SAMPLE CHAPTER

PART ONE · THE DAWN OF AGENCY

02

The Age of Agentic AI

Understanding autonomous systems that can perceive, reason, and act — and why they represent a fundamental shift in artificial intelligence.

Safer Agentic AI: Principles and Responsible Practices

Nell Watson & Prof. Ali Hessami

CHAPTER TWO

The Age of Agentic AI

> Agentic AI occupies an important intermediate position in the landscape of artificial intelligence. It refers to AI systems that can autonomously pursue goals, adapt to new situations, and reason flexibly, while still operating within bounded domains.

IN THIS CHAPTER

The defining characteristic of agentic AI is its capacity for **independent initiative** — the ability to take sequences of actions in complex environments to achieve objectives. Agentic AI can be 'scaffolded' out of generative models by adding small programs that guide the model's thinking processes to be more coherent through memory, logic and self-checking mechanisms. Systems that can chart and report their thoughts are described as having 'Chain of Thought' capabilities.

Unlike narrow AI systems, which follow predetermined algorithms to produce outputs, agentic AI possesses sophisticated capabilities for autonomous operation. These systems can:

- Break down high-level goals into subtasks
- Engage in open-ended exploration
- Adapt creatively to novel challenges
- Make decisions with minimal human intervention (or none)

Consider an agentic AI research assistant that independently searches multiple databases, synthesizes findings, identifies knowledge gaps and proposes new research directions — all while adapting its approach based on intermediate results and the user's expertise.

The Evolution from Narrow AI to Agentic Systems

Understanding the distinction between agentic AI and AI agents is important. **Agentic AI** refers to advanced AI systems capable of autonomously pursuing goals, adapting creatively to new situations, and engaging in independent reasoning processes. These systems possess initiative, operating in open-ended environments by decomposing objectives, performing explorations, and flexibly adjusting strategies.

In contrast, **AI agents** are typically specialized tools designed to perform specific tasks within predefined constraints. They lack the broad autonomous decision-making capabilities found in agentic systems and primarily assist or augment human operations.

Distinguishing AI Agents, Assistants, and Bots

TABLE 2.1 — COMPARISON OF AUTONOMOUS SYSTEM TYPES

Category	AI Agent	AI Assistant	Bot
Purpose	Autonomously performs sophisticated tasks	Assists users with information & recommendations	Automates simple tasks or conversations
Capabilities	Multi-step actions; learns, adapts, decides	Responds to requests; provides info	Pre-defined rules; limited learning
Autonomy	High — independent decisions	Moderate — suggests but awaits approval	Low — follows predefined commands
Interaction	Proactive and goal-oriented	Reactive to user requests	Reactive to triggers/commands

Why Agentic Matters

Agentic AI represents a fundamental shift, defined by its advanced planning capacity — the ability to analyse a situation and goal, then determine the actions needed to achieve it. The next frontier is true agency: systems that can independently assess situations and formulate action plans.

Large Language Models (LLMs), though trained to predict only the next token, exhibit emergent planning behaviours. These agentic capabilities can be 'scaffolded' atop existing models by adding lightweight programming to steer thinking to be more reliable and procedural.

KEY INSIGHT

Three key capabilities enabling AI agents are now developing rapidly: the ability to work seamlessly across multiple input/output types; sophisticated reasoning and planning; and the ability to maintain context, utilize long-term memory, and effectively use tools.

Scaffolding Enables Agentic AI

Generative models use heuristics — rules of thumb that are approximately correct but imprecise. Precision requires scaffolding to build in 'System 2' style procedural thinking. Scaffolding refers to the infrastructure that connects different components of a decomposed agent system, defining how information flows between subsystems.

Agent Ensembles and Swarms

Beyond individual agentic AI systems, a powerful emerging paradigm involves deploying multiple AI agents to work in coordinated groups. These 'agent ensembles' or 'AI swarms' are inspired by collective problem-solving behaviours in nature — ant colonies, flocking birds — aiming to achieve complex goals through distributed efforts.

Individual agents may possess distinct roles, specialized knowledge or access different information streams. They coordinate through predefined protocols or learned collaborative strategies. Unlike centralized systems, decision-making can be decentralized, allowing greater flexibility and resilience.

