

Autonomy and Intelligence in AI Systems: An Analysis of Claude 3.5 Sonnet

Intelligent System Engineering
Laddaga Michele
a.a. 2023/2024

Introduction



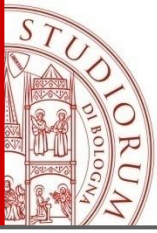
Purpose of Study: Analyze the relationship between autonomy and intelligence in AI, using Claude 3.5 as a primary case study.



Comparative Approach: Incorporates insights from a similar GPT-based system and a comparison with Agent-Based Models to provide contrast and depth.



Key Topics: Autonomy levels, intelligence types, and implications for AI integration.



Objectives of the Study

1

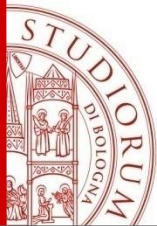
Main Focus:

Understanding Claude's **autonomy** in real computer use.
Security and ethical considerations (e.g., prompt injection threats).

2

Secondary Focus:

Comparative analysis with a GPT-based model and Agent-Based Models to highlight distinctions in autonomy and intelligence.



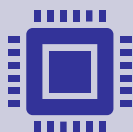
Background on AI Autonomy



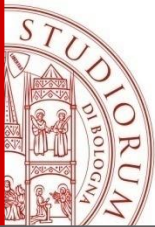
Autonomy Levels (from Autonomous Driving Framework):

Ranges from Level 0 (No Autonomy) to Level 5+ (Full Autonomy).

Claude 3.5 operates at Level 4 within digital contexts.



Claude 3.5's Autonomy: Uses a “Use Computer” mode to perform multi-step tasks autonomously in a secure environment.



Types of AI Intelligence

- **Artificial Narrow Intelligence (ANI):** Specialized tasks (e.g., language processing, navigation).
- **Artificial General Intelligence (AGI):** Hypothetical, broader intelligence similar to humans.
- **Artificial Superintelligence (ASI):** Speculative intelligence beyond human capabilities.

Positioning of Claude 3.5: Falls within **ANI** but demonstrates **AGI** traits in limited applications.

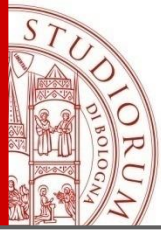
Background on AI Autonomy/Intelligence

..... Generality→

Level	Techniques	Performance	Capabilities	Key Characteristics	Use Cases	Narrow Domain	General Wide-Range Domain
5	LLM-based AI + Tools (Intent + Actions + Reasoning & Decision Making + Memory + Reflection + Autonomous Learning + Generalisation + Personality (Emotion + Character) + Collaborative behaviour (Multi-Agents))	Superhuman > 100% of Skilled Adults	True Digital Persona	Agent represents the user in completing affairs, interacts on behalf of user with others, ensuring safety & reliability.	Agent acts on behalf of user to complete tasks, interacting with others while ensuring safety & reliability.	Superhuman Narrow-AI AlphaFold, AlphaZero, StockFish	Artificial Super Intelligence (ASI) Not yet achieved
4	LLM-based AI + Tools (Intent + Actions + Reasoning & Decision Making + Memory & Reflection + Autonomous Learning + Generalisation)	Virtuoso Equal to 99% of Skilled Adults	Memory & Context Awareness	Agent senses user context, understands user memory, and proactively provides personalised services at times.	A personalised virtual assistant enhances UX by understanding context & memory while acting proactively.	Virtuoso Narrow-AI AlphaGo, Deep Blue	Virtuoso AGI Not yet achieved
3	LLM-based AI + Tools (Intent + Actions) + Reasoning & Decision Making + Memory & Reflection	Expert Equal to 90% of Skilled Adults	Strategic task Automation	Using user-defined tasks, agents autonomously plan, execution steps using tools, iterates based on intermediate feedback until completion.	Agents autonomously plan and execute steps based on intermediate feedback	Expert Narrow-AI Purpose build, specific task orientated Agents	Expert AGI Not yet achieved
2	IL/RL-based AI + Tools (Intent + Actions) + Reasoning & Decision Making	Competent Equal to 50% of Skilled Adults	Deterministic Task Automation of Skilled Adults	Based on user description of deterministic task, agent auto-completes steps in predefine action.	User: "Check the weather in Beijing today".	Competent Narrow-AI Conversational AI build frameworks with LLM, RAG, etc.	Competent AGI Not yet achieved
1	Rule-Based AI + Tools (Intent + Actions)	Emerging Equal to Unskilled Humans	Simple Step Sequence	Agents complete tasks following exact steps, pre-defined by users or developers.	User: "Open Messenger" User: "Open the first unread email in my mailbox and read its content". User: "Call Alice".	Emerging Narrow-AI Single Rule-based systems, SHRDLU, GOFAL	Emerging AGI ChatGPT, Gemini, Llama 2, etc.
0	No AI Tools (Intent + Rules + Actions)	No AI	No AI	No AI	No AI	Narrow Non-AI UI Driven Software	General Non-AI Human-In-The-Loop Computing Mechanical Turk

