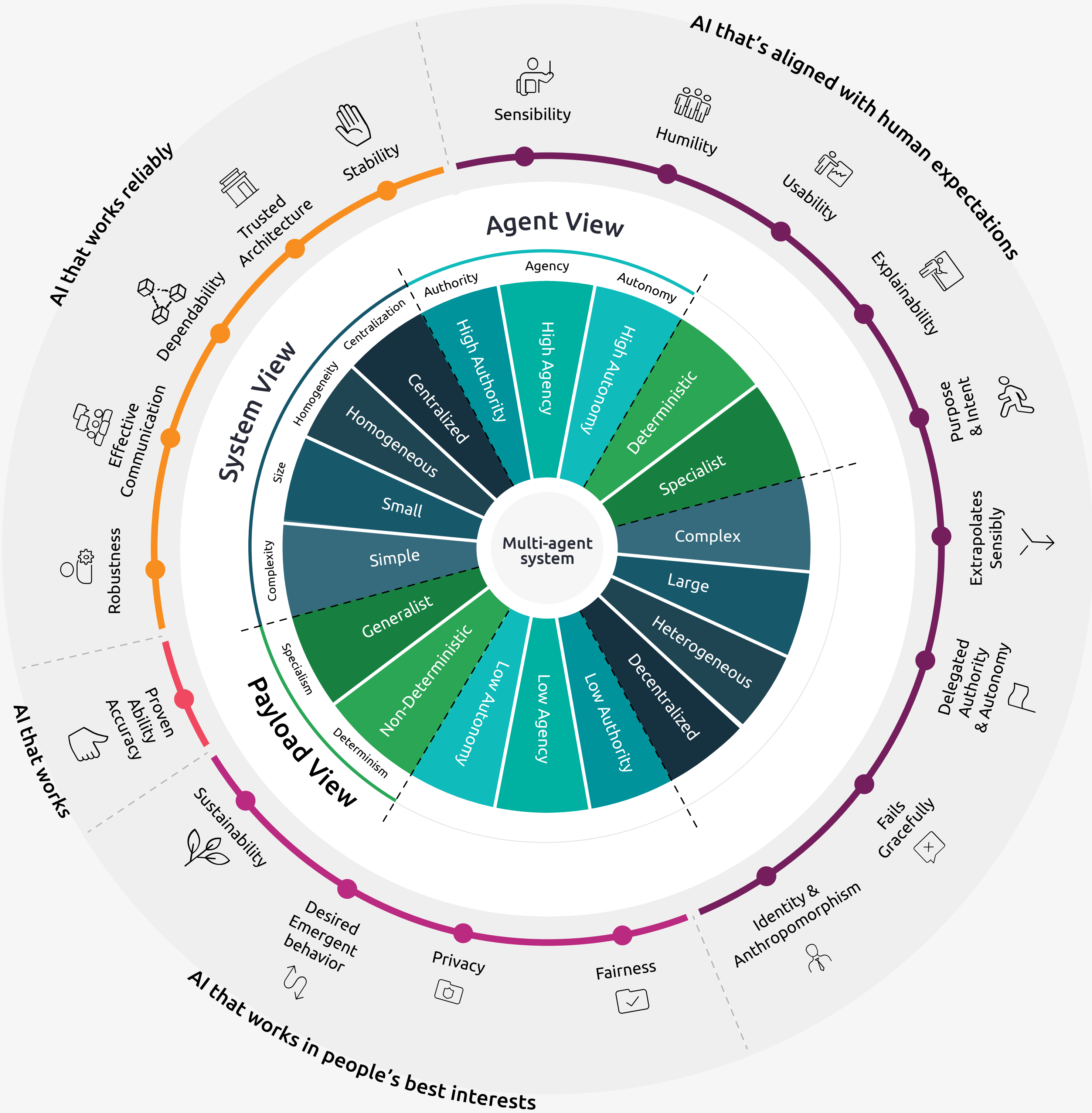
Perhaps most overlooked is the human element. Change management processes must be thoughtfully designed to help users adapt to working alongside autonomous systems, addressing concerns and building confidence through education and transparent communication. Finally, these systems are not "set and forget" solutions. Continuous monitoring and improvement mechanisms should be established to ensure they evolve with changing business needs and technological capabilities, creating a virtuous cycle of increasing value and reliability.

