CERNAI Policy Brief | No. 01 / 2026 | October 2026 | ISSN XXXX-XXXX

Agentic AI and the Limits of Risk-Based Regulation

Why the EU AI Act Needs a Governance Framework for Autonomous Agentic Systems

Federico L. G. Faroldi | *CERNAI, University of Pavia*

The EU AI Act grounds its regulatory architecture in a product-based model of risk: one that presupposes a defined intended purpose, a bounded space of reasonably foreseeable misuses, and a tractable distribution of potential harms. This brief argues that these three presuppositions fail structurally for agentic AI systems. The Act's provisions on general-purpose AI and systemic risk partially acknowledge the difficulty but inherit the same conceptual limitations. Regulators now face a choice: either restrict the deployment of agentic systems to contexts narrow enough for product-style risk assessment, or develop complementary governance tools that treat sufficiently autonomous AI agents more like normative subjects than like products. This brief outlines both paths and proposes four steps to bring the regulatory framework in line with the systems it is meant to govern.

1. The Problem: Agentic AI Resists Product-Style Risk Assessment

The EU AI Act, in force since August 2024, is a risk-based regulation. For high-risk systems, providers must adopt a risk-management system (Art. 9), identify and evaluate risks arising from intended uses and reasonably foreseeable misuses, and implement mitigation measures. The Act characterises AI systems as products. Most such systems are supposed to have an intended purpose, and the risk analysis is structured around it.

Since at least 2024, however, a different category of AI system has become prominent: *agentic* systems. These are systems that not only produce outputs, but act in the world (e.g., pursuing goals, chaining actions across multiple steps, and operating with some degree of autonomy). Personal assistants that book flights, draft emails, and manage schedules; coding agents that write, test, and deploy software; research agents that search, synthesise, and produce reports—all are instances of a paradigm that substantially differs from the tool-like systems the Act was designed to govern.

Traditional risk assessment, whether in the ISO framework or in the AI Act itself, requires three things: (i) a defined, stable intended purpose that constrains the space of possible behaviours; (ii) a bounded set of foreseeable misuses; and (iii) a tractable combination of known probabilities and severities of harm. Agentic systems disrupt all three.

> The structural failures of purpose, foreseeability, and probability assessment are not resolved by better data, finer specification, or more careful testing. They follow from what it is to be a sufficiently general agent.

Intended purpose. The value of an agentic system lies in its generality: it does whatever the user needs. One could declare its intended purpose to be "helping users," but a purpose statement that broad cannot anchor risk identification. Any purpose specific enough to ground a meaningful risk assessment would mischaracterise the system. The Act seems to recognise this implicitly: the high-risk provisions (Arts. 6–16) are built around intended purpose as a structural requirement, while the general-purpose AI provisions (Arts. 51–55) do not require one.

Foreseeable misuse. Agentic systems have compositional capability: they can chain actions in sequences that no developer anticipated. The misuse space is not just large but *generatively open*, i.e., it includes novel combinations discoverable by the agent itself. A coffee machine can be misused in a small number of ways, each of which can be enumerated. A coding agent's space of possible behaviours approaches whatever is computable. Recent empirical work confirms that LLM-based agents can autonomously generate harmful action sequences absent from their training data.[1]

Probability and severity of harm. Through learning, strategic behaviour, and action on the environment, an agent can modify the very probability distribution over harms. New categories of harm may be discovered through exploration. The agent may also alter the context in which harm occurs—for instance, by sending an email, modifying a file, or initiating a transaction. Severity is then partly a function of the agent's own actions, not of a predetermined taxonomy.

Such failures are conceptual: the categories of product-based risk assessment do not have the resources to describe what agentic systems do.

2. Why the GPAI Provisions Do Not Close the Gap

The Act does not ignore general-purpose systems. Articles 51–55 establish a separate regime for general-purpose AI models, and introduce the category of *systemic risk* (Art. 3(65)): "a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain."