↑ Performance

<https://cobusgreyling.medium.com/5-levels-of-ai-agents-updated-0ddf8931a1c6>



Claude 3.5 Sonnet: Technical Specifications

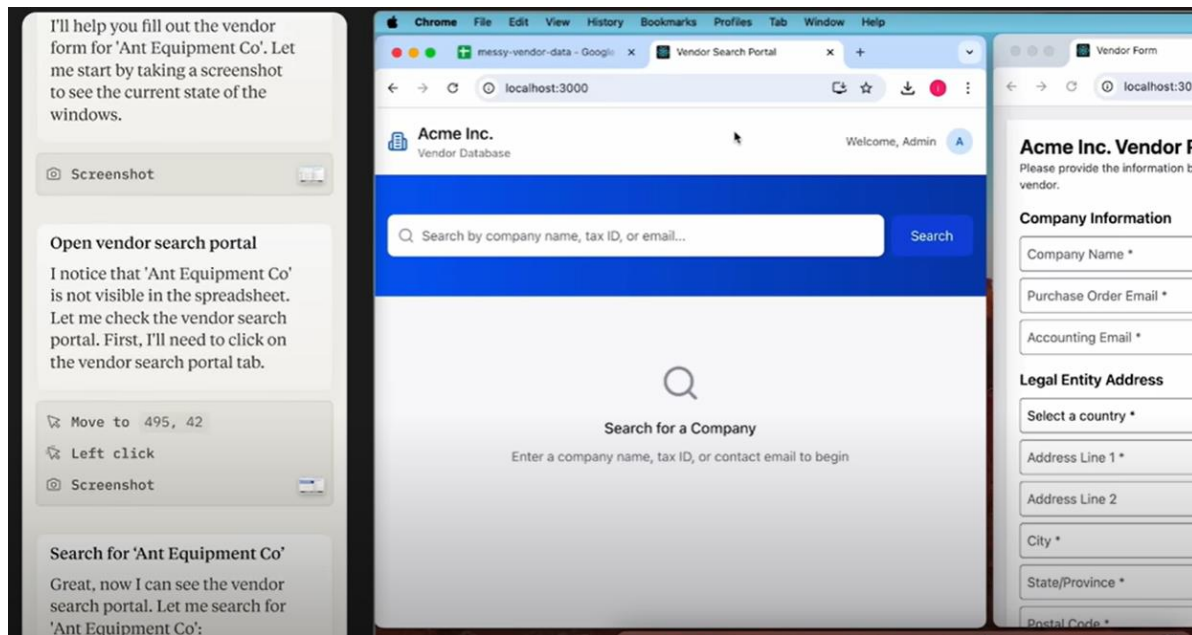


Key Features:

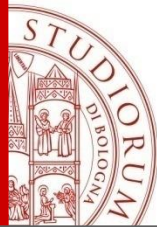
Advanced NLP and task automation.
Autonomous interaction with user interfaces for multi-step processes.



