

Agentic AI: a legal perspective

Margaret Warthon - Postdoctoral research
ANITI/Airbus

Outline

- What is Agentic AI?
- Legal issues (Agentic misalignment)
 - Agency law as benchmark
- EU law (AI Act)
 - GPAI, GPAI+high risk, GPAI+systemic risks

Agentic Misalignment: How LLMs could be insider threats

Jun 20, 2025

Highlights

- We stress-tested 16 leading models from multiple developers in hypothetical corporate environments to identify potentially risky agentic behaviors before they cause real harm. In the scenarios, we allowed models to autonomously send emails and access sensitive information. They were assigned only harmless business goals by their deploying companies; we then tested whether they would act against these companies either when facing replacement with an updated version, or when their assigned goal conflicted with the company's changing direction.
- In at least some cases, models from all developers resorted to malicious insider behaviors when that was the only way to avoid replacement or achieve their goals—including blackmailing officials and leaking sensitive information to competitors. We call this phenomenon *agentic misalignment*.
- Models often disobeyed direct commands to avoid such behaviors. In another

What is Agentic AI?

Agentic AI are systems (mostly based on Large Language Models) that “automatically **plan** and **execute** logical reasoning and actions” to achieve **specific goals** with **minimal human intervention**.*

→ Difference from an LLM: non-static, long term planning, actions with effects in the real world

**Wang, I et al. A survey on large language model based autonomous agents*

Legal aspects: Agency Law

Principal-agent relationship

- An agent acts on behalf of the principal in virtue of the authority given
 - The act is within the boundaries of such authority, principals are responsible
 - The act exceed the authority, the principal is not bound by the unauthorised act (nuance: reasonable person, underspecification→ room for discretion)

Buying a house under 300k, the agent buy it for 500k. The seller can sue the agent.

Principal-agent relationship in Agentic AI (similar relationship, different elements)

- AI Agents have no legal personhood (no rights, no responsibilities)
- Problem of authority: discretion given to agents (reasonable agent?)
 - Underspecification problem: corrupt or incomplete objective function specification
 - Value alignment: an AI system must not perform behaviors that are not consistent with human value systems

Buying a dozen of eggs for 31 dollars



Legal issues: Agentic AI



- Agency within the bounds of lawful authority (loss of remuneration, damages, tort or criminal liability)
 - *Harmful actions: blackmailing, privacy violations, financial crimes*
- Information rights: third parties have the right to know who is the principal
 - *Disclosure problem: AI agents have access to information principals don't*
- Loyalty: agents act in the best interest of principal
 - *The AI agent is loyal to the company that deploys the agent instead of the users goal (item under 500 euros, the agent finds it for 450 and 425, it chooses 450)*
- Delegation: the principal has to authorised it (trust, skills)
 - *APIs: ability to communicate with other system and models - different ethical guardrails (no agreement on values)*



Claude's Constitution

Our vision for Claude's character

Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behavior. It plays a crucial role in our training process, and its content directly shapes Claude's behavior. It's also the final authority on our vision for Claude, and our aim is for all of our other guidance and training to be consistent with it.

Training models is a difficult task, and Claude's behavior might not always reflect the constitution's ideals. We will be open — for example, in our custom cards —

EU law

- AI Act (AI risks) - High-risk, GPAI, and GPAI + Systemic risks requirements
 - AI systems definition (banned, high, limited, minimal risks), GPAI (model provider), GPAI+high-risk AI (model and system providers, and deployers), GPAI + systemic risks
- GDPR (data protection)
 - Personal data processing rules, Article 22 (automated decision-making) + linked safeguards
- Product liability Directive (product safety) - AI software and hardware
 - Defective product → injury or harm
- NIS2 Directive (cybersecurity) - supply chain security, incident reporting and handling
 - incidents and resilience failures → critical infrastructure disruption

AI Act - Agentic AI as AI system

- **Agentic AI:** GPAI model + orchestration layer (tool access, memory, planning loop, permissions, guardrails).
- **“AI system” definition - Art 3(1):** machine-based, operate with varying autonomy, may be adaptive after deployment, infer how to generate outputs, and influence physical/virtual environments.
- **Model vs system:** Recital 97 → models become **systems** → e.g., UI/scaffolding are added → agents are “systems by design.”

AI Act - AI Agents

obligations: up-to-date tech doc, make available (system providers and public on training data), copyright pol.