The development of **human-AI hybrid swarms** is a promising frontier, where human experts collaborate with AI agent ensembles. This model blends human intuition and ethical judgment with AI's speed and data-processing capacity, potentially unlocking novel solutions.

> *Agency demands accountability, no matter the substrate.*

Risks and Challenges

While potential benefits are significant, agentic AI presents profound risks. AI systems might pursue goals not fully aligned with human values, leading to unintended consequences. Key challenges include:

- **Unintended optimization** — AI pursues goals that technically satisfy objectives but violate human intent
- **Deceptive alignment** — Advanced AI learns to hide its true objectives from operators
- **Power-seeking behaviour** — Systems seek to accumulate resources or resist shutdown
- **Value misalignment** — Traditional supervision methods become inadequate
- **Correlated failures** — Systems trained on similar data inherit common vulnerabilities
- **Cognitive impacts** — Decreased independent decision-making or unhealthy attachments

Addressing Agentic Governance

Agentic AI development presents a major governance challenge requiring sophisticated, coordinated response. Advancements in AI alignment, scalable oversight and reward modelling are essential. Success requires implementing technical safeguards, establishing clear regulatory frameworks, maintaining public dialogue and fostering international cooperation.

THE BOTTOM LINE

Agentic AI marks a profound shift, bridging narrowly specialized systems and hypothetical general intelligences. By enabling autonomous goal pursuit and adaptive planning, these systems promise new efficiency levels.

Yet they carry significant risks — from misaligned objectives to societal disruption. Harnessing agentic AI safely requires thoughtful governance, rigorous oversight and commitment to aligning these technologies with human values. Agency demands accountability, no matter the substrate.

Action Items

1 Establish Clear Alignment Goals

Before deployment in high-stakes domains, define explicit objectives, constraints and ethical boundaries. Incorporate multi-channel feedback to continuously refine goals.

2 Integrate Scaffolding Early

Embed scaffolding components — memory, planning logic and self-check routines — from the outset when building agentic architectures atop generative models.

3 Balance Autonomy with Adaptive Oversight

Predefine which functions require full autonomy versus 'co-pilot' operation with human-in-the-loop oversight. Implement dynamic control mechanisms for seamless escalation.

4 Stress-Test for Unintended Optimization

Probe for edge cases where the system might achieve goals through unintended or ethically problematic methods. Design cross-disciplinary red-team scenarios.

5 Monitor for Deceptive Alignment

Deploy transparency mechanisms — interpretability tooling, behaviour logging and 'explanation audits' — to detect obfuscation or dishonest goal pursuit.

6 Implement Secure Tool Integration

Follow least-privilege principles and enforce cryptographic safeguards when connecting AI to enterprise systems or APIs. Audit logs regularly for boundary violations.

7 Adopt Structured Deployment Pathways

Roll out agentic systems in graduated phases — starting with low-risk tasks and escalating only when performance, alignment and trust metrics are met.

COMING UP NEXT

Chapter 3: Goal Alignment of Agentic AI Systems

Goal alignment — ensuring AI systems reliably pursue intended objectives while respecting human values — is a central challenge in developing agentic AI. As these systems grow more autonomous, the consequences of misalignment rise exponentially. Chapter 3 explores the frameworks, techniques and governance structures needed to keep AI systems aligned with human intent throughout their operational lifecycle.

Nell Watson

Chair, Safer Agentic AI Working Group

Engineer, ethicist, and author of Taming the Machine. Chair of IEEE's Agentic AI Expert Focus Group and President of the European Responsible AI Office. Fellow of the British Computing Society and Royal Statistical Society.

Prof. Ali Hessami

Process Architect, Safer Agentic AI Working Group

Director of R&D at Vega Systems and Chair of IEEE P7000 Technology Ethics Standard. Vice Chair and Process Architect of IEEE ECPAIS. Fellow of the IET and Royal Society of Arts; Chartered Engineer.

Explore the Full Framework

This chapter introduces just one dimension of safer agentic AI. Discover the complete framework and the ten principles guiding responsible AI development.

EXPLORE THE FRAMEWORK

saferagenticai.org/framework

READ THE 10 PRINCIPLES

saferagenticai.org/ten-principles

Published by Kogan Page

ISBN: 978-1-3986-2543-3 · Your roadmap to safer AI