## The complete view: technical and business integration

A truly comprehensive understanding of autonomous and agentic systems emerges only by integrating technical and business perspectives into a unified whole. This holistic view encompasses internal technical architectures, external technological ecosystems, and broader organizational and societal impacts.

# Confidently deploying autonomous and agentic systems

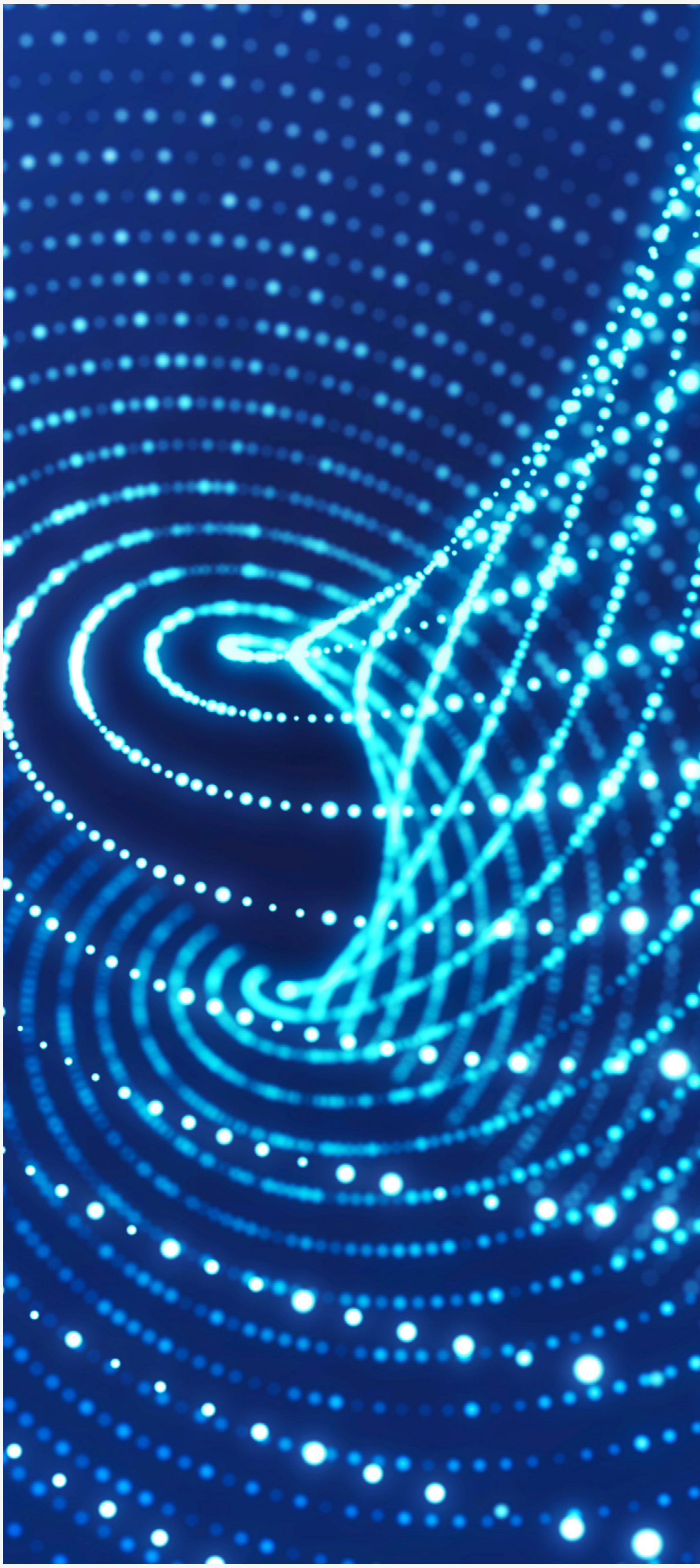
## A unified framework



This integrated perspective brings together diverse elements into a single, coherent picture. Technical architecture and capabilities form the functional basis, while value alignment secures meaningful relevance. Governance structures and risk management processes create a governance framework, complemented by ethical considerations that address broader societal interests.