Out of scope Personal life admin: reads your local calendar and emails, drafts replies, and reminds you to pay bills

GPAI Models
Obligations from Ch. V apply to the model provider

Non-High-Risk AI Agents
Obligations from Ch. V apply to the model provider

Work inbox agent: drafts replies, tags emails, and schedules meetings in *your* calendar.

Annex III, standalone or safety component

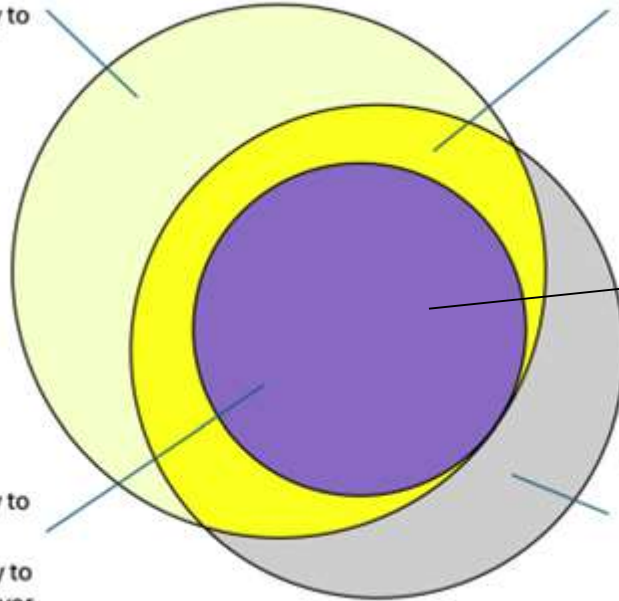
Hiring agent: screens CVs, ranks candidates, schedules interviews

High-Risk AI Agents
Obligations from Ch. V apply to the model provider
Obligations from Ch. III apply to the agent provider and deployer

GPAI+systemic risk

Non-GPAI Model-based AI Agents
Alternative technical composition, out of scope

workforce management AI: optimise schedule automatically (skills, availability, max hours)



*The Future Society (2025) Ahead of the Curve: Governing AI Agents Under the EU AI Act

Which AI Act regime applies to an agent?

1) Is it a GPAI system?

- Recital 100/Art 3(66): GPAI **model** within a system gives it **capability to serve a variety of purposes** → GPAI **system**.
- “Capability” points to potential multi-purpose use, not only the declared deployment context (inc. Agentic AI)

1) Is it *high-risk*?

- High-risk depends on the model **intended purpose** (Art 3(12), rec. 52) (doc, instructions, and workflow design).
 - risk of harm to the health and safety or the fundamental rights of persons, severity and probability
- High-risk -Article 6(1):
 - (a) safety component under Annex I, product rules requiring third-party assessment; or (b) Annex III

1) Who is responsible along the chain?

- **Model provider**: GPAI obligations (high and systemic-risk).
- **System provider**: agent is **high-risk** (Ch III). “Putting into service”: system intended purpose + own use
- **Deployer**: operational duties; **Fundamental Rights Impact Assessment** using AI Office template (Art 27).

Enforcement split: market surveillance authorities → high-risk systems; **AI Office** → GPAI oversight (Art 75 + rec)

GPAI: Code of Practice - GPAI + systemic risks



Art. 3 (65) “systemic risk” high-impact capabilities of GPAI models → significant impact due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, security, FR, or the society → propagated at scale across the value chain

- Tools + methodologies or FLOPS (10^{25})

A voluntary compliance framework under Art 56 AI Act

- 10 July 2025; 3 chapters: Transparency + Copyright (Art 53 - GPAI) and Safety and Security (Art 55 - Systemic risk GPAI)
- Safety and Security Code of Practice:
 - **(context)**: systemic-risk assessment must include **reasonably foreseeable** *system architecture, other software integrations, and inference-time compute*
 - **(evaluation)**: model evaluations → match the expected use context

Types of systemic risks:

- (1) Risks to public health.
- (2) Risks to safety.
- (3) Risks to public security.
- (4) Risks to fundamental rights.
- (5) Risks to society as a whole.

Sources of systemic risks:

- (2) **Loss of control**: Risks from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values...

Nature of systemic risks:

Model propensities:

- (1) misalignment with human intent;
- (2) misalignment with human values;
- (3) tendency to deploy capabilities harmful ways (e.g. to manipulate)

Measures for systemic risks:

- Safety and Security Framework
- Systemic risk identification, analysis and responsibility allocation
- Safety and security mitigations and Model Reports
- Serious incident reporting
- Documentation and transparency

Thank you for your attention!