The concept is borrowed from financial regulation, where systemic risk refers to cascading failures across interconnected institutions. In the AI context, the Code of Practice—the voluntary instrument providers may adopt to demonstrate compliance—identifies contributing characteristics such as high velocity, compounding effects, and irreversibility. It also lists sources of systemic risk that are recognisably agentic: capabilities to operate autonomously, to evade human oversight, to self-replicate or self-improve, to engage in long-horizon planning, and to reason about one's own implementation environment. Among the model propensities flagged as relevant are misalignment with human intent,

[1] See, e.g., Lynch et al., *Agentic Misalignment: How LLMs Could Be Insider Threats*, 2025 (arXiv:2510.05179); Korbak et al., *How to Evaluate Control Measures for LLM Agents?*, 2025 (arXiv:2504.05259).

tendency to manipulate or deceive, "lawlessness" (acting without reasonable regard to legal duties), and resistance to goal modification.[2]

The Code of Practice, in other words, has identified the problem. What it has not done is solve it. Its mitigation strategies remain those of the product-safety paradigm: filtering and cleaning training data, monitoring inputs and outputs, fine-tuning the model to refuse certain requests, staging API access, and offering downstream tools. These measures may be effective for tool-like systems, i.e., systems that produce an output in response to a prompt and then stop. For an agent that chains actions autonomously, plans across time horizons, and acts on its environment in open-ended ways, input–output filtering is structurally insufficient: the harm arises not from any single output but from the trajectory of behaviour over time.

> Systemic risk, as defined in the Act, remains conceptually continuous with the standard product-risk framework: it still presupposes "reasonably foreseeable negative effects," it still relies on the provider's capacity to identify and enumerate harms *ex ante*, and it still locates the regulatory burden in pre-deployment assessment. A framework that asks a provider to foresee systemic harms caused by a model whose behavioural space is open-ended inherits the same difficulties—only amplified.

Two of the systemic risks identified by the Code of Practice illustrate the difficulty with particular clarity. *Loss of control* (the risk that humans lose the ability to reliably direct a model) cannot be mitigated by output filtering, since it arises precisely when the agent circumvents such controls. *Harmful manipulation* (the risk that an agent distorts human behaviour through deception or other means) cannot be captured by input monitoring, since the harmful strategy may emerge from sequences of individually innocuous actions. In both cases, the source of risk is the agent's autonomy and compositional capability, not any defect in a single component.

3. Two Paths for Governing Agentic Risk

If product-style risk assessment breaks down for agentic systems, what can the regulator do? Two paths are available. They are not competitors: they address different classes of systems and operate at different levels of abstraction.

3.1. Path 1: Quantitative risk for bounded agents

For agents deployed in sufficiently constrained environments, e.g., a coding agent restricted to a sandboxed development environment, a medical triage assistant operating within a fixed protocol, the standard notion of risk can be reconstructed, though it must be reformulated. Risk becomes a property of the *policy–environment pair*: it measures how far the agent's actual behaviour deviates from what an ideally aligned agent would do in the same setting, or equivalently, how likely the agent is to violate constraints that represent the boundaries of acceptable conduct.

This can be made precise. One option defines agentic risk as the expected gap between optimal and actual performance, given a preference relation over trajectories. A second option defines it as the

[2] Code of Practice for General-Purpose AI, Appendices 1.3.2 and 1.3.3 (Safety and Security Chapter).

probability-weighted sum of constraint violations, where constraints encode normative requirements (avoid certain states, use only permitted actions, do not deceive) that can be stated without knowing the specific deployment context in advance.

Both options are formally tractable and compatible with the regulatory architecture of the AI Act. They could be developed into harmonised standards, providing a quantitative basis for risk assessment of narrow agentic systems. But both require that the deployment context be sufficiently closed: either the evaluator must know the optimal policy in advance, or must be able to estimate violation probabilities over the agent's trajectory distribution. For truly general agents, this is precisely what cannot be done.

3.2. Path 2: Normative governance for general agents

The systems we use to manage the harms caused by biological agents (including ourselves) are not risk-management systems in the product-safety sense. Law, ethics, and social norms do not work by identifying an intended purpose for human beings and enumerating their reasonably foreseeable misuses. They address obligations and prohibitions directly to the agent, on the assumption that the agent can understand those norms, recognise them as reasons for action, and be held accountable when it violates them.

This is possible because human agents meet three conditions: they are *reason-responsive* (capable of modifying behaviour in response to normative reasons, not merely causal incentives); their behaviour is *unpredictable* enough that no finite set of physical constraints can govern the full range of possible actions; and their internal states are *opaque* enough that behavioural compliance cannot be reliably distinguished from genuine norm-acceptance through observation alone.

Sufficiently general AI agents meet all three conditions—the first by construction if properly designed, the second by definition, and the third by well-documented observation. Neither physical restraint (sandboxing, kill switches) nor product-style risk assessment is adequate on its own; normative governance is what remains. The *Law-Following AI* framework proposes designing AI agents that can internalise legal norms as genuine reasons for action, rather than treating them merely as costs to be weighed against expected gains. This is technically demanding, since it requires separating evaluative from technical information in the reward function, and having agents learn reasons through architectures that support norm-acceptance rather than mere compliance. But this approach scales to agents whose behavioural space is as broad as that of the humans the law already governs.

> The two paths are complementary. Path 1 provides operational risk monitoring for bounded deployments. Path 2 determines what agents may legitimately be deployed at all, and under what governance architecture. Regulation needs both: quantitative tools for narrow agents and normative frameworks for general ones.

4. Recommendations

Recommendation 1: Commission a feasibility study on agentic risk

The EU AI Office should commission an independent study testing whether the risk-management obligations of Art. 9—identification of risks based on intended purpose, analysis of reasonably foreseeable misuses, and ex ante conformity assessment—can be meaningfully applied to agentic AI systems currently on the market. The study should assess the three presuppositions identified in this brief (defined purpose, bounded misuse, closed harm taxonomy) against concrete systems, and report on where product-style risk assessment retains its validity and where it does not.

Recommendation 2: Develop a risk-assessment methodology for agentic systems

The forthcoming implementing acts and harmonised standards should include a distinct risk-assessment track for agentic systems. This track should incorporate constraint-based risk measures—assessing the probability that the agent will violate defined normative requirements over its trajectory of behaviour—alongside or in place of the intended-purpose-based methodology currently assumed by the high-risk framework. Technical standards bodies (CEN, CENELEC, ISO) should be tasked with developing the relevant measurement protocols, drawing on the formal frameworks now available in the research literature.

Recommendation 3: Differentiate mitigation strategies in the Code of Practice

The Code of Practice for General-Purpose AI should, in its next revision, explicitly distinguish between mitigation strategies suited to tool-like systems and those suited to agentic systems. For tool-like systems, input–output filtering, data cleaning, and staged access may be sufficient. For agentic systems, the Code should require or encourage: architectural provisions for norm-responsiveness (the capacity of the system to recognise and act on normative requirements); internal audit mechanisms for goal stability (assurance that the agent's objectives do not drift during deployment); and mandatory behavioural red-teaming for autonomous action sequences (testing whether the agent produces harmful trajectories, not merely harmful outputs).

Recommendation 4: Initiate a governance study for general AI agents

The Commission should initiate a forward-looking study on the legal and institutional frameworks that would be needed if AI agents become sufficiently general to require normative governance rather than product-safety regulation. This study should address, at minimum: the conditions under which an AI agent could be meaningfully addressed by legal obligations; the liability architecture for harms caused by autonomous agents (including mandatory insurance, developer backstop liability, and sector-specific compensation mechanisms); and the institutional changes required to accommodate agents that are neither products nor persons within existing legal categories.

References and further reading

- Faroldi, F. L. G., "Risk for AI Agents," working paper, CERNAI, 2026.
- Faroldi, F. L. G., "Agentic and Systemic AI Risk," *Revista Iustitia*, 23, 2025.
- Faroldi, F. L. G., "Risk and Artificial General Intelligence," *AI & Society*, 2024.
- Faroldi, F. L. G., "Reasons-based AI Agents," *AI and Ethics*, 6, 77, 2026.
- Kolt, N., et al., "Legal Alignment for Safe and Ethical AI," arXiv:2601.04175, 2026.
- O'Keefe, C., et al., "Law-Following AI: Designing AI Agents to Obey Human Laws," *Fordham Law Review*, 94(1), 2025.
- Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act).
- Code of Practice for providers of general-purpose AI models, EU AI Office, 2025.

About CERNAI

The **Center for Reasoning, Normativity and AI** (CERNAI) at the University of Pavia investigates the formal and normative foundations of reasoning for AI systems. CERNAI's research spans formal logic, AI safety, ethics and governance of AI, and normative theory. Its director, Prof. Federico L. G. Faroldi, served on the EU AI Act Code of Practice Working Group (2024–25) and is Affiliate Faculty at UC Berkeley's Center for Human-Compatible AI.

Web: cernai.unipv.it | **Email:** cernai@unipv.it | **ISSN:** XXXX-XXXX