Autonomous agents rarely function in isolation, requiring them to become integrated components within complex ecosystems. Such integration demands specialized capabilities that organizations must willingly develop or acquire. The human dimension remains important, as change management and user adoption ultimately determine whether systems are embraced or not. Across all dimensions, autonomous AI systems exist in continuous development rather than static deployment, requiring ongoing adaptation to maintain relevance and effectiveness.

By using this all-encompassing perspective, organizations can approach autonomous and agentic systems with a good balance between ambition, responsibility, and transformation potential while implementing the right safeguards.Such an approach enables businesses to leverage the capabilities of autonomous AI systems while ensuring proper alignment with human values, intentions, and objectives.



# Conclusion

Though agent architectures are often presented as being straightforward implementation patterns in need of minimal specialized knowledge or governance, the analysis presented here shows a much more nuanced reality, with autonomous and agentic systems in need of many technical, operational, ethical, and organizational considerations that require systematic approaches to design, implementation, and governance.

Autonomous AI systems and AI agents will certainly shape the future technological landscape, transforming how organizations operate and create value. Understanding the multi-dimensional nature of agency provides a basis for navigating this complex domain with confidence and responsibility. The insights from this white paper create a conceptual framework for the more detailed explorations that follow in subsequent chapters.

The journey ahead explores practical implementation strategies, reliability and alignment challenges, and domain-specific considerations across varied business contexts. By combining technical precision with business acumen, organizations can harness the transformative potential of autonomous and agentic systems while effectively managing risks and ensuring genuine value creation.

The autonomous future has arrived – not as a distant possibility, but as an immediate reality demanding thoughtful engagement. This white paper provides the conceptual tools and practical guidance necessary to navigate this new landscape with both ambition and responsibility.

# The AI Futures Lab

We are the AI Futures Lab - expert partners that help you confidently visualize and pursue a better, sustainable, and trusted AI-enabled future. We do this by understanding, pre-empting, and harnessing emerging trends and technologies. Ultimately, making possible trustworthy and reliable AI that triggers your imagination, enhances your productivity, and increases your efficiency. We will support you with the business challenges you know about and the emerging ones you will need to know to succeed in the future.

Build your AI advantage, layer by layer. Backed by extensive research and collaboration, we're best placed to help you navigate the AI landscape, and establish AI solutions that herald a step change in how we can solve business problems, holistically. Engage with us – let us surprise you with our visionary mix of what's to come.

## Lab Contacts

**Robert Engels**
Head of AI Futures Lab
robert.engels@capgemini.com

**Mark Roberts**
Deputy Head
of AI Futures Lab
mark.roberts@capgemini.com

**Jonathan Aston**
jonathan.kirk@capgemini.com

**Monika Byrtek**
monika.byrtek@capgemini.com

**Bora Ger**
bora.ger@capgemini.com

**Niharika Kalvagunta**
niharika.kalvagunta@capgemini.com

**Johan Müllern-Aspegren**
johan.mullern-aspegren@capgemini.com

**Henk Vermeulen**
henk.vermeulen@capgemini.com

**James Wilson**
james.l.wilson@capgemini.com

## About Capgemini

Capgemini is a global business and technology transformation partner, helping organizations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 340,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fueled by its market leading capabilities in AI, generative AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2024 global revenues of €22.1 billion.

**Get the future you want | www.capgemini.com**

Capgemini