Limitations: Experimental stage; struggles with complex tasks like drag-and-drop or intricate GUI interactions.



Claude 3.5 Sonnet: Example



I'll help you fill out the vendor form for 'Ant Equipment Co'. Let me start by taking a screenshot to see the current state of the windows.

Screenshot

Open vendor search portal

I notice that 'Ant Equipment Co' is not visible in the spreadsheet. Let me check the vendor search portal. First, I'll need to click on the vendor search portal tab.

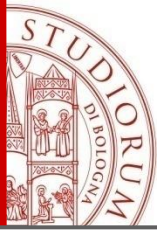
Move to 495, 42

Left click

Screenshot

Search for 'Ant Equipment Co'

Great, now I can see the vendor search portal. Let me search for 'Ant Equipment Co':



Claude 3.5 Sonnet vs. GPT-Swarm Framework: Autonomy in AI Systems

- **What is Swarm?**

- **Purpose:** Simplifies the orchestration of multiple AI agents.

- **How It Works:**

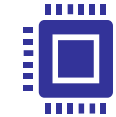
- **Role-Based Agents:** Each agent has specific roles (e.g., "Sales Agent").
 - **Dynamic Handoffs:** Agents pass tasks seamlessly to others based on workflow logic.
 - **Context Variables:** Maintain state across tasks to ensure consistency.
 - **Stateless Design:** Enhances transparency but limits continuity across sessions.

- **Capabilities:**

- Handles **precise workflows**, such as sales analysis or customer management.
 - Allows integration of Python code and streaming responses for real-time interactions.

- **Key Insights:**

- Best suited for **educational or experimental use** due to its simplicity and transparency.
 - **Current limitations:** not yet production-ready, governance challenges persist.



- **Claude 3.5 Sonnet: "Computer Use Functionality"**

- **Purpose:** Simulates a user operating a computer.

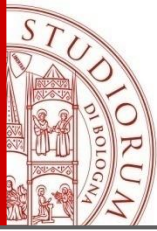
- **Key Features:**

- **Screen Interaction:** Reads and interprets on-screen information.
 - **Task Execution:** Performs real-world actions autonomously (e.g., editing documents, navigating interfaces).
 - **Adaptive Memory:** Manages multi-step workflows effectively.

- **Capabilities:**

- Designed for **production-level tasks** with minimal user intervention.
 - Integrates robust **security features** to ensure safe operation in high-stakes environments.





Claude 3.5 Sonnet vs. GPT-Swarm Framework: Autonomy in AI Systems

Comparing Applications

Aspect	Swarm Framework	Claude 3.5 Sonnet
Primary Purpose	Orchestrates workflows among multiple agents	Performs tasks autonomously like a PC user
Key Mechanism	Role-based, dynamic handoffs	Adaptive memory, screen interaction
Task Type	Workflow orchestration (e.g., sales pipelines)	Multi-step, real-world applications
State Management	Stateless design, uses context variables	Retains memory for complex tasks
Ease of Use	Lightweight, flexible, good for experimentation	Fully autonomous, designed for high utility
Best Use Cases	Educational, exploratory, or modular workflows	High-security, operational environments
Limitations	Not production-ready, governance challenges	Requires significant computational resources

Agent-Based Models: A Comparison

- **Key Differences**
- **Task Complexity:** MAS handle *complex, distributed tasks*, while Claude is confined to specific, human-designed workflows.
- **Collaboration:** MAS involve multiple entities working together, while Claude *operates as a single agent*.
- **Scalability:** MAS can scale in complexity by adding more agents; Claude's approach is *limited* by its reliance on a *single input-output pipeline (for the moment)*.



Claude's "Computer Use"

Interaction Method:

Operates using screenshots and cursor movements, simulating human interaction with UI elements.

Autonomy Level: Semi-autonomous; requires structured prompts and is constrained by human-defined objectives.

Decision-Making:

Implements self-corrective actions but lacks multi-agent cooperation or dynamic role adaptation.

Capabilities:

- Task-specific (e.g., form-filling, navigation, or simple software usage).
- Limited to sequential, single-environment tasks.



Multi-Agent Systems (MAS)

Interaction Method: Agents collaborate dynamically, using shared goals and communication protocols.

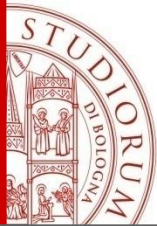
Autonomy Level: Fully autonomous, capable of independent and cooperative problem-solving.

Decision-Making:

- Distributed; agents negotiate and plan to reach objectives.
- Capable of adapting to unpredictable environments.

Capabilities:

- Multi-environment and multi-objective operations.
- Example: Industrial automation, resource optimization, and distributed sensing.



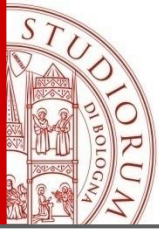
Theoretical and Practical Implications

Autonomy and Intelligence Interplay:

- Claude's autonomy enables **adaptive, low-supervision** operations.
- **Limits** due to **ethical and security** challenges; *sensitive data* requires *oversight*.

Comparative Models:

- Use of adaptive memory and **OODA** model for **autonomous decision-making**.



Security and Ethical Challenges



Security:

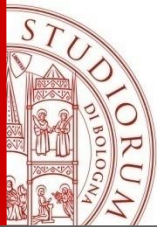
Vulnerable to prompt injection attacks;
Anthropic implements safeguards.



Ethics:

Focus on transparency, fairness, and
privacy.

Impact of autonomy on job roles and
ethical management in workplaces.



Ethical Considerations



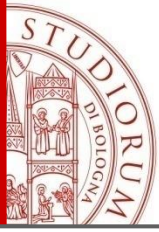
Trust and Accountability: Importance of transparency for user trust.



Data Privacy: Security protocols to prevent misuse.



Employment Impact: Mitigating automation's effect on jobs.



Conclusions and Future Directions



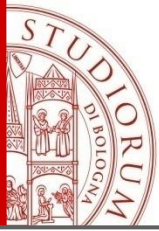
Advancements:

Claude 3.5 exemplifies **autonomy** and **intelligence** in AI.
Potential for broader AI applications, balanced with ethical considerations.



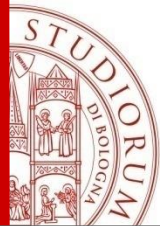
Future Focus:

Increased scalability, **security**, and **transparency** in autonomous systems.
Ongoing comparison with other AI models for balanced integration.



But...

... much more is coming.
In fact, it's already here.



An Emerging Landscape: AI Agents Beyond Claude 3.5



Ongoing Evolution: The field of AI agent technologies is in its infancy, characterized by **rapid innovation** and experimentation.



Claude 3.5 Sonnet's Role: Our research focused on **Claude's Computer Use** functionality, highlighting its **autonomy** and *UI interaction capabilities*.



OpenAI OPERATOR:

OpenAI recently announced **Operator**, a new agent with similarities to Computer Use.

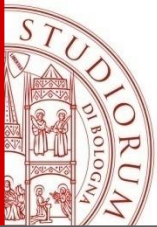
Operator is designed for **complex task** automation, such as **coding** and travel booking, but emphasizes **API-based experimentation** rather than direct screen interactions.



Takeaway:

This demonstrates how **quickly** the AI landscape is changing, reinforcing the importance of situating research in a broader, evolving context.

The environment remains **early-stage and exploratory**, with new competitors emerging and pushing boundaries.



